CHAPTER 7

ACHIEVEMENT TESTING

Lynda J. Katz
Gregory T. Slomka

INTRODUCTION

The second edition of the *Handbook of Psychological Assessment* appeared in 1990 and included a chapter on “Achievement Testing” by these authors. Developments and trends in the field are updated in this chapter.

During the mid-1970s the use of standardized tests among a variety of elementary, secondary, and post-secondary educational programs came under severe criticism. The use of standardized achievement tests in particular involved over 80 percent of American school children, with some of these children taking 26 achievement tests during a school career (National School Boards Association, 1977). And yet, as recent as 1992, 80 percent of the state system-wide tests given to some 14.5 million students were achievement tests. It has been postulated that the nation-wide concern with the use of standardized tests resulted from competition among the “baby boom” generation children of the late 1940s and early 1950s for “scarce slots in the choicest schools and businesses,” so that their stakes of doing well or poorly on tests went up. Second, those same baby boomers were looking back on their experiences with years of taking standardized tests and were very sensitive to the perceived abuses of such testing (Strenio, 1981, p. xviii). The main criticisms of these tests, however, have centered around the equality of the tests themselves; the use to which they are put; the behavior of the testing industry, with some 40 to 50 test publishers responsible for 90 percent of the tests used in the country today (Haney, Madaus, & Lyons, 1993); and the consequences for society of the misuse of these tests. In addition, major court cases and federal legislation for exceptional children have addressed specifically the use of tests and testing as part of the overall assessment process (Larry P. v. Riles and P.L. 94-142), again in response to these same criticisms.

In November 1975, at a conference on testing sponsored by the National Association of Elementary School Principals and the North Dakota Study Group on Evaluation, 25 national organizations, including the U.S. Office of Education, drafted the following statement:

We believe that the public, and especially educators, parents, and children, need fair and effective assessment processes that can be used for diagnosing and prescribing for the needs of individual children....

In regard to standardized achievement tests, we have agreed on the following recommendations:

1. The profession needs to place a high priority on developing and putting into wide use new processes of assessment that are more fair and effective than those currently in use and that more adequately consider the diverse talents, abilities, and cultural backgrounds of children.
2. Parents and educators need to be much more actively involved in the planning and processes of assessment.
3. Any assessment results reported to the public must include explanatory material that details the limitations inherent in the assessment instruments used.
4. Educational achievement must be reported in terms broader than single-score national norms, which can be misleading.

5. Information about assessment processes should be shared among the relevant professions, policy makers, and the public so that appropriate improvements and reforms can be discussed by all parties.

6. Every standardized test administered to a child should be returned to the school for analysis by the teachers, parents, and child.

7. Further, the standardized tests used in any given community should be made publicly available to that community to give citizens an opportunity to understand and review the tests in use.

8. The professions, the public, and the media need to give far greater consideration to the impact of standardized testing on children and young people, particularly on those below the age of ten.

9. A comprehensive study should be conducted on the actual administration and use of standardized tests and the use of test scores in the schools today. (National School Boards Association, 1977, p. 18).

In 1983 A Nation at Risk was published, one of the most widely publicized education reforms reports of the 1980s. In that report the authors warned “the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a nation and a people” (p. 5).

The National Commission on Excellence in Education went on to recommend that “standardized tests of achievement (not to be confused with aptitude tests)...be administered at major transition points from one level of schooling to another and particularly from high school to college or work” (p. 28). The purposes of testing would be to certify a student’s credentials, identify needs for remedial instruction and identify opportunities for accelerated work.

In 1990, U.S. President Bush and the National Governors Association announced the “America 2000” strategy for educational reform (National Education Goals Panel, 1990). That reform called for new achievement tests in the core subjects of English, mathematics, science, history, and geography. These tests were to differ from traditional norm-referenced assessments and focus instead on problem solving or task performance. The author of one recent review article has suggested that over-reliance on multiple-choice tests in the 1980s “led teachers to emphasize tasks that would reinforce rote learning and sharpen test-taking skills, and discouraged curricula that promote complex thinking and active learning (Wells, 1991, p. 55).

In addition, the Individuals with Disabilities Act (IDEA) of 1990 (Amendments to the Education for all Handicapped Children Act of 1975) called specifically for nondiscriminatory testing and multidisciplinary assessment (Hardman, Drew, Egan, & Wolf, 1993) for children with disabilities, explicitly supporting a major role for testing in the Individual Educational Plan (IEP). It has been estimated that between 8 and 20 million tests were used for special-education testing alone in the late 1980s (Haney, Madaus & Lyons, 1993). These estimates are based on 4.4 million students aged 3 to 21 years who served in special education programs in elementary and secondary schools between 1984 and 1985 (Snyder, 1987), and for whom an average of five to ten tests were used for initial assessment and one to two tests were used at least every three years thereafter. The majority of these tests were tests of achievement.

Thus, it remains both relevant and timely more than a decade later to

• review the historical development, classification, and psychometric properties of traditional achievement tests;
• update their status and use in terms of contemporary educational and clinical research and practice;
• consider the relationship of achievement testing to ecological and sociocultural variables and their use with special population groups; and
• take a futuristic look at the impact of modern computer technology on test construction and utilization.

Such a discussion may determine whether recommendations made 20 years ago regarding the use of achievement tests have been or will continue to need to be addressed.

### Historical Development of Achievement Tests

The standardized objective achievement test based on a normative sample was first developed by Rice in 1895. His spelling test of 50 words (with alternate forms) was administered to 16,000 students in grades 4 through 8 across the country. Rice went on to develop tests in arithmetic and language, but his major contribution was his objective
and scientific approach to the assessment of student knowledge (DuBois, 1970). Numerous other single-subject-matter achievement tests were developed in the first decade of the twentieth century, but it was not until the early 1920s that the publication of test batteries emerged; in 1923, the Stanford Achievement Test at the elementary level, and in 1925, the Iowa High School Content Examination (Mehrens & Lehmann, 1975). Since the 1940s, there has been a movement toward testing in broad areas as well, such as the humanities and natural sciences rather than in specialized, single-subject-matter tests. Moreover, attention has been directed toward the evaluation of work-study skills, comprehension, and understanding, rather than factual recall per se. In the 1970s, standardized tests were developed that were keyed to particular test books, the use of “criterion-referenced” tests (CRTs) emerged (their dissimilarity from norm-referenced tests will be addressed in the next section), and the development of “tailored-to-user specifications” tests (Mehrens & Lehmann, 1975, p. 165) was initiated.

Early in the 1990s, the literature on achievement testing was concerned with latent-trait theory, item-response curves, and an assessment of learning achievement that is built into the instructional process. With the later 1990s, concerns have tended to focus on the intrinsic nature of the achievement test itself. Computer-adaptive testing is not the computerization of standardized norm-referenced paper-and-pencil tests but a radically different approach. The approach is based on a concept of a continuum of learning and where a particular child fits on that continuum so that his or her experience with testing is one of success rather than failure.

In addition to computer-adapted testing, the use of alternative assessment tools has taken a front-row seat (Improving America’s Schools, Spring, 1996). This performance based assessment approach involves testing methods that require students to create an answer or product that demonstrates knowledge or skill (open-ended or constructed-response items, presentations, projects or experiments, portfolios). As Haney & Madaus (1989) have pointed out, these alternatives to multiple-choice tests are not new; and in fact, multiple-choice testing replaced these alternative forms of assessment in the late 19th and early 20th centuries because of the expense involved, the difficulties with standardization, and their use with large numbers of people. To appreciate fully this dramatic shift in the conceptualization of the assessment of achievement, it is first necessary to understand (a) the nature of tests which fall under the domain of achievement; (b) the psychometric underpinnings of achievement tests; (c) the basis for criterion-referenced as opposed to norm-referenced measurement; and (d) special issues which arise when achievement tests are used for particular purposes.

Classification of Achievement Tests

Achievement tests have generally been categorized as single-subject tests, survey batteries, or diagnostic tests and further dichotomized as group- or individually administered tests. Reference to the Ninth Mental Measurement Yearbook (Mitchell, 1985) reveals the prevalence of multitudinous published objective tests, and elsewhere it has been reported that some 2,585 standardized tests are in use (Buros, 1974). Table 7.1 is a listing of the most commonly used achievement tests. They have been categorized as (a) group administered, (b) individually administered, and (c) modality-specific tests of achievements, which can be either group or individually administered.

Typically one administers achievement tests in order to obtain an indication of general academic skill competencies or a greater understanding of an individual’s performance in a particular area of academic performance. In this regard achievement tests are specifically designed to measure “degree of learning” in specific content areas. There are several distinct applications of achievement tests which vary as a function of the setting in which they are applied. Tests such as the Metropolitan Achievement Tests, Stanford Achievement Tests, California Achievement Tests, and Iowa Tests of Basic Skills represent instruments that typically consist of test-category content in six or more skill areas. The benefit of the battery approach is that it permits comparison of individual performances across diverse subjects. Because all of the content areas are standardized on the same population, differences in level of performance among skill areas can reflect areas of particular strength or deficit. Many of these instruments provide a profile as well as a composite score that allows ready comparison of levels of performance between tests. The representative content of these batteries typically includes core assessment of language, reading, and mathematics abilities. The extensiveness of the coverage of allied curricula, that is, science,
### Table 7.1. Commonly Used Achievement Tests

<table>
<thead>
<tr>
<th>Group Administered Achievement Tests</th>
<th>California Achievement Tests</th>
<th>Iowa Test of Basic Skills</th>
<th>Metropolitan Achievement Test</th>
<th>Stanford Achievement Test</th>
<th>SRA Achievement Services (SRA)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Individually Administered Achievement Tests</th>
<th>Basic Achievement Skills Individual Screener (BASIS)</th>
<th>Kaufman Test of Educational Achievement</th>
<th>Peabody Individual Achievement Test-Revised</th>
<th>Wide Range Achievement Test 3</th>
<th>Woodcock Johnson Psychoeducational Battery-Revised</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Modality Specific Achievement Tests</th>
<th>Classroom Reading Inventory</th>
<th>Diagnostic Reading Scales</th>
<th>Durrell Analysis of Reading Difficulty</th>
<th>New Sucher-Alred Reading Placement Survey</th>
<th>Gates-MacGinitie Reading Tests</th>
<th>Gray Oral Reading Tests</th>
<th>Nelson-Denny Reading Test</th>
<th>Stanford Diagnostic Reading Test</th>
<th>Woodcock Reading Mastery Tests-Revised</th>
</tr>
</thead>
</table>

(continued)
Mathematics

Enright Diagnostic Inventory of Basic Arithmetic Skills

Keymath Revised

Sequential Assessment of Mathematics Inventories

Stanford Diagnostic Mathematics Test

Test of Mathematical Abilities

Language

Spellmaster

Test of Written Language-3

Woodcock Language Proficiency Battery - Revised

Written Language Assessment Test

humanities, and social studies, varies significantly. Sax (1974) provides a description of the major differentiating characteristics of 10 of the most commonly used achievement test batteries.

In contrast to the “survey” type tests or screening batteries described above are the more content-focused diagnostic achievement tests. Although any of the survey instruments is available to identify areas of academic strength or weakness (Radencich, 1985), they are not in themselves sufficient for diagnostic or remediation-planning purposes. Their use in screening large groups helps to identify those individuals in need of more specific individualized diagnostic evaluation. Through the use of a diagnostic battery, an area of identified deficit is examined in a more extensive fashion to determine what factors contribute to the academic dysfunction. Typically, these tests include a broad enough sampling of material so that areas of need are specified in order to develop remedial instructional objectives. For example, the Woodcock Reading Mastery Tests-Revised (Woodcock, 1987) provides five subtests which examine component processes associated with overall reading ability. These include Letter Recognition, Word Attack, Word Recognition, Word Comprehension, and Passage Comprehension. More in-depth examination at this level permits hypothesis generation regarding the nature of the specific academic deficit to be further tested. Similar tests are available to assess other aspects of academic performance: mathematics, spelling, writing, language skills, etc. Refined assessment at this level is necessary for differential diagnosis and remedial intervention. Screening batteries simply do not permit sufficient evaluation of an area for this kind of decision making to take place.

Although most achievement tests have the potential to be used as screening instruments to identify individuals in need of remedial instruction, fewer instruments actually appear to have been used for diagnostic purposes. In a national survey conducted in the early 1980s, Goh, Teslow, and Fuller (1981) reported that the Wide Range Achievement Test and the Peabody Individual Achievement served as the general achievement batteries most commonly utilized by school psychologists. At that point in time, in the area of specific achievement tests, the Key Math Diagnostic Achievement Test, the Illinois Test of Psycholinguistic Abilities (ITPA), and the Woodcock Reading Mastery Tests ranked as the instruments used most frequently for the assessment of specific academic content areas. However, in the late 1990s,
one rarely, if ever, encounters reference to the ITPA either in reported research studies or in diagnostic test reports used as part of an Individualized Education Plan.

**Criterion-Referenced versus Norm-Referenced Achievement Tests**

One other highly significant dichotomy must be addressed when discussing the classification of achievement tests and certain of their psychometric properties, namely, the distinction between criterion-referenced tests (CRTs) and norm-referenced tests (NRTs). While it is not possible to differentiate one from the other in terms of visual inspection (a criterion-referenced test can also be used as a norm-referenced test: for example, Basic Achievement Skills Individual Screener), there are intrinsic differences between the two approaches to achievement testing. Traub and Rowley (1980) described the decade of the 1970s as a time when "the notion of criterion-referenced measurement captured and held the attention of the measurement profession unlike any other idea" (p. 517). Mehrens and Lehmann (1975) asserted that the issues of accountability, performance contracting, formative evaluation, computer-assisted instruction, individually prescribed instruction, and mastery learning created a need for a new kind of test, the criterion-referenced test.

The concept of criterion-referenced achievement measurement was first detailed in the 1963 paper by Robert Glaser entitled "Instructional Technology and the Measurement of Learning Outcomes: Some Questions." In that landmark publication Glaser wrote:

Underlying the concept of achievement is the notion of a continuum of knowledge acquisition ranging from no proficiency at all to perfect performance. An individual’s achievement level falls at some point on this continuum as indicated by behaviors he displays during testing. The degree to which his achievement resembles desired performance at any specified level is assessed by criterion-referenced measures of achievement or proficiency... Criterion levels can be established at any point in instruction....

Criterion-referenced measures indicate the content of the behavioral repertory.... Measures which assess student achievement in terms of a criterion standard...provide information as to the degree of competence attained by a particular student which is independent of reference to the performance of others. (p.519)

Glaser further stated that achievement measures are appropriately used to provide information regarding a student’s capability in relation to the capabilities of his or her fellow students as well. Where an individual’s relative standing along the continuum of attainment is the primary concern, the appropriate achievement measure is one that is norm referenced. Whereas both CRTs and NRTs are used to make decisions about individuals, NRTs are usually employed where a degree of selectivity is required by a situation, as opposed to situations in which concern is only with whether an individual possesses a particular competence and there are no constraints regarding how many individuals possess that skill. Thus, at the core of the difference between the two kinds of tests is the issue of variability. “Since the meaningfulness of a norm-referenced score is basically dependent on the relative position of the score in comparison with other scores, the more variability in the scores the better” (Popham, 1971). This obviously is not a requirement of the criterion-referenced measure.

Because of basic differences in the theories underlying test construction, there have been several hundred publications on CRTs dealing with such issues as test reliability, determination of test length (Millman, 1973), score variability (Hambleton & Cignor, 1978; Hambleton, 1980), and test validity (Linn, 1982). The psychometric properties of CRTs have undergone close scrutiny, and one of the most critical dimensions reviewed has been the issue of validity. In the words of Linn (1980):

Possibly the greatest short-coming of criterion-referenced measurement is the relative lack of attention that is given to questions of validity of the measures. The clear definitions of content domains and well-specified procedures for item generation of some of the better criterion-referenced measures place the content validity of the tests on much firmer ground than has been typical of other types of achievement tests. Content validity provides an excellent foundation for a criterion-referenced test; but...more is needed to support the validity of inferences and uses of criterion-referenced tests. (p. 559)

In their review of 12 commercially prepared criterion-referenced tests, Hambleton and Cignor (1978) did not find a single one that had a test manual that included satisfactory evidence of validity (Hambleton, 1980). Validity has too often been assumed by both developers and users of criterion-referenced tests. This is no more acceptable for a criterion-referenced test than it is for any other test. It is time that questions of validity of the uses and
interpretations of criterion-referenced tests be given the attention they deserve.

Despite these criticisms from the point of view of traditional test-construction theory, criterion-referenced measurement has been found to have major utility with respect to the development of computer-assisted, computer-managed, and self-paced instructional systems. In all of these instructional systems, testing is closely allied with the instructional process, being introduced before, during, and after the completion of particular learning units as a monitoring, diagnostic, and prescriptive mechanism (Anastasi, 1982). Moreover, it has had practical applications with respect to concerns with minimum competency testing (Hunter & Burke, 1987; Lazarus, 1981) and mastery testing (Harnisch, 1985; Kingsbury & Weiss, 1979).

Curriculum-Based Measurement

In addition to criterion-referenced and norm-referenced tests of achievement, one additional “hybrid”—which appears to be surfacing, particularly in the area of special education—curriculum-based measurement (CBM), merits a brief note in this review. From the Institute for Research on Learning Disabilities at the University of Minnesota, Deno (1985) and his colleagues have proposed a method of measurement which lies somewhere between the use of commercialized tests and informal teacher observations. Their initial research with the procedure in the areas of reading, spelling, and written expression, and concerns with reliability, validity, and limitations are reviewed by Deno. Among the limitations are its utility only with the domain of reading at present, its lack of stability estimates as indicative of reliability, and its lack of generality that enables aggregation across curricula.

However, one aspect of CBM that appears to mark a distinct embarkation from traditional achievement testing is the concept of frequent measurement. In addition to the work of Mirkin, Deno, Tindal, and Kuehnle (1982) on the measurement of spelling achievement with learning disabled students, LeMahieu (1984) reported on the extensive use of a program of frequent assessment known as the Monitoring Achievement in Pittsburgh (MAP) which began in 1980 and involved 81 schools with a total enrollment of 40,000 students. Students were tested every six weeks with curriculum-based measures developed by committees of teachers. Serious risks in this kind of achievement testing involve the potential for teachers to narrow the curriculum and to teach to the assessment instrument as well as for students themselves to develop and refine test-wise behaviors as opposed to attaining specific academic skills.

USE OF ACHIEVEMENT TESTS

Achievement Tests in Education

Within the context of educational programs there is a continual process of evaluation that also includes teacher-made tests and letter-grade performance standards. The continuous monitoring of student performance within a particular academic content area provides means not only to assess student progress but also to link instructional strategies and learning objectives with identified student learning needs or skill deficits. Out of a concern for the performance of public schools, statewide minimum competency testing programs proliferated in the 1990s. “Policymakers reasoned that if schools and students were held accountable for student achievement, with real consequences for those that didn’t measure up, teachers and students would be motivated to improve performance” (Improving America’s Schools, 1996, p.1). Traditional achievement tests were judged to be “low-end” tests (p.1), and the advent of standards-based reform was seen as impetus to revamp methods of student assessment, a revamping which is ongoing at the time of this writing.

In a similar vein, a study by Herman, Abedi, and Golan (1994) assessed the effects of standardized testing on schools. They surveyed 341 elementary teachers in 48 schools, although the location of the schools was not identified. In their study, classes in which disadvantaged students were the majority were more affected by mandated testing than those serving their more advantaged peers. Results suggested that teachers serving disadvantaged students were under greater pressure to improve test scores and more driven to focus on test content and to emphasize test preparation in their instructional programs.

Despite such criticisms with respect to the misuse or inappropriate use of these tests, the periodic administration of achievement tests has traditionally been viewed as an educationally
Table 7.2. Achievement Tests: Purpose and Outcome

<table>
<thead>
<tr>
<th>PURPOSE OF TESTING</th>
<th>OUTCOME CRITERION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screening:</td>
<td>Identification of students potentially eligible for remedial programming.</td>
</tr>
<tr>
<td>Classification/Placement:</td>
<td>Specific academic deficiencies have been ascertained. Question now arises regarding whether student meets eligibility criteria.</td>
</tr>
<tr>
<td>Prescriptive Intervention:</td>
<td>A specific developmental arithmetic disorder is manifest in a child identified with visuo-perceptual processing problems. What curriculum adjustments appear warranted?</td>
</tr>
<tr>
<td>Program Evaluation:</td>
<td>Administrators seek to evaluate benefits of an accelerated reading program for gifted students.</td>
</tr>
</tbody>
</table>

sound procedure by professionals in the field. From a positive perspective, Anastasi (1988) provided a summary of their usefulness in educational settings. First, their inherent objectivity and uniformity provide an important tool in assessing the significance of grades. While individual classroom-performance measures can be susceptible to fluctuation because of a number of variables, their correlation with achievement-test scores provides a useful comparative validity criterion for grades. They are especially useful in the identification of students whose limited progress in a content area will require remedial intervention. Within this context, individualization of specific needs can be identified so that individual and group curricula can be modified. In this regard, the use of achievement tests prior to the initiation of training can become particularly efficacious. When these measures are utilized at the end of an instructional period they have the potential to serve as a means for assessing the quality of instructional programming and aiding in programmatic evaluation.

In general, then, achievement tests are used to make decisions, decisions which may involve instructional, guidance, or administrative issues. For example, what is the efficacy of a particular method of instruction? What are the specific outcomes of learning? Is there a need for remediation? Are grading practices accurate? Is the curriculum responsive to the acquisition of basic and specific academic skills? Is counseling appropriate for any given student? Is appropriate placement a concern? Thus, the breadth of the assessment will be predicated upon the rationale for the use of particular achievement measures. Table 7.2 illustrates the types of questions or problems that may be addressed and the expected benefit(s) to be derived from the testing process.

Achievement versus Aptitude

One further point, which any review of achievement tests must certainly address with respect to their classification and use, is the notion of aptitude versus achievement. This contrast dates back to the preoccupation of educational psychologists in the 1920s and 1930s with the role of heredity versus environment in the learning arena. This early simplistic notion that innate capacity or potential could be measured by aptitude tests independent of an individual’s learning history or “reactional biography” (Anastasi, 1984, p. 363) has been disavowed. Replacing the traditional concepts of aptitude and achievement in psychometrics is the concept of “developed abilities,” the level of development attained by an individual in one or more abilities (Anastasi, 1982, p. 395). In line with this conceptualization of the measurement of abilities, Anastasi provides a continuum of testing in terms of the “specificity of experimental background” that particular tests presuppose. The continuum ranges from course-oriented achievement tests to broadly oriented achievement tests to verbal-type intelligence to “culture-fair” tests. This continuum more accurately reflects the overlapping of aptitude and achievement tests. This analysis has been demonstrated empirically over and over in terms of the high correlations between achievement and intelligence tests. “In some instances, in fact, the correlation between achievement and intelligence tests is as high as the reliability coefficients of each test” (Anastasi, 1982, p. 395).

Finally, Anastasi notes that the continued labeling of some tests as aptitude or achievement measures has led to misuses of test results—in particular, the identification of certain children as underachievers when their respective achievement-test scores are lower than their scholastic aptitude- or intelligence-test scores. In the words of Anastasi (1982):
Actually, such intraindividual differences in test scores reflect the universal fact that no two tests...correlate perfectly with each other....Among the reasons for the prediction errors in individual cases are the unreliability of the measuring instruments, differences in content coverage, the varied effects of attitudinal and motivational factors on the two measures, and the impact of such intervening experiences as remedial instruction or a long illness. (p. 396)

Scoring Systems Associated with Tests of Academic Achievement

Before further discussion of the application of traditional achievement-test data, it is necessary to consider how the results of these tests are conveyed. Raw scores derived from achievement tests are typically converted to age- or grade-equivalent scores, standard scores, or percentile scores. Hoover (1984) makes a useful distinction between two scoring dichotomies. Developmental scores compare individual performance to that of a series of reference groups that differ systematically and developmentally in average achievement, with developmental scores being expressed as age- or grade-equivalent scores. Status scores compare test performance with a single normative reference group and are expressed as standard scores and percentiles. It is important to distinguish between the two types of measurement as each has unique strengths and limitations.

Developmental Scores

Age-Equivalent Scores. Educational Age (EA) represents a scoring criterion which has come under significant criticism and is used very infrequently in reporting educational test data. The scaling of items on some achievement tests is presented in a developmental sequence such that a particular score represents mean level of performance for a specific-age reference group. An individual who attains a specific score on the test is reported to function at a particular age level. This system of score reporting is useful for descriptive purposes, especially for “measuring growth.” As in grade-equivalent scores, which will be discussed next, serious flaws are encountered when one attempts to utilize such scores for comparative purposes.

Grade-Equivalent Scores. A grade-equivalent score (GE) reflects the presumed level of performance of an average student at a particular grade level. For example, if the mean score of a group of sixth graders on an achievement test is reported as M = 6.2, children who attain the same score are imputed to function at a level of performance commensurate with sixth graders in general. Although it is quite important to have available a continuous scale describing developmental level as a means to demonstrate progress in attainment and growth, the GE represents one of the most frequently misinterpreted sources of educational data. First, it should be noted that GE scores are reported in a format that reflects both grade level and month. The typical school year is approximately 10 months. Hence, scores of 6.2 and 6.9 contrast levels of performance commensurate with the beginning and end of the school year. There are, however, limitations on direct interpretation of GE scores. The scaling of achievement-test data is rarely a continuous process. Scores for many grade equivalents are frequently extrapolated or interpolated and consequently do not reflect actual derived scores. They are, in fact, estimations based on a hypothetical grade-equivalent curve. The use of such a scale also presumes that the teaching of such skills is a continuous process reflected across grades. This is not, however, reflected in the reality of the educational experience. Gains made by students are more realistically seen as a combination of spurts and plateaus, and not as a continuous process as is mathematically interpolated in scale construction.

The most significant limitation in the use of GE scores appears to arise because they are ordinal measures. The difference between a one-year gain in proficiency at a lower grade level in comparison to that same gain at a higher grade level may be significant. Further, because most of the basic core academic competencies are taught within the first through eighth grades, one cannot presume that grade-equivalent scores associated with the terminal stages of the educational career are equivalent. Finally, it must be noted that relatively small differences in performance can result in exaggerated differences in grade-level equivalency owing to the nature of scale construction.

The most frequently cited problem with using GE scores is the potential for misinterpretation of significant differences in level of performance. For example, a fourth grader obtains a score of
6.7 in reading. One cannot directly compare this youngster to other sixth graders. It is an erroneous assumption to state that this child’s reading ability is commensurate with that of a sixth grader. His reference group remains fourth graders. He clearly demonstrates well-above-average performance in comparison to this reference group. One cannot, however, compare him to sixth graders, who by the nature of their development and experience with reading, are different from our fourth grader. Because of the inherent potential for parents to set inappropriate standards of performance for their children based on such scores, the use of scores has been abandoned in many quarters.

**Status Scores**

A wide variety of standard score methodologies are available for reporting test results. These represent scores scaled along a continuum which permits one to ascertain where a particular score may fall in comparison to other scores in a distribution. There are two distinct advantages to the utilization of this scoring system. Standard scores permit the opportunity to compare individual performance to a normative standard, and they make possible the comparison of individual performance across two or more different tests. The latter represents an important criterion for the application of achievement tests within the context of a larger test battery.

**Percentiles**

Percentile rank represents a point in a distribution at or below which the scores of a given percentage of subjects fall. If a student scored at the 95th percentile, this would mean his or her score was better than 95 percent of the other students who took the same test. When clearly conveyed in the context of a psychological report, this scoring methodology represents one of the most readily understandable forms of test description. The potential for inappropriate comparisons of level of performance, as reflected in the GE score example, is significantly reduced.

**Standard Scores**

Standard scores represent raw scores that have been scaled relevant to a constant mean and standard deviation. As a function of the magnitude of the standard deviation, one can, through linear transformation, readily ascertain how far from the mean performance lies. Most tests standardize scores within defined age groups. Therefore, regardless of the age of the subjects under evaluation, a specific standard score will have the same meaning. For example if two students, ages 8 and 10 years, obtain the same standard score on a reading test, relative to the normal curve, one can readily distinguish that in comparison to their age mates, they are functioning at equal distance from the mean. Standard score conversions also include z scores, t scores, and occasionally stanine scores which can be interpreted in like manner. In general, standard scores are considered the more accurate and precise means of reporting test results. Finally, it is not uncommon for test developers to provide multiple methods for performance description. For example, the Wide Range Achievement Test-3 provides grade equivalent scores, age-based standard scores, percentiles, normal-curve equivalents, and absolute scores.

**Achievement Test Scores and the Diagnosis of Learning Disabilities**

The relevance of understanding the scoring systems utilized in the interpretation of achievement-test results can be dramatically illustrated when one considers the educational diagnosis of a specific learning disability. Learning disabilities have become the dominant handicap of school-age children in the country, with some 42 percent of all students ages 3 to 21 years in special education programs diagnosed as learning disabled (Databank, 1985).

A basic assumption underlying learning disabilities is the failure of the student to acquire primary academic skills at levels expected for age, grade placement, and level of intellectual functioning. The identification of individuals with learning disabilities has traditionally been based on the notion of a “significant discrepancy” between ability level and demonstrated academic skill attainment. Regardless of which of the many formulas is used to diagnose a learning disability, all require data from standardized achievement tests. Thus, the use of achievement testing has become an integral component in the differential diagnosis of learning disabilities. In this regard, the concept of “significant discrepancy” has been an important one, for it forms the basis for distinguishing specific learn-
ing-disability diagnoses from conditions such as underachievement or mental retardation.

Under Public Law 94-142, the Education for All Handicapped Children Act of 1975, it was specified that a team could render a determination of specific learning disability if a child did not achieve at his or her ability level when provided with appropriate educational instruction and if a severe discrepancy existed between intellectual ability and achievement in one or more of seven areas of achievement, including oral expression, listening comprehension, written expression, basic reading skills, reading comprehension, mathematics calculation, or reasoning. Specifically excluded along with mental retardation were other factors which could impinge on limited academic proficiency, such as peripheral sensory or motor handicaps, emotional disturbance, or socioeconomic or cultural disadvantage. The actual specification of the means of ascertaining discrepant performance is left vague in this definition. Algozzine, Ysseldyke, and Shinn (1982) emphasize that the field of learning disabilities has always suffered a definitional dilemma. Federal guidelines have not appreciably corrected this situation. No clear consensus across school districts exists nationally for arriving at workable definitions of learning disability diagnoses (Shaw, Cullen, McGuire, & Brinkerhoff, 1995).

In spite of the lack of consensus regarding definition, the notion of severe discrepancy has been defined most frequently by the use of an ability-achievement discrepancy. Inherent in this conceptualization of learning disability is the potential for at least average intellectual functioning with academic performance well below expectations. A number of strategies have been applied in an attempt to operationalize criteria representative of a severe discrepancy.

Deviation from Grade Level

A commonly encountered criterion used to define a potential learning disability might be "grade level performance in academic achievement two grade levels below expectation for age." This criterion has been criticized as inadequate for a number of reasons. First, as previously discussed, grade-level equivalents represent the weakest psychometric criterion upon which to base comparisons of academic performance. Second, utilization of such a constant criterion fails to take into consideration the significance of discrepant performance at various points in the continuum of educational programming. For example, performance two grade levels below expectation in a third grader can be far more significant than the same magnitude of score deficit in an eighth grader. Further, in the assessment of adult populations, the efficacy of grade-equivalent scores loses predictive validity. It is extremely difficult to ascertain whether eighth-grade academic skills in a 40 year old are indicative of any significant disparity in level of performance.

Finally, problems have been identified with potential identification of students who may be learning-disabled. Use of grade-level discrepancy criteria tends to overidentify children whose intellectual functioning is below average and to underidentify those students who may be above average. A student with an IQ of 82 might in fact be functioning at a grade level which is not discrepant for his or her overall level of intellectual functioning. On the other hand, a fourth grader who is reading at or just below grade level, but who has an IQ in the superior range and who should clearly be reading at well above grade level expectations, would be excluded.

Standard Score Discrepancy Models

The process of comparing standard scores derived from academic and intelligence tests holds apparent benefits over grade-discrepancy scores on purely psychometric grounds. Typically, a criterion level is arbitrarily selected, a 1 or 2 standard-deviation-point discrepancy between general ability and achievement test score. This methodology can, however, also impose bias into the discrimination process. Many such models do not take into consideration the regression of IQ on achievement. One cannot assume direct correspondence between IQ and standard score equivalents. It can be demonstrated that academic achievement-test scores fall somewhat short of IQ for individuals, manifesting above-average performance, and in lower-functioning individuals, academic achievement-test scores are actually higher. The use of a simple discrepancy-score formula implicitly assumes a perfect correlation between general ability and achievement tests which in fact does not exist. It would also require that each test be based on the same standard-score distribution.
Regression Equations

The most sophisticated methodologies available for determining significant score discrepancies are based upon complex computations or tables designed from formulas based on regression equations. A number of strategies have been developed, each with unique distinguishing properties. A number of reviews are available (Forness, Sinclair, & Guthrie, 1983; Reynolds, 1984; Wilson & Reynolds, 1984) that describe the characteristics of these methodologies. There remains, however, no one mathematical model that is commonly accepted or in fact utilized.

Reynolds (1984) reported on the findings of the Work Group on Measurement Issues in the Assessment of Learning Disabilities, a study section formed in 1983. This group was delegated the responsibility of addressing questions directed toward identification of “best practice” solutions to the learning disabilities definitional dilemma. In their findings, models of discrepancy analysis based upon grade-equivalent scores were rejected outright. Factors related to their imprecision and their ready misinterpretation were noted. Most critical, however, was the inherent lack of the mathematical properties necessary for conducting comparative analyses that are associated with this scoring system. The group concluded that age-based standard-score discrepancy models represent potentially the best methodology available. However, while developmental standard scores are to be preferred over grade-level or status-standard scores, their value has been challenged also because they require greater growth for below-average children than for average or above-average children (Clarizio & Phillips, 1986).

One cannot, however, focus exclusively on the concept of discrepancy as the sole basis for the diagnosis of a learning disability. To quote Reynolds (1984), “The establishment of a severe discrepancy is a necessary but insufficient condition for the diagnosis of a learning disability” (p. 468). A host of factors other than a specific learning disability (LD) can contribute to significant academic underachievement. Among these are limited sociocultural opportunity, dysmotivation, sensory-perceptual dysfunction, or functional psychiatric impairment. It is Reynolds’ bias, however, that only when a severe discrepancy can be demonstrated is a child considered eligible for a diagnosis of LD. This bias has come under severe criticism of late because identification, and therefore remediation, must wait until the student fails (Shaw, Cullen, McGuire, & Brinkerhoff, 1995). In contrast, research studies consistently support the efficacy of early intervention. Studies supporting the identification of reading problems as early as the pre-school years, with programs in kindergarten that include a focus on phonological and orthographic awareness, are compelling in this regard (Foorman, Francis, Beeler, Winikates, & Fletcher, 1997; Wasik & Slavin, 1993; Lundberg, Frost, & Petersen, 1988).

Some Thoughts on the Validity Issue

There are, among educators and researchers, those who question the focus on the reliability of the IQ and achievement discrepancy versus its validity (Shepard, 1983). In a study by Shepard and Smith (1983), which evaluated the identification practices of psychologists and teachers within the state of Colorado involving 1,000 student files and 2,000 specialists, 50 percent of those professionals surveyed were unaware that an IQ of 90 falls at the 25th percentile. For children with IQs of 90, the expectation was that achievement would be at grade level (the 50th percentile) because the IQ was “in the normal range” (Shepard, 1983). The authors continued that these specialists were unaware also that after the first or second grade, it is not uncommon for large numbers of children to have grade-equivalent scores below their grade placement.

Other technical problems identified in this study further complicate the identification of LD. Most of the tests used in the diagnosis of LD were technically inadequate with the exception of the WISC-R and one or two achievement batteries. Many clinicians were unaware of the differences between technically adequate and inadequate tests. Specialists often selected technically inadequate measures even when more valid instruments were available; their choices tended to follow traditional preferences associated with each professional group. Many clinicians continued to apply inaccurate conventional wisdom regarding the symptoms of the disorder (relying on interpretations of subtest scatter, underestimating normal patterns of difference, etc.) (Shepard & Smith, 1983).

Reynolds (1984) and the Task Force advanced a number of recommendations which attempted to bridge this validity-reliability gap with respect to the diagnosis of LD:
1. Instruments applied should meet criteria defined in PL 94-142.
2. Well-standardized national norms should form the basis for statistical comparison of individual levels of performance.
3. Normative comparisons should be based upon co-normed samples. The ideal scenario is one in which the two tests compared are normed on the same sample. Where this is not possible, the two normative groups should be clearly comparable.
4. Only individually administered tests of achievement and intellectual ability should be utilized.
5. Age-based standard scores based upon a common scale represent the most statistically robust means for score comparison.
6. Measures employed should conform to acceptable criteria for validity and reliability.
7. Special technical considerations should be addressed when using performance-based measures of achievement (e.g., writing skill).
8. Bias studies should have been conducted and reported.

In summary, while the psychometrics involved in scoring and interpreting the results of achievement tests can be fraught with complexity and controversy, as illustrated in the case of the diagnosis of learning disabilities, the consequences of the resolution of the issues involved are even further reaching. Consider the effects of labeling, the contraction of teacher competence to deal with a variety of learning styles in the classroom, the allocation of resources available to those students with the most severe disability, and the costs of providing for special education resources themselves (Shepard, 1983). All of these can be viewed as negatives. It is not difficult nor unrealistic to extrapolate these same issues to include diverse groups of students in educational programs today. Thus, we are left with ethical responsibilities to insure the appropriate utilization of achievement tests based on the most current thinking and research available, which is macrocosmic rather than microcosmic in nature.

Messick (1980) has argued this point in his “Test Validity and the Ethics of Assessment.” He had written earlier, with specific reference to the measurement of personality, that tests should be evaluated not only in terms of their measurement properties but also in terms of their potential social consequences (Messick, 1965). Messick emphasized the importance of construct validity, arguing “that even for purposes of applied decision making reliance upon criterion validity or content coverage is not enough” (Messick, 1975, p. 956), and that “the meaning of the measure must also be comprehended in order to appraise potential social consequences sensibly” (Messick, 1980, p. 1013). He defined test validity as an overall evaluative judgment of the adequacy and appropriateness of inferences drawn from test scores, opining that values questions arise with any approach to psychological testing, whether it be norm-referenced or criterion-referenced, a construct-based ability test, or a content-sample achievement test. This evaluative judgment of test validity is based on (a) convergent and discriminate research evidence as to the test scores interpretability in terms of the particular construct under review; (b) an appraisal of the value implications of that interpretation; (c) justification of the relevance of the construct and its utility of the particular application proposed; and (d) dealing with the potential social consequences of the proposed use as well as the actual consequences upon implementation of the testing procedure.

Intervening in the model between test use and the evaluation of consequences is a decision matrix to emphasize the point that tests are rarely used in isolation but rather in combination with other information in broader decision systems. The decision process is profoundly influenced by social values and deserves, in its own right, massive research attention beyond the good beginning provided by utility models. (Messick, 1980, p. 1025)

Messick concluded his remarks by paraphrasing Guion (1976): “The formulation of hypotheses is or should be applied science, the validation of hypotheses is applied methodology, but the act of making...(a) decision is...still an art” (p. 1025).

The Use of Achievement Tests in Clinical Practice

Achievement testing conducted with clinical populations is generally regarded as an extension of intelligence and aptitude testing. It provides one further means to ascertain “general ability level.” Results are typically utilized for drawing inferences regarding the capacity of the individual under evaluation to apply knowledge or native intelligence in practical problem-solving situa-
tions. One equates intelligence and exposure to educational opportunity with the ability to conform with the demands of achievement testing at commensurate levels of success. Typically, one is not engaging in achievement testing with this population in anticipation of identification of potential performance discrepancies, but to gauge overall adaptive competency. The identification of any significant discrepancies would of course result in further clinical investigation. Cognitive as well as noncognitive variables would then be explored.

Achievement Test Results Applied in Neuropsychological Evaluation

Achievement tests play a definitive role in the administration of standard neuropsychological test batteries. For example, a number of extended versions of the Halstead-Reitan Neuropsychological Test Battery include an administration of the Wide Range Achievement Test or another age-appropriate screening battery within the test protocol. Data derived from such tests offer clinical utility beyond discrepancy analysis. They can be used as a method to infer an estimated level of premorbid intellectual functioning (Lezak, 1983). As basic academic skill competencies are generally not susceptible to significant deterioration in mild-to-moderate generalized cerebral dysfunction, standard scores derived from general achievement-test measures offer one means to interpolate a coarse estimation of premorbid functioning when other means of documentation are not available.

Achievement-test results can be incorporated into the pattern analysis of other neuropsychological test variables to aid in the specification of the effects of focal-lesion processes. For example, problems exclusively with the spatial components of arithmetic processes in an individual manifesting no evidence of linguistic defects would help suggest a post-Rolandic lesion of the right cerebral hemisphere, when other markers of right hemisphere dysfunction are present. It is not uncommon to consider achievement-test performance within the context of a formal aphasia examination as a means to extend the assessment to the integrity of lexical-skill functions and writing ability.

Beyond their application in the documentation of the effects associated with focal-lesion processes, such test results hold even greater potential utility in aiding in the development of hypotheses regarding functional limitations associated with cerebral dysfunction. As primary academic-skill competencies are intimately related to aspects of autonomous functioning in a number of instrumental activities associated with daily living, the degree of preservation of such primary skills as reading and arithmetic abilities can be important prognostic indicators associated with long-term recovery and adaptation.

Achievement Test Results Applied to Rehabilitation Assessment Methodologies

In the areas of both psychiatric and vocational rehabilitation, the specification of the degree to which core academic competencies are developed holds a number of prognostic implications. With low-level functioning individuals, the specifications of primary literacy skills is an important determinant of the level of complexity of programming in which they might participate. The degree to which a learning curriculum might emphasize effective reading comprehension might be potentially exclusionary, for example.

An important component of the rehabilitation assessment is determination of the degree to which any remedial intervention might be required prior to implementing programming. Inadequate educational opportunity or underachievement related to psychosocial factors must be distinguished from developmental academic disorders and conditions which cause a loss of previously attained ability. Intervention strategies to remediate or supplant deficient academic skills are determined by the thorough analysis of their cause. Prognostically, it is important to identify those individuals functioning at their plateau versus those who have the potential to develop these skills further.

In summary, with the use of achievement testing in clinical settings the focus is typically divested towards two lines of inquiry: (a) obtaining knowledge of the degree to which basic academic skill competencies are developed in a particular individual, and (b) examining individual performance within a particular area of academic performance. The basic referral question in large measure determines what armamentarium of techniques will be brought to bear in the assessment. It will also influence how test scores will be compared and interpreted.
COGNITION, METACOGNITION, AND ACHIEVEMENT TESTING

The application of cognitive-theory research to educational psychology can be traced back as early as 1960 with the publication of David Ausubel’s paper “The Use of Advance Organizers in the Learning and Retention of Meaningful Verbal Material,” the later work of Rothkopf (1965) on mathemagenic behaviors, Ausubel’s (1968) text, Educational Psychology: A Cognitive View, Anderson’s (1972) work on how to construct achievement tests to assess comprehension, and the work of Marton and Saljo (1976a, 1976b) who argued that a description of what is learned is more important than a summary of how much is learned (Clarke, 1982). Glaser (1981) reviewed current research in cognitive and developmental psychology addressing its potential influence on the development of new psychometric methodology. He cited Bartholomae’s (1980) work on error analysis with college students in remedial-writing programs and Siegler’s (1976) work on rule assessment in the acquisition of scientific concepts as illustrative of the “necessary interrelationships between the analytical assessment of performance and effective instruction” (Glaser, 1981, p. 929). Interest in the assessment of mastery or competence can be traced also to developments in cognitive psychology, artificial intelligence, and language understanding. Herein the works of Chase and Simon (1973) on the chess master and the work of Larkin, McDermott, Simon, and Simon (1980) on problem solving in the area of elementary physics were cited by Glaser.

Finally, research in the realm of metacognition—the knowledge, regulation, and management of one's own cognitive processes and products—(Flavell, 1976) has led to a concern with the measurement of these self-regulatory skills in terms of predicting successful problem solving which then leads to learning. Metacognitive abilities develop with maturity, and current research in learning instruction has demonstrated that these skills may be less well developed in those individuals who have learning disabilities.

Thus, it becomes quite clear that an understanding of the learning process and its assessment can yield more fruitful data than those traditionally obtained by achievement tests. This is particularly important in light of the social-educational demands outlined by Glaser (1981) which will shape and mold the future of educational assessment:

- the shift from a selective educational system to one designed to help individuals succeed in educational programs (zero-reject system);
- the requirement for improved levels of literacy and problem-solving ability in a variety of knowledge and skill domains (minimum competency and mastery certification);
- the need to understand individual differences in the process of measurement so that abilities can be improved to facilitate further learning (cognitive, sociocultural, gender specific).

The application of cognitive and metacognitive principles with respect to the measurement of learning have been detailed in the areas of reading (Curtis, 1980; Curtis & Glaser, 1983), spelling (Henderson & Beers, 1980; Nolen & McCartin, 1984), and foreign language (Fischer, 1981; Stevenson, 1983; Terry, 1986). Curtis and Glaser (1983) describe the current level of understanding regarding the theoretical framework utilized to study the process of learning to read, a process which involves a complex of interrelated skills (word decoding, accessing semantic word-information, sentence processing, and discourse analysis), proficiency in one affecting success in the others. The results of traditional reading-achievement tests have made it impractical to diagnose reading problems in terms of remediation or instructional strategies thus far. However, current theory on efficiency in word identification, the qualitative features of semantic knowledge, and research on schemata can be utilized as a form of construct validity and thus allow measurement of achievement that reflects both the development of competence and the process of instruction. “With developing knowledge of reading it should be possible to establish standards of performance...[and]...combined enterprise representing test design based on knowledge of human learning and performance, psychometric requirements, and studies of test use should improve our ability to link testing and instruction” (Curtis & Glaser, 1983, p.144).
Diagnostic Application of Achievement Test Results

As an illustration of the application of cognitive and metacognitive strategies in the process of achievement testing, the remainder of this discussion focuses on an expanded level of analysis that can be undertaken in the clinical setting for purposes of both diagnosis and remediation interventions.

Reading

Converging lines of research (Fletcher et al., 1994; Shaywitz, Fletcher, & Shaywitz, 1996) have emphasized the primacy of core phonological-processing deficits in disabled readers. Unlike language competencies which unfold naturally in a fairly predictable fashion, reading represents an acquired skill. As such, not only constitutional but environmental determinants may contribute to failures in reading. Without an explicit model of normal reading development, patterns of impairment cannot be described.

Reading development has been traditionally dichotomized across two component skills which must be mastered. These have been described within a dual-route model of reading as involving (a) a phonological, and (b) a direct lexical route in which whole-word or orthographic-recognition skills facilitate active word-recognition (Morton, 1969). In the earliest stages of the development of pre-reading competencies, the logographic stage, mastery of the visual-orthographic properties of letters, memorization of the visual gestalts of a limited repertory of words, and utilization of visual associative skills to foster word recognition from pictures that accompany text, are accomplished. The early reader is simultaneously developing sound-symbol associative skills. Such abilities are predicated upon a child’s ability to first decompose speech into component structures (phonological awareness). Via these processes, the early reader is learning to associate visual symbols (graphemes) with their corresponding sound equivalents (phones). With increasing experience, direct orthographic-associative competency is established and familiar words are decoded on “sight”. Confrontation with low frequency, novel words is presumed to require some combination of orthographic- and phonological-coding abilities (Coltheart, 1978).

Skilled readers are presumed to have developed automatized orthographic skills in reading by the fourth grade. As such, phonological skills are relegated primarily to the processing of less familiar words. A competent reader is presumed to have developed equivalent proficiency in both aspects of word analysis. In addition, with advancing age, the capacity to incorporate morphological cues as well as lexical-semantic, and other contextual cues, further contributes to the act of reading. As such, words rich in meaning tend to be decoded with greater faculty than more ambiguous function words.

The complexity of reading-skill acquisition expressed over time requires recognition of the reciprocal contributions of higher-level processing systems, beyond the dual-route model. Cognitive models of reading development (see Chase & Tallal, 1991, for review) take into consideration increased capacity to bring on-line higher cognitive faculties (the simultaneous development of not only “bottom-up” but “top-down” processing) to the development of reading fluency and comprehension (McClelland & Rumelhart, 1981). Progression in reading proficiency requires the act of word recognition becoming automatized (Liberman, Liberman, Mattingly, & Shankweller, 1980). Thus, on-line cognitive activities are directed less at the act of decoding and permit the use of metalinguistic awareness, selective attention and working memory to support semantic encoding and comprehension monitoring. Although phonological-coding skills (word-analysis abilities) represent the most widely studied aspect of reading development, more recent research has focused upon the contribution of other types of linguistic-rule knowledge (semantic, morphological, and syntactic conventions) to the development of higher-level reading skills, i.e., comprehension skills.

Associated with research on phonetic processing in deficient readers has been the identification of specific linguistic deficiencies associated with reading disabilities. In a review by Mann (1994), deficiencies associated with naming and verbal productivity, expansion of semantic knowledge, auditory sequential memory, sentence recall, and grammatical and syntactical analysis skills (particularly comprehension demands requiring the processing of more complex grammatic structures) have been identified. Catts (1989) further identified higher rates of oral-speech deficits, including early history of articulation inefficiencies among the reading disabled. A commonality linked to these
deficits involves the processing of sound patterns of language. These deficits have been broadly conceived as evidence that specific maturational lags in the systems supporting language development represent a secondary source of the cognitive deficit expressed by disabled readers. The specificity of language dysfunction related to reading inefficiency is conveyed by Tallal’s finding (1987) that 85 percent of children exhibiting language disorders in the preschool years develop language-related learning disabilities (i.e., reading problems).

In addition to phonological and language-processing abilities, visual-feature analysis is also required in the act of grapheme-phoneme correspondence. Although low-level visual deficits have been identified among reading-disabled populations (Lovegrove, Martin, & Slaghuis, 1986), their impact as factors significantly impinging upon reading-skill development appears minimal (Hulme, 1988; Vellutino & Scanlon, 1987). In addition, select memory inefficiencies have been identified in some impaired readers. A sparse network of associations in working memory and retrieval deficits have been hypothesized in these instances. Employing hierarchical-regression analysis, Vellutino, Scanlon and Tanzman (1994) utilized measures of phonological coding and analysis abilities, verbal-memory measures, semantic- and syntactical-analysis tests, and visuo-perceptual task demands as dependent variables in predicting word-analysis proficiency. Phonological-processing skills accounted for the majority of the variance associated with word-identification proficiency. Semantic and syntactic measures were identified as intermediate predictors of reading proficiency and visual abilities.

Findings such as these have reshaped conventional wisdom applied to the assessment of reading disabilities. Dissociations between normal versus impaired readers have been traditionally specified by decision rules, i.e., aptitude-achievement test discrepancies. Children with a discrepancy between IQ and an objective reading measure are classified as disabled and deemed eligible for special-education supports. Low-achieving readers are conceptualized as reading below normative standards for age. But because of associated lower level IQ scores and imputed, more generalized cognitive inefficiencies, these low-achieving readers are not presumed eligible or appropriate for special-education services. These long-standing practices have led to a bimodal conceptualization of reading deficiencies, with reading disabilities representing a hump on the lower tail of this distribution (Shaywitz, Escobar, Shaywitz, Fletcher, & Makuch, 1992). Data derived from the Connecticut Longitudinal study (Shaywitz et al., 1992) support the contention of Stanovich (1991) that there actually may be no qualitative differences between disabled and low-achieving readers, and that phonological-processing deficits represent a core deficit indentifiable in both groups. These investigators have argued that reading abilities exist on a continuum which includes superior, average, and impaired readers. This model argues against the use of any arbitrary cut point indicating normality versus disability (i.e., discrepancy models) and instead suggests that intervention for any individual with reading inefficiencies be driven by identification of his or her unique processing or associated cognitive dysfunction.

These findings have aided in validating Stanovich, Nathan, and Zolman’s (1988) initial hypothesis regarding the variability expressed among impaired readers. This model presumes all disabled readers manifest a phonological-processing deficit. The most severe forms of reading disability are characterized by a fundamental or “core” deficit in the ability to establish grapheme-phoneme correspondence. As the nature of manifest impairment extends beyond core phonological-processing deficiencies, the term “variable” is attributed to the idiosyncratic manifestation of other language, attention, memory, or perceptuo-integrative skill deficits that may be additionally expressed. The model takes into account the remarkable heterogeneity expressed in reading deficiencies and why categorical models of reading, e.g., subtyping schemas, may not satisfactorily characterize the unique attributes expressed in individual cases.

Conceiving of reading problems in this fashion emphasizes the importance of defining the individual array of strengths and weaknesses expressed by any reader. This represents an alternative to models which posit more discrete subtypes of disabled readers and permits a means to conceive of reading on a continuum from normal variability in reading.
proficiency to the heterogeneous expression of impaired reading development.

Assessment

A major portion of diagnostic reading assessment focuses on the sophistication and accuracy of decoding skills. This assessment is accomplished through the presentation of reading material as isolated phonemes, nonsense words, familiar and unfamiliar words, as well as words presented “in context,” that is, in the form of sentences or complex paragraphs. At a first level of analysis the rule-out of basic visuo-perceptual dysfunction is necessary. The reader must be able to appreciate fully the visuo-symbolic configuration of letters and words. Here one is concerned with the rule-out of visual-sequential and modality-specific attentional deficits which could prevent the accurate assimilation of the written material. Perceptual errors such as reversals (reading “b” for “d” or “p” for “q”) would also be excluded.

With the rule-out of primary perceptual dysfunction, analysis of grapheme-phoneme correspondence is undertaken. Basic decoding ability is ascertained for vowels, consonants, and consonant blends of letter combinations. Increasing the level of complexity of syllabic blends permits analysis of any sequential information-processing deficits that may be present. One is interested in the capacity not only to analyze and decode written material sequentially but aural material as well.

There are tasks which tap auditorization or syllabication, that is, the ability to decode the component phonetic properties of a word. On the Auditory Analysis Test, for example, one is asked to say “Germany” without the “ma” sound, thus transforming the remaining syllables to “journey.” Some individuals, who on the Word Attack subtest of the Woodcock Reading Mastery Tests are reasonably successful in reading isolated phonemes, have great difficulty blending these same sounds into their appropriate phonological expression when confronting them in complex words. For example, when asked to read “phonological” the student struggles to isolate—“pho”...“no”...“loge”...“ee”...“cal”—only to pronounce the word then as “phonograph,” a word more embedded in auditory memory. Frequently the effort required to analyze words laboriously in this fashion is exacted at great expense in terms of comprehension and memory for material read.

Assessment techniques that require rapid identification of words serve as a means to assess sight-recognition vocabulary. Speed of recognition is not factor-controlled in many types of reading tests. “Automatic recognition” represents the most sophisticated and efficient means of reading. Reading performed at this level taxes working memory minimally and frees the reader to focus on the semantic organization of the material for greater understanding and for committing textual information to memory. There are, however, individuals who have not attained adequate levels of sight-recognition skills. They maintain a more labored phonologically based reading style. These individuals may present a variety of deficits that impede their ability to process complex visuo-symbolic material. This might involve visual inattention, visuo-perceptual processing problems, spatial- or gestalt-recognition deficits, or weak visual memory. An analysis of the approach taken during “word attack” can be helpful in isolating the contributing deficit or deficits.

Within this context, the overall complexity of the word presented can be important. Errors encountered with relatively simple reading material can suggest problems in processing the basic visual morphology of written material. In terms of the simultaneous processing of visual input, there may be a finite limit on how complex a word can be for it to be realized. In attempts to compensate, some children “guess” at the whole word by processing only the prefix or first few syllables. Poor visual-gestalt functions or whole-word recognition skills are usually typified by gross lexical “word substitution” errors. Here words that share a similar visual gestalt to the word at hand are substituted, often resulting in flagrant misreading. In this regard it is necessary to rule out impulsivity as a contributing factor. The absence of other evidence of attention-deficit-disorder symptoms in ancillary testing or observation is particularly helpful.

Finally, comparisons of the relative efficiency of oral and silent reading under timed conditions can be potentially useful. A sample of oral reading of both word-recognition material and passage material can be extremely beneficial. Dramatic improvement in passage versus isolated-word reading immediately suggests the potential for the reader to compensate via the use of semantic cues. There are students whose oral-reading efficiency can be significantly compromised by anxiety or
inhibition. Far greater efficiency can be expressed by them in silent reading.

Reading Comprehension

Examination of reading comprehension is generally undertaken via the reading of a paragraph and the answering of questions about the content. A quantitative score is applied based on the number of correct responses and an estimation of reading-comprehension level is ascertained. This procedure does not in itself reflect the myriad factors which can contribute to comprehension difficulties. Level of investment can be a significant factor. Motivation can be influenced by interest in the factual material presented as well as general investment in reading as a preferred learning modality. Basic reading proficiency in terms of adequate word-recognition skills will also influence comprehension. Without strategies for the decoding of unfamiliar or complex reading material, adequacy of understanding will suffer. There are also a number of higher cognitive skills that influence performance, including linguistic proficiency, memory, cognitive flexibility, and semantic-organization skills. In order to ascertain where on a continuum of contributing factors comprehension problems lie, a number of informal strategies have been recommended to augment the reading-comprehension examination (Aaron & Poostay, 1982; Levine, 1987).

These strategies focus on the reading of restricted passages of known grade-level difficulty with the examiner focusing on a number of direct questions that permit an informal task analysis of potential contributions to comprehension failure. For example, Levine (1987) recommends beginning with the oral reading of simple sentences as the starting point. Limiting the amount of information to be assimilated restricts the degree to which active memory and semantic organizational skills are required, thus permitting direct access to potential problems based on decoding lexical information. At this level, basic questions regarding word-recognition errors, limited functional vocabulary, and problems with understanding morphology and syntax can be ascertained. Increasingly more complex lexical material is then presented. With each passage a number of profiles are presented in which the reader is asked to recall details, sequence events, and identify main ideas. More sophisticated demands can be made, such as summarizing the overall content of the passage. Responses can be evaluated on a continuum of literal to inferential depending on their level of complexity. It is also of value to compare general level of performance on oral comprehension and memory tests to determine whether reading comprehension is related to more generalized cognitive impairment.

The comprehensive evaluation of reading competencies requires utilization of diverse methodologies that typically involve an amalgam of standardized tests. In addition to academic achievement tests, language measures which tap lexical retrieval, semantic knowledge, linguistic short-term memory, as well as auditory comprehension are required. The Test of Language Development-Primary (Newcomer & Hammill, 1988), Test of Language Development-II (Intermediate) (Hammill & Newcomer, 1988), the Clinical Evaluation of Language Fundamentals-R (Semel, Wiig, & Selord, 1995), and Test of Adolescent and Adult Language-3 (Hammill, Brown, Larson, & Wiederholt, 1994) represent important adjunctive measures.

Thus, the primary role of the diagnostician should be geared less towards the documentation of any aptitude-achievement disparity and more upon the multivariate description of underlying cognitive processes contributing to reading impairment. While categorical diagnosis remains a requirement for eligibility determination within most classificatory systems, the potential explanatory power of assessment lies not with the discrepancy analysis, but with detailed multi-variate description of the constituent cognitive processes subserving reading. In this fashion, the cognitive basis of reading impairment can be linked empirically to remediation strategies.

Mathematical Abilities

In the past, an at-minimum fourth-grade math competency was considered adequate for adult functioning. Adaption to an increasingly technological society requires greater fluency in mathematics (Semrud-Clikeman & Hynd, 1992). The complexity of the subject matter, predominance of the use of a spiral curriculum, and other factors related to instructional technology have been related to trends noted in the preceding two decades that reflect lower overall math achieve-
ment in American children. Math-skill development is the subject of renewed interest.

Success in skill acquisition varies as a function of developmental stage, mastery of acquired-skill competencies, as well as a variety of intrinsic and extrinsic factors. Among intrinsic factors associated with underachievement, anxiety, negative self-attributions towards mathematics, and other motivational factors have been identified. Additional intrinsic or constitutional factors, such as the potential influence of heritability, remains essentially unknown. Multiple cognitive deficits have been imputed as potentially adversely impacting math-skill development. Early studies (Larsen & Hammill, 1975; McLeod & Crump, 1978) found general intellectual ability, visual perceptual and visuo-motor competencies, memory for visual sequences, verbal abilities, sequential-information processing, and comprehension and reasoning skills to be correlated with success in math performance.

Among those identified with math disabilities, Baker and Cantwell (1995) note comorbidity for reading disorders, disorders of written expression, expressive and receptive language disorders, and developmental coordination disorders. Attention-Deficit Hyperactivity Disorder (ADHD) represents the most commonly occurring Axis I diagnosis. Greater overall risk for social immaturity, school or personal adjustment problems, social skill deficits, anxiety, and depression have also been identified as risk factors expressed in this population.

Compared to the investigation of reading disabilities, the specification of integrated cognitive and neuropsychological models of math disability are lagging. Spreen and Haaf (1986) as well as Rourke and colleagues (see Rourke, 1993, for review), utilizing empirically derived clustering methodologies, have identified subtypes of mathematics impairment. There are, however, multiple sources of variability that may impinge on acquisition of mathematics competencies over the course of development. These include problems with decoding symbols; writing and copying numbers, appropriate sequencing and alignment of numbers; fact mastery; acquisition of the semantics of mathematics; memorization; capacity to convey multi-step, sequenced cognitive operations; monitoring the quality of on-going performance; higher level linguistic competencies related to both reading and segmental language processing; development of spatial processing; as well as reasoning and abstract conceptual abilities (Semrud-Clikeman & Hynd, 1992). Heterogeneity in the expression of mathematics disabilities tends to be the rule rather than the exception.

Assessment

An excellent overview of the history of mathematics assessment as well as suggestions for assessment strategies can be found in the paper by Bryant and Rivera (1997). The authors suggest that in the field of learning disabilities, norm-referenced mathematics instruments play a vital role in developing a profile of strengths and weaknesses but that they are not intended as tools for instructional planning. “Rather, other math assessment practices, such as criterion-referenced testing, curriculum-based measurement, error analysis, clinical interviews, and so forth, can be used...to develop appropriate mathematics progress and to document student progress” (p. 66). Consistent with the remarks by Bryant and Rivera, the major objective of conducting diagnostic standardized testing in mathematics abilities is to ascertain areas of strength and deficits relative to developmental level. That is, assessment of math abilities should be developmentally driven. It requires knowledge of the developmental sequences of math concept development and an understanding of the curriculum demands faced by a youngster relative to age or grade-based expectations. Informal mathematical concepts and skills are acquired via spontaneous interaction with the environment during the preschool period, for example, acquisition of concepts of “more” and “less” conservation, additive and subtractive qualities within events, rudimentary counting and enumeration (Ginsburg, 1987). Levine, Jordan, and Huttenlocher (1992) demonstrated that as early as age 4 years, math strategy utilization is driven via well-established nonverbal conceptual abilities. It is not until age 5 or 6 years that conventional number fact or story problems can be assimilated. Thus, potential dissociations between nonverbal versus verbal conceptual abilities may contribute to divergence in developmental pathways associated with mathematics-skill development by the time a youngster reaches the primary grades. With the attainment of school age, mastery of basic conventions of number facts (counting and grouping), the alphanumeric symbol code of integers, as well as number alignment and place value are established. These skills permit mastery of written calculation. With advancing
age, mastery of more complex algorithms is achieved and the curriculum includes greater emphasis on concept development, mathematical reasoning and problem solving.

As previously noted, at different age levels, differing cognitive styles may be brought to bear in problem resolution. More than one means to go about solving a particular problem may be chosen. The level of sophistication of the processes brought to bear in task resolution can in itself be diagnostic. Even though a correct answer is ultimately obtained, the strategies utilized in reasoning may be developmentally deficient, hence affecting overall efficiency in performance. The capacity for greater “automatization” and use of “formal operations” with maturity is anticipated. Lack of expression of efficient problem-solving strategies can be diagnostically important. Standardized achievement tests are helpful, therefore, in identifying both the failure to develop appropriate numerical reasoning or problem-solving strategies as well as in identifying their types. As multiple pathways can lead to the expression of developmental arithmetic problems, it is important for the assessment of mathematical abilities to be tied to the larger domain of higher cognitive functioning.

According to Fleischner (1994), a core battery of achievement tests appropriate for the assessment of math disability should provide: (a) coverage in areas of conceptual understanding, conceptual proficiency and skill applications as well as (b) a means for the integration of qualitative error analysis and clinical interview procedures. The former measures permit normative comparisons requisite for the establishment of any severe discrepancy criteria as well as pattern analysis of errors. These later strategies offer insights into errors in thinking and strategy utilization. Criticisms have been leveled against a number of standardized assessment methodologies because the preponderance of coverage is limited only to computational measures, i.e., the Wide Range Achievement Test-3 (WRAT-3. Such screening measures are insufficient for the comprehensive diagnosis and description of mathematics disabilities (Romberg, 1992).

Examples of more comprehensive assessment methodologies include the Key Math-Revised: A Diagnostic Inventory of Essential Mathematics (Connolly, 1988), the Test of Mathematical Abilities (Brown, Cronin, & McEntire, 1994), and the Test of Early Mathematics Ability (2nd ed.) (Ginsburg & Baroody, 1990). These tests offer the opportunity for multiple observations of performance across developmentally age-appropriate measures of conceptual understanding, computational skills, and applied problem solving. Utilization of a diagnostic battery further permits the profiling of a pattern of performance which can then be correlated with cognitive and neuropsychological data.

In addition, a number of informal strategies should be objectified to aid in furthering the analysis of procedural deficits or problem analysis. Clinical interview as well as “testing of limits” procedures (Ginsburg, 1987) provides additional insights into the nature of identified problems with analysis and conceptualization. Levine (1994) provides one such structured interview that can be utilized to further pinpoint a subject’s appreciation of the nature of his or her underlying math inefficiencies.

The augmentation of achievement-test data with the utilization of neuropsychological measures provides a means to correlate the nature of the manifest skills with any cognitive deficits that may contribute to learning problems. Further, such augmentation helps to define the procedural math deficits from an information-processing perspective as well as to establish a data base from which remediation or accommodations can be developed (Fletcher, Levin, & Satz, 1989).

**Disorders of Written Expression**

Disorders of written expression are identified by a demonstrable lag in the development of one or more components of writing. These could include deficits in spelling, punctuation, grammatical form and structure, or composition and organization. (Manifestations of only spelling deficits or legibility problems do not currently constitute a differential diagnosis of this disorder based on DSM-IV diagnostic criteria.) Objectification of criteria defining this form of learning disability has been challenging owing to the multi-factorial nature of writing. Intra-individual expression of a writing disorder may vary as a function of impairment of expression in any of the following areas: handwriting, spelling, language, attention and memory, written-narrative organizational skills, or metacognitive abilities.

No data are available that formally characterize the prevalence of this disorder are available. Gender influences remain ill-defined. Hooper et al.
suggest that prevalence may likely parallel the expression of a developmental language disorder. Determination of whether the deficits are primary or secondary to language or reading disabilities is integral for differential diagnosis. The occurrence of isolated disorders of written language is comparatively rare. The multi-dimensional nature of the neuropsychological underpinnings of writing suggest that deficits in any of the areas sufficient to impact writing would also impact other domains of academic performance. Thus, the evaluator must be prepared to identify a variety of potential co-morbid conditions. These include not only other learning disabilities but ADHD, depression, low academic self-esteem, anxiety, or thought disorder.

Developmentally, writing has been conceived of as the final pathway in the ontogeny of language (Johnson & Myklebust, 1967). Levine (1987) provides a heuristic model describing the progression of writing skills. Stage one begins with the establishment of basic graphomotor control (including drawing, tracing, and coloring abilities of preschoolers), attempts at proto-writing via pretend activities, and the initiation of writing training in first grade. Stage two concerns the honing of the basic orthographic skills related to letter and word formation as well as the establishment of greater graphomotor control. Stage three, associated with late second grade, is characterized by the progressive incorporation of skills such as capitalization, punctuation, syntax, and grammar (cursive writing is subsequently introduced). In the automatization stage, progressive mastery of primary competencies permits greater capacity for self-monitoring of the written product, expansion in the length of written expression, and utilization of more complex grammatical forms. In addition, planning and organizational skills begin to be incorporated into writing. In the elaboration stage (grades 7 through 9), the act of writing is sufficiently automatized in order to permit its use as a means to support the development of ideation. Greater capacity for ideational integration is expressed, and summarization skills are subsequently displayed. Capacity to form and express a viewpoint develops. In the final stage (9th grade and beyond), diversification and early development of a writing style are achieved. Writing progressively increases in versatility within such a model in order to not only augment communicative effectiveness (oral as well as written) but to support reasoning skills and creativity.

Although substantial intraindividual specification of deficits underlying writing abilities can be identified utilizing a comprehensive assessment approach, the identification of interindividual variability, or how patterns or subtypes of writing disorder manifest themselves remains more incompletely understood. The investigation of writing disability subtypes is in its infancy. One of the few comprehensive attempts is reflected in the work of Sandler et al. (1992). This factor notwithstanding, attempts have been made to formalize assessment methodologies based upon emerging empirical studies of writing disability (Berninger, 1994). Six components of writing assessment have been identified: (1) handwriting quality (legibility); (2) writing fluency (number of words copied within time constraints); (3) spelling from dictation; (4) spelling fluency within composition; (5) compositional accuracy; (6) compositional fluency (number of words produced within time constraints); and (6) compositional quality (content, cohesiveness, and organization of written narrative material).

The presence of identified hand-writing deficiencies requires assessment of motor, perceptual, and visuo-motor integrative competencies, as well as rule-out of any other pervasive developmental output failures. (Refer to specific recommendations for assessment of handwriting disorders by Bain, 1991). Skills essential in spelling include the mastery of grapheme-phoneme correspondence, overlearning the orthographic representation of word structures, and development of morphological knowledge. Strategies for assessment in this area have been covered in the reading section of this chapter (refer to phonological awareness and linguistic measures). Moats (1994) in her “Assessment of Spelling in Learning Disability Research” suggests among other things that a well-designed measure of spelling would sample “the broad domain of orthographic patterns, sound-symbol relationships, and morphophonemic patterns that must be learned by the writer of English” (p 335), while containing a wide enough range of items to accurately measure incremental development.

In addition to accrual of graphomotor samples and analysis of spelling errors, an evaluation of basic language competencies and reading skills is required. Proficiency in oral language has traditionally marked the starting point for the investigation of written language. As such, the language measures previously cited for use in assessment of linguistic underpinnings of reading disabilities
offer a means to assess semantic and general linguistic competencies (syntax) which are prerequisites for the conveyance of ideas within written form.

Developmental output failure in writing may also be a reflection of deficits in other aspects of higher cognitive functioning (Berninger, 1994). Attentional inefficiencies may be expressed as a function of input or output faculties. Deficits in the capacity to monitor quality of on-going cognitive performance impact writing demands which place a premium on simultaneous information-processing abilities. So too, executive function deficits would be expected to impact the planning and organizational skills essential in orchestrating complex ideation.

Larsen (1987) provides a review of 13 individual and group administered achievement tests based on methods of administration, test format, and coverage which are applicable to the assessment of writing abilities. The majority of these instruments provide only a cursory evaluation of writing abilities. Among traditional broad-focused academic achievement tests, the Woodcock Johnson Tests of Achievement-Revised (Woodcock & Johnson, 1989) offer the broadest coverage of conventional skills underlying writing. The Wechsler Individual Achievement Test (WIAT), while less comprehensive in assessment of core competencies related to the mechanics of writing, also offers a means to evaluate written expression within a standardized format.

Most psychoeducational screening batteries applied to the assessment of learning disabilities are limited in their sensitivity to the identification of developmental writing disorder features. When a writing disorder is suspected, a focused screening battery would consist of at minimum a spelling test (e.g., WRAT-3), the Proofing and Writing Samples subtest of the Woodcock Johnson Battery, and a sample of thematic writing, the WIAT Narrative Writing subtest. Vulnerabilities expressed on any of these measures would identify areas for more comprehensive testing. Unfortunately, the range of standardized tests available to assess written language remains restricted. No single standardized assessment tool comprehensively evaluates the heterogeneous language and cognitive deficits that characterize this disorder (Gregg, 1992). Of those systems available, many are time consuming and challenges associated with scoring can be daunting.

Two of the most frequently utilized standard written language tests are briefly summarized herein. The Test of Written Language-3 (TOWL-3) (Hammill & Larsen, 1995) was standardized for use with children from ages 7-6 through 17-11. It utilizes samplings of both spontaneous and contrived writing abilities. Eight subtests tap spelling, punctuation and capitalization, applications of semantic knowledge, syntax and grammatic cohesiveness. Factors relevant to story construction and thematic maturity are tapped. It provides an overall index of written language competency which can be contrasted against other standard scores related to intellectual ability or language mastery. The Test of Early Written Language (Hresko, Herron, & Peak, 1995) represents a downward extension of the TOWL-3. It was designed to assess emergent writing abilities. In addition to sampling linguistic skills, it taps discrimination of verbal and nonverbal representational forms as well as handwriting abilities. Like the TOWL-3, it provides a ready means for profile analysis.

Given the labor intensity of these direct assessment methodologies, an alternative to these fixed battery approaches is Berninger’s Core Battery for Writing Assessment (1994). It utilizes the Writing Samples and Dictation subtests of the Woodcock Johnson to assess handwriting, the WRAT-3 Spelling Test to assess handwriting fluency and spontaneous spelling, and a variety of hybrid measures, for which norms are available for grades one through nine, to assess compositional fluency and quality.

Acknowledgement of writing disorders as conditions worthy of neuropsychological investigation has been slow in development. Although literature is established relating acquired agraphia to aphasia or apraxia in adulthood, this literature is not generalizable to developmental disorders. Within this context, writing problems continue to be conceptualized as conditions secondary to language, motor, attention, reading, or other deficits and not as multi-dimensional disorders worthy of investigation in their own right. As yet, a comprehensive model delineating the ontogeny of writing abilities remains to be developed.
ACHIEVEMENT TESTING WITH SPECIAL POPULATIONS

Exceptional Children

Under the educational opportunity safeguards included within Section 504 of the Rehabilitation Act, P.L. 94-142 and its amendments are specific components dealing with the process of evaluation. What is mandated by law is that all students who potentially have an educational disability receive a comprehensive evaluation that fairly assesses their abilities and does not discriminate against them because of cultural or racial factors or a disabling condition. Moreover, in all areas of exceptionality, federal and state legislation require the development of individualized education plans (IEPs) for handicapped students. Educational assessment data from standardized tests provide one necessary source of information used in the development of strategies for diagnostic prescriptive teaching. Here diagnostic achievement testing plays a particularly important role not only in identifying areas in need of remediation but also in placement and classification decisions. With the importance attached to assessment in the identification, diagnosis, placement, and instruction of children with disabling conditions, it is no surprise that the use of achievement tests, particularly the use of norm-referenced measures, has come under increasing criticism (Fuchs, Fuchs, Benowitz, & Barringer, 1987; Fuchs, Fuchs, Power & Darley, 1985; LaGrow & Prochnow-LaGrow, 1982; Ysseldyke, Algozzine, Regan, & Potter, 1980; Ysseldyke & Shinn, 1981).

Fuchs et al. (1987) conducted an extensive study of the 27 most well-known and commonly used tests in special education in order to determine the degree of participation of children with handicaps in the creation of test norms, and item selection, and in the establishment of their reliability and validity. Fourteen of these tests were measures of achievement classified as either screening (battery) or diagnostic (content specific). The user manual and/or technical supplement of each test was then analyzed in terms of (a) norms, (b) item development, (c) internal and test-retest reliability, and (d) concurrent and predictive validity. In only two of the achievement measures were children with handicaps included in the norming process and on only one measure were they included in item development. Otherwise, no other information was available. Such findings led the authors to state: "[I]f, in fact, test constructors have not validated their instruments for use with handicapped people, they 'should issue cautionary statements in manuals and elsewhere regarding confidence in the interpretation' based on these tests" (p. 269. Note: The quotation in Fuchs is taken from Standard 14.2, p. 79, the Standards for Educational and Psychological Testing, 1985).

Numerous studies have analyzed the performance on standardized tests of academic achievement of students with learning disabilities (Caskey, 1986; Estes, Hallock, & Bray, 1985; McGue, Shinn, & Ysseldyke, 1982; Shinn, Algozzine, Marston, & Ysseldyke, 1982; Webster, 1985), behavioral disturbances (Altrows, Maunula, & LaLonde, 1986; Eaves & Simpson, 1984), and hearing impairments (Allen, White, & Karchmer, 1983; Karchmer, Milone, & Wolk, 1979; Trybus & Karchmer, 1977), as well as students who are gifted (Karnes, Edwards, & McCallum, 1986). The findings from these studies and others demonstrate empirically (a) the variability in test results across achievement measures; (b) particular item biases where low socio-economic status (SES) is a factor; (c) the influence of the examiner on the testing process; (d) the differential effect of diagnosis and (e) the roles of time pressure, anxiety, and sex (Doolittle, 1986; Plass & Hill, 1986). It is critical that the professionals who utilize these tests be aware of the significant validity issues involved when assessing persons with disabilities or other areas of exceptionality.

Minority Children

Cautionary comments have been made also by those persons concerned with the standardized testing of minority students. Critics of the testing movement assert that tests which purport to measure achievement, among other things, are biased against certain ethnic/racial groups. Those in favor of testing regard test misuse as the real problem. Underlying the debate is the belief by the critics that the model used to assess performance and competence in society is monocultural. "A main criticism is that the model ignores the relevance of culturally different experiences that foster other equally important competencies essential to the survival of the group or individual" (Williams, 1983, p. 192). Similarly, Green and Griffore (1980) report that in one study 46 percent of the errors made on the Gray Oral Reading Test by
minority children were due to dialect differences. Others have suggested that lack of “test-wiseness” (Millman, Bishop, & Ebel, 1965) may serve to lower the scores of minority students on tests of aptitude and achievement. Johnson (1979), commenting about the variables that may invalidate test scores for African-Americans and other minorities, wrote:

Many factors operate to attenuate or lower test scores, and these factors tend to have their greatest effects on Blacks and other minority applicants. These include factors which affect the actual performance of individuals on the test, such as socioeconomic status, differences in educational opportunity, motivation, narrowness of content of the tests, atmosphere of the testing situation, and the perceived relevance of the test to success. They also include factors that affect the test score more directly such as the composition of the group used for item tryouts and item selection and analysis which precede the actual standardization, composition of the standardization or normative group, and the techniques and procedures employed in item construction. Also, the validity or appropriateness of tests often differ for Black and white applicants, in relation to the same future performance of criterion. (p.3)

In addition, it has been substantiated that minority and white children are exposed to different curricula through the practice of ability tracking (Coleman, 1966; Findley, 1974; Green & Griffore, 1980; McPartland, 1969). Reviewers of the hundreds of ability grouping studies conducted since the 1920s have concluded that while superior students may benefit from this method of curricular offering, students with lower class ranking may not. The primary areas of concern are exposure to undemanding curricula and the social stigma attached to students in low-ability groups.

In a study by Abadzi (1985), the effects on both academic achievement and self-esteem of students placed in ability grouping classrooms were investigated with a population of 767 students from grades 4 to 8 in a large Texas school district. Contrary to earlier studies, her findings were that high-ability students did not maintain in the long run the performance gains made in the first year of grouping. Only the lower-level high-ability students in grouped classes were to benefit from the educational and social opportunity provided the highest-ability students. Students near the cutoff score in all groups were the ones most influenced by grouping in terms of both achievement and self-concept. Support for these findings was provided in spite of a general downward trend in performance at the end of elementary school that was characteristic of the school district’s test scores and those of other districts as well. The author hypothesized that the steady drop in scores with the high-ability students may have been the result of reduced achievement motivation brought on by a “sense of invincibility, which the high status of the program combined with nonexistent exit criteria helped reinforce” (Abadzi, 1985, p. 39).

The concept of achievement motivation raised in Abadzi’s conclusions has been systematically studied since the publication of David McClelland’s The Achievement Motive (1953). This concept has been defined as a learned motive, unconscious in nature, resulting from reward or punishment for specific behavior. While studies utilizing this definition of achievement motivation have been conducted across racial groups, they have been criticized because of their ethnocentric design, methodology, and instrumentation. Castenell (1984) suggests that future research incorporate the definition espoused by Katz (1969) and Maehr (1974) which posits that (a) achievement motivation is conscious, (b) the need to achieve is universal to all groups, but (c) “because different groups have different life experiences it is likely that situations or a set of tasks will evoke different group responses” (p. 442).

While concerns have been raised with respect to standardized testing with minority students in general, we have not addressed the issues involved in standardized achievement testing with language minority students. In this regard we defer to a thorough discussion of this topic by Lam (1993) in which he suggests guidelines to consider in exempting limited English proficiency (LEP) students from standardized achievement testing and for the development of special testing for these students.

This section on special populations concludes with guidelines set forth by Williams (1983) that are highly reminiscent of the recommendations put forth in 1975 and cited at the beginning of this chapter. They would appear to encompass concerns regarding the use of achievement tests regardless of students’ race, color, national origin, or handicap.

- Test constructors should foster an awareness of the limitations of the tests and the meaning attributed to test scores.
• Test constructors should educate their consumers in selecting tests in terms of particular goals and objectives of educational evaluation.

• Test constructors should bear responsibility for including minorities in all aspects of test development and not limit this to the standardization sample.

• Test consumers must assume some responsibility for developing skills in administering tests and interpreting results in light of the culturally diverse experiences that pupils bring into the testing situation.

• The educational community should minimize or eliminate intelligence testing or substitute approximately modified assessment techniques and interpretive procedures that consider cultural differences.

• The educational community should focus on achievement rather than intelligence or aptitude testing to eliminate pernicious connotations and unfair placement practices that limit future educational attainment and opportunity (p. 205).

THE FUTURE OF ACHIEVEMENT TESTS

Computer Adaptive Testing

The final section of this chapter is a discussion of the growth and impact of computerized adaptive testing on the measurement of achievement and what this product of modern technology means to the field of measurement. This is a fitting topic to conclude the previous narrative because computer adaptive testing is the direct result of advances in the fields of psychometrics, mathematics, cognitive learning theory, educational measurement, human engineering, and science technology. It relies as heavily on Glaser’s criterion-referenced measurement as it does on Ausubel’s cognitive approach to learning, Deno’s curriculum-based measurement, Messick’s concern with test validity, and Anastasi’s continuum of testing.

Overall, educational research and development is most currently preoccupied with enhancing the instructional value of tests, or as Haney (1985) describes it, “making testing more educational” (p.4). He states that one need not be a dyed-in-the-wool social Darwinist to recognize that the use of standardized testing is increasing because it serves some important social functions. However, certain deficits that currently exist tend to negate the value of these tests: (a) Most testing programs violate the one nearly universal desideratum in all learning theories—in order to learn, an individual needs to receive rapid and specific feedback. (b) Most standardized tests have a very uncertain relationship to the specific teaching and learning that occurs in particular schools and classrooms. (c) The frequent concern to keep standardized testing programs secure limits their educational utility. It is these deficits, both narrowly and broadly defined, that the process of adaptive testing can be seen to rectify.

Adaptive testing is based on the premise that a measurement continuum should parallel a learning/teaching continuum, and if this learning continuum could be adequately measured by an underlying scale extending through its entire range, a student could enter and exit the measurement continuum at points appropriate to his or her current development regardless of age or grade levels (Forbes, 1986). This test development system is based on a measurement model popularly named the Rasch model after its originator. This model is also referred to as a one-parameter model in contrast to three-parameter models of latent traits which are based not only on item difficulty (single parameter) but also on item discrimination (slope of the difficulty) and on the level of change performance (guessing).

All item-response theory models must have an item data bank from which test items are drawn in the process of test construction. These items are computer stored and are then retrieved following a logical format. Utilizing a computer, the test can be presented to the student on a video screen with the computer keyboard serving as the response mechanism. Under such a procedure, the computer represents one pre-constructed test selected from a group of such tests. The test is tailored so that the computer “jumps” the person to the appropriate item-difficulty range and then gives a preselected sequence of items based on the correctness or incorrectness of the previous response. Generally, fewer items are required to measure performance at a predetermined level of measurement error than is the case with traditional testing procedures. Computerized adaptive tests have been shown also to take less than half the testing time required by traditional achievement tests and to provide more precise ability estimates across the entire ability range. Because the ability estimates and the item parameters are calibrated on a common scale, these estimates are theoretically
independent of the particular sample of persons taking the test and the particular sample of items selected by each examiner.

Seminal work done by Weiss (1980) focused on applying computerized adaptive testing to the measurement of achievement, using a methodology to extend beyond the aptitude measurement to which this type of testing had been limited previously. In addition to extending the use of item-characteristic curve theory (ICC) methods from ability testing to the problems of achievement testing, the project was also concerned with developing solutions to unique problems raised in achievement testing, that is, assessment in multiple content areas, mastery testing, the issue of stability of measurement over time, and the effects of immediate feedback as to the correctness or incorrectness of test responses. The findings of this three-year research project supported the use of ICC theory and methods and computerized adaptive testing for the measurement of achievement. However, many new questions were raised in addition to those originally addressed by the research that were in need of further study.

One of the first studies to compare and equate achievement scores from three alternative methods of testing, paper-administration, computer-administration, and computerized adaptive testing, was conducted by Olson (1986) with all students in grades 3 to 6 within three California school districts. A total of 575 students were involved in the study. Results of the study indicated that (a) analysis of variance showed no significant differences among the three measures in terms of the comparability of measurement precision; (b) computerized adaptive testing (CAT) required only one fourth of the testing time required by the paper-administered test; (c) the computerized adaptive test provided a more precise ability estimate with smaller variance than either of the other two measures; and (d) the ability estimates calculated from a 20-item CAT tended to show more precision than tests of 55 to 62 items used with the other two measures.

Since that time, work has been done to investigate an innovative application of item-response theory (IRT) in computerized testing known as self-adapted testing (SAT). With this model, the difficulty levels of items administered are chosen by the student rather than by a computer algorithm, with positive results (Rocklin & O'Donnell, 1987). Rocklin and O'Donnell (1991) later reported that anxiety influenced student performance less on the SAT than on the CAT. Their results were later substantiated by Wise, Plake, Johnson, and Roos (1992) and Roos, Plake, and Wise (1992). Wise, Kingsbury, and Houser (1993) then experimented with a restricted form of SAT to provide students with control over the testing situation while preventing large mismatches between item-difficulty choice and proficiency level, which had shown itself to be a factor with a limited number of students in previous studies. At this point, use of the RSAT procedure is yet to be empirically evaluated.

This section on adaptive testing concludes with the futuristic predictions raised by Hsu and Sadock (1985) in their review, Computer-assisted Test Construction: The State of the Art. The authors foresaw the following as commonplace in testing of the future:

1. The development of item construction theories that take advantage of artificial intelligence and the phrase recognizability of the computer.
2. The development of item banks in the area of criterion-referenced achievement tests and in conjunction with textbook publication.
3. Item calibration and test design available on microcomputer.
4. The regular use of computers in test administration.
5. The application of IRT in test design by non-measurement specialists.
6. The use of computerized adaptive and diagnostic testing in the classroom.

Writing about achievement tests in the 1984 edition of the Handbook of Psychological Assessment, Fox and Zerkin concluded: "[While] standardized tests are not perfect and can be misused and misunderstood...they are currently the best instruments educators have available for assessing the quality of curriculum and for individualizing and improving instructional programs for each child" (p. 130). These conclusions no longer hold.

It is no longer possible to call these standardized measures of achievement the "best" instruments available. With the 1970s, criterion-referenced tests were touted as useful alternatives to norm-referenced tests. In the 1980s the new fields of cognitive sciences and computer technology were cited as likely sources for new and better test development. And now, in the late 1990s, performance assessment is popularly viewed as a remedy to the past ills of standardized tests. While numerous social problems are associated currently with the
use of more traditional testing procedures, and in particular multiple-choice tests, Haney, Madaus, and Lyons (1993) suggest that the negatives associated with their use may have more to do with the myriad functions that standardized tests are expected to perform. “To the extent that we regain more balanced approaches to assessment, reflecting a wider range of the modes by which we ought to judge student learning...to that extent will the distortions now associated with standardized tests be reduced” (p. 294).

It is hoped that the present discourse has led the reader to question practices of the present because of knowledge of the past and to look to the future with eager anticipation. Tests can be a flexible passport into that future or a rigid barrier bound to the past. It is our job as professional educators, in the broadest sense, to insure the former. When describing the failure of the testing profession to inform the public about the meaning of “objective” standardized tests, Strenio (1981) states: “At a minimum, testers have an obligation to avoid placing their particular jargon in any context that makes it even harder for the layman to interpret than it already is” (p. 65). The authors of this chapter hope that they have not been guilty of this same failing.

"Then you should say what you mean," the March Hare went on.

"I do," Alice hastily replied; "at least—at least I mean what I say—that’s the same thing, you know."

"Not the same thing a bit!" said the Hatter; "why, you might just as well say that ‘I see what I eat’ is the same thing as ‘I eat what I see!’".

—Lewis Carroll

Alicia’s Adventures in Wonderland

REFERENCES


that dyslexia may represent the tower tall of a normal distribution of reading ability. New England Journal of Medicine, 326, 145–150.


Aptitude assessment and intervention have a long and distinguished role in the clinical evaluation of children and youth with learning and behavioral problems. This chapter will review the major approaches to assessment of aptitudes with emphasis on ways the assessment information is used in decisions about diagnosis, placement, and treatment. The chapter content does not include information on aptitude assessment with adults, nor the uses of aptitude information in career counseling or vocational guidance.

A critical theme in our review is treatment validity, that is, is treatment effective (or more effective) when based on a conception of aptitude and guided by the associated assessment procedures? This criterion requires consideration of models, conceptions, assessment procedures, treatment approaches, and intended outcomes. The latter is especially central in this review. Intended outcomes, or criteria for validity, must be an integral part of the evaluation of different aptitude models.

Several traditional models, although intellectually attractive and logically persuasive, fall short on the treatment-validity criterion. Their use in educational or clinical situations must be regarded as questionable if benefits to children and youth cannot be demonstrated. Several models will be described and critically reviewed. Major attention will be devoted to mediated learning, a model with enormous potential that has been examined by scholars in several western nations.

Chapter 8

EVALUATION OF APTITUDES

Daniel J. Reschly
Carol Robinson-Zañartu

CONCEPTS OF APTITUDE, INTELLIGENCE, AND ACHIEVEMENT

Aptitude, intelligence, and achievement as psychological constructs or types of tests are not easily distinguished. The traditional distinction was that achievement tests reflected the effects of past learning, whereas aptitude and intelligence reflected the individual’s potential for success. In this traditional view, both aptitude and intelligence were seen as relatively enduring traits of the individual, not easily modified by experience or special training. In some instances both aptitude and intelligence-tests results were regarded as indications of innate capacity.

These traditional meanings of aptitude, intelligence, and achievement tests were rejected in all the leading measurement texts published in the last decade (Anastasi, 1997; Brown, 1983; Cronbach, 1990). All now are viewed as tests of developed abilities, that is, they reflect the effects of experience, and, as maximum performance measures, it is assumed that the individual is encouraged to try as hard as possible to do well.

The most important differences among aptitude, intelligence, and achievement have to do with how they are used and with assumptions about antecedent experiences (Anastasi, 1997; Brown, 1983). Achievement tests are assumed to measure past learning that occurred in a specific teaching or instructional situation. In contrast, aptitude has a future reference. The aptitude concept involves inferences about performance in future learning or
training situations. Intelligence is usually seen as between achievement and aptitude on the continua of test use and antecedent experiences. Intelligence has a present reference as a reflection of the effects of general, broad learning experiences. When intelligence tests are used in diagnoses, a future reference is assumed because predictions typically are made about the continuing status of the individual.

As a construct, aptitude often is used quite broadly, especially in theory and research on aptitude by treatment interactions (Cronbach & Snow, 1977; Snow, 1980, 1992). Here, aptitude is virtually any psychological characteristic of the person that predicts differences among people in later learning or training situations. Included in this very broad conception of aptitude are cognitive abilities and processes and personality and emotional characteristics (Snow, 1992). Although this broad conception of aptitude is used in this chapter, more of the content will come from examination of cognitive abilities and processes than from emotional or personality characteristics.

COMMON FEATURES OF APTITUDE MODELS AND ASSESSMENT

The varying aptitude models used with children and youth have a number of common features. Perhaps the most basic commonality is the goal of improving diagnosis, placement, and treatment decisions with children and youth who exhibit varying degrees and kinds of learning problems.

Expanded Consideration of Cognitive Processes

Proponents of different aptitude models share a commitment to assessing cognitive processes that are not directly represented on conventional measures of intelligence and achievement. Numerous observers have noted the rather narrow range represented or the limited opportunity to observe distinct cognitive processes on traditional measures (e.g., Naglieri, 1989; Woodcock & Mather, 1989). Widely used measures of intellectual functioning such as the Wechsler Scales (Wechsler, 1974, 1991) reflect a rather narrow range of cognitive processes. Further, the Wechsler items involve a wide variety of complex tasks that require the simultaneous use of two or more cognitive processes rendering difficult the observation of distinct processes.

Some more recently developed measures, such as the Kaufman Assessment Battery for Children (K-ABC) (Kaufman & Kaufman, 1983), provide a slightly broadened array of cognitive processes and better opportunities to observe specific processes. The K-ABC and similar instruments, however, remain primarily as measures of general intelligence with limited specific information on processes that are drawn from a single model of aptitude. The recently published Cognitive Assessment System (Naglieri & Das, 1997) likewise reflects a single cognitive model and a limited array of processes.

Intra-individual versus Inter-individual Differences

The aptitude models and measures typically focus more heavily on intra than inter-individual differences. Intra-individual differences involve variations in the pattern of cognitive processes within the individual. For example, the individual’s mean on several measures often is used as the point of comparison in assessing strengths and weaknesses in cognitive processes. Substantial variations from this mean then may be translated into goals for cognitive training or recommendations for appropriate instructional methodology. The inter-individual interpretation method typically used with conventional measures of intelligence and achievement examines differences between individuals. Here, the individual’s variation from a population mean is the primary reference point for interpreting performance. Inter-individual approaches to interpretation are useful for determining level of performance in comparison to others while intra-individual approaches yield information on pattern of performance within the individual.

Improved School Achievement

Aptitude assessment and intervention models share the common goal of improving academic achievement. Indeed, aptitude assessment and intervention usually is initiated because an individual student is having difficulty with acquiring basic literacy skills in reading, writing, or mathematics. Although different aptitude models use
Quite different assessment and treatment procedures (see next section), an overall goal is improved academic functioning.

**Improved Overall Cognitive Functioning**

The aptitude models have the common goal of improved overall cognitive functioning, although it should be noted that there are quite different assumptions about how best to produce improved general functioning (see next section). Treatment, based on assessment of aptitudes, is thought to lead to improved functioning either through direct changes in cognitive processing or through the contribution to cognitive functioning that occurs with improved acquisition of academic skills. Rapidly changing technology leading to the dawn of the “information age” throughout the world has made advanced thinking processes, cognitive flexibility, and rapid acquisition of new skills imperative goals for all children and youth in modern societies. Aptitude assessment and treatment attempts to address these needs as well, through direct or indirect procedures for improving overall cognitive functioning.

**DIFFERENCES AMONG APTITUDE MODELS**

Aptitude models vary significantly regarding assessment procedures, treatment techniques, and intended outcomes. These variations involve fundamental elements, including (a) assumptions about stability or modifiability; (b) product versus process of cognitive activities; (c) transfer of effects; (d) standardized versus dynamic assessment procedures; and (e) applications to decisions.

**Stability versus Change in Aptitudes**

The most basic difference is the assumption about whether aptitudes are relatively stable traits of the individual or whether aptitudes can be changed through treatment. In the former view, aptitudes are important individual characteristics that form the basis for relatively enduring recommendations for program placement or instructional methodology. Since the cognitive or neurological processes that underlie the aptitudes are believed not to change, better cognitive and academic performance is sought through matching the individual’s permanent aptitude characteristics to curricula or methodology that capitalize on intact processes or realistic goals. If aptitudes are viewed as unchanging, then treatments to change aptitudes (remediate deficits) is regarded as futile (Reynolds, 1981, 1986, 1992).

In contrast, those who contend that aptitudes can be changed through treatment use initial assessment results as the starting point for the design of experiences that will modify the individual’s basic ways of thinking. Here, the goal is nothing less than the modification of thinking and, thereby, the development of aptitudes previously not observed. Expanding the cognitive repertoire is seen both as possible and essential. It is difficult to think of a more basic difference among models than whether aptitudes are changeable. This fundamental point leads to numerous additional differences.

**Product versus Process Orientation**

The primary observational unit varies among aptitude models. Product-oriented models focus on whether or not the individual can correctly perform certain tasks that are assumed to be reflections of underlying aptitudes. For example, can the individual correctly solve problems that make primary demands on visual-spatial processing or presumed right cerebral cortex functions? In contrast, aptitude models that are more process oriented attempt to examine underlying cognitive processes or thinking skills that the individual uses to achieve right or wrong answers to tasks. In process orientation the answer itself often is less important than the thinking skills that produced the answer. The process-product distinction is related both to the stability-modifiability assumption and the questions of transfer of training, kind of assessment, and intended outcomes.

**Transfer Questions**

The transfer question involves the issue of how the aptitude information is used and the effects of its use. In models that assume stability of aptitudes with a focus on products of cognitive activity, the transfer question is: do aptitude strengths translate to more successful academic learning through matching aptitude strengths to instructional methodology, or through designing curricula to match...
aptitude levels and patterns? It is assumed that the aptitudes that are assessed do underlie instruction or control responses to different curricula. If these aptitudes do not transfer from assessment to school-learning tasks, the aptitude assessment as well as the instructional and curricular recommendations lack treatment validity.

The transfer question is different for the models that assume that aptitudes are modifiable. In these models cognitive processes observed in the individual's efforts to solve task-specific problems then become goals for cognitive modification treatments. Near transfer is demonstrated subsequent to treatment if the individual can perform more successfully new examples of the same or highly similar tasks. Such positive change suggests that cognitive processes have been modified, at least within a specific kind of task.

The cognitive modification is, however, limited unless far transfer can be demonstrated. Far transfer is the question of whether cognitive modifications in the individual lead to more successful performance in different problem-solving situations that involve new tasks or stimulus properties. Ultimately, most cognitive modifiability models seek better performance in generalized problem solving and learning, including the acquisition of academic skills. For obvious reasons, it is considerably easier to produce near than far transfer.

Standardized versus Dynamic Assessment

The rules that are established for valid and reliable assessment differ substantially among aptitude models. In modifiability models, the examiner engages in a complex interaction with the student wherein the purpose is to establish how the individual thinks and the strategies that individual uses to approach new learning, solve problems, and communicate ideas and findings. Accomplishing this purpose necessarily requires a format that can be at best partially structured. The assessment is "dynamic" in that what the examiner does is determined by the kinds of thinking processes used by the student, interacting with the student to test hypotheses regarding the reasons for correct and incorrect problem solutions, and using near-transfer tasks to further refine hypotheses and explore effective interventions. Although modifiability clinicians may use a limited array of problem solving tasks, the presentation of the tasks and the interaction with the student nearly always varies from case to case.

Models that assume aptitude stability typically use conventional standardized procedures; that is, they attempt to achieve uniformity in task presentation, evaluating responses, summing scores, and interpretation of score meaning. The relationship of clinician and student is assumed to be standardized as well, and similar to the relationship that is used in individual assessment of intelligence or achievement.

Application to Decisions

The major aptitude models are used primarily in cases where the child or adolescent is not acquiring academic skills to the degree expected. Assessment of aptitudes leads to decisions about these students, but the kinds of decisions vary significantly by model. Application of the stability models typically leads to decisions about the kind of instruction that should be used, most often in reading. Clinicians working with teachers attempt to match aptitude strengths to instructional methodology, assuming that the best match will lead to more efficient learning. In other instances, aptitude strengths and weaknesses may be used as part of classification and placement decisions, for example, to diagnose low achievement as stemming from a learning disability or dyslexia and, then, based on this diagnosis, placing the student in some kind of remedial or special education program.

Modifiability models are less likely to be used in classification and placement decisions, or in the prescription of specific curricula. Decisions subsequent to application of assessment are more likely to involve prescriptions for particular kinds of training in thinking processes and problem-solving strategies and their application to teaching methodology. That methodology embeds cognitive processes in the teaching of academic skills. The assessment, however, continues as an integral part of the cognitive training. Teaching new thinking processes or strategies occurs simultaneously with ongoing assessment of problem-solving competencies such that the teaching-assessment process is continuous and inseparable.
MODELS OF APTITUDE ASSESSMENT AND INTERVENTION

Three models of aptitude assessment and intervention are reviewed in this section. Each of the models is described using the common features and differences discussed in the previous section, after which the criterion of treatment validity is applied to each.

Psycholinguistic and Perceptual-Motor Models

Psycholinguistic (PL) and perceptual-motor (PM) assessment and intervention are the oldest, best researched, and most controversial of the aptitude models used currently with children and youth. The models make similar assumptions about cognitive processes, interventions, and anticipated effects on academic achievement. Due to these similarities, both are discussed in this section.

The primary PL model (Kirk, McCarthy, & Kirk, 1968), based on a communication theory (Osgood, 1957), had three major components: (a) channels of communication (auditory-vocal and visual-motor); (b) communication process (reception, association, and expression); and (c) levels of organization (representational and automatic-sequential). Kirk and colleagues developed the Illinois Test of Psycholinguistic Abilities (ITPA) to assess these components of language processes and several volumes were published to guide intervention efforts (Bush & Giles, 1977; Kirk & Kirk, 1971; Minskoff, Wiseman, & Minskoff, 1972).

The ITPA subtests and the associated intervention procedures attempted to address the following cognitive processes: auditory reception, visual reception, auditory association, visual association, verbal expression, manual expression, grammatic closure, visual closure, auditory sequential memory, visual sequential memory, auditory closure, and sound blending. Adequate functioning on these processes was assumed to be required for acquisition of literacy skills in reading, writing and, to a lesser extent, mathematics.

The perceptual-motor assessment and intervention model emphasized processes such as visual and auditory discrimination and perception, visual-motor coordination and integration, visual-auditory integration, and motor skills. Some of the models claimed direct relationships between neuro-logical functioning and motor-perceptual awareness, leading to interventions such as walking on balance beams, vestibular stimulation from movement of the entire body in space, and precise large-motor exercises. Other PM-model variations placed more emphasis on pencil-and-paper tasks designed to improve visual-motor skills or auditory exercises to improve discrimination and recognition of sounds. All variations of PM assumed relationships between these skills and academic achievement.

The PL and PM models are used to identify intra-individual differences in processes presumed to underlie overall cognitive functioning and school achievement. The models generally are used with younger children (aged 2 to 10 years) who have been identified as delayed in cognitive development, or as experiencing difficulty in acquiring beginning literacy skills, especially reading. Intervention is usually guided by careful, standardized assessment of perceptual-motor or psycholinguistic strengths and weaknesses, followed by specific teaching activities designed to overcome weaknesses.

Clearly, the PL and PM models assume that basic cognitive processes could be identified accurately and improved through systematic instruction. Transfer to improved school achievement is assumed in a chain of logic that proceeds through the following assumptions: (a) PL or PM processes underlie and are a prerequisite to successful school learning; and (b) untreated PL- or PM-process deficits would remain as barriers to, and improved PL or PM processes will lead to, more successful achievement. Many educational programs for young children with learning problems continue to make these assumptions and emphasize PL training.

In addition to remedial programming for young children, the PL and PM models are highly influential in the diagnosis of specific learning disabilities (SLD). The most widely used SLD conceptual definition contains the following language, “a disorder in one or more of the basic psychological processes involved in understanding or using language, written or spoken” (Mercer, King-Sears, & Mercer, 1990; Reschly & Gresham, 1989). This conceptual definition appears in U.S. law and is adopted in many states. Although, the classification criteria described in federal and state law typically do not require identification of PL- or PM-processing deficits as part of the diagnosis of SLD (Mercer et al., 1990), the PL and PM models con-
continue to be highly influential regarding thought about the causes of SLD.

The PL and PM models dominated thought and practice in SLD until about 1980. Reviews of research on the assessment of PL processes and the outcomes of PL interventions began to appear in the mid-1970s, leading to diminished use. The Illinois Test of Psycholinguistic Abilities and various tests of PM processes were severely criticized on psychometric criteria, especially the low reliabilities on subtests that were used to diagnose weaknesses and prescribe interventions (Salvia & Ysseldyke, 1988). If the subtest reliabilities were low, then intra-individual strengths-and-weaknesses results were, by necessity, inaccurate and the prescriptions for interventions were based on faulty information. Even more devastating, though more controversial, were a number of separate reviews of the effects of PL and PM interventions. Hammill and Larsen (1974, 1978) and Newcomer, Larsen, and Hammill (1975), concluded that PL interventions had little positive effects on improving PL processes, and no documented positive effects on school achievement. The conclusions about the effects of PL interventions were disputed by Minskoff (1975), and later by Lund, Foster, and McCali-Perez (1978), in a series of increasingly heated debates with Hammill and his associates. Later examination of the same body of literature using the technique of meta-analysis (Glass, 1983) led to slightly more positive conclusions about interventions with some PL processes (verbal expressions, manual expression, visual closure, and auditory association) (Kavale, 1981, 1990), but no solid evidence has been provided confirming that improved PL processes lead to improved academic achievement.

The treatment-validity evidence regarding the PM model is even more negative. Kavale and Mattson's (1983) meta-analysis of some 180 studies led to the disappointing conclusion that PM interventions had no effect on PM processes and no identifiable beneficial effects on school achievement. PM assessment and training appears to be a waste of teachers’ and students’ valuable time (Kavale, 1990).

The PM and PL models have strong intuitive appeal. Most of the cognitive processes identified in these models are logically related to school achievement and overall cognitive functioning. Several possible explanations exist for the failure to unequivocally establish gains in either the processes or school achievement from PL and PM interventions. First, the theory simply may be wrong; that is, the PL and PM processes may not be essential for overall cognitive functioning and school achievement. If so, it would not be the first time that an intuitively appealing idea was incorrect. Second, the essential PL and PM processes may not be assessed accurately by current measures. Several authors have noted the psychometric deficiencies of measures in these areas. Third, the interventions may not be sufficiently powerful to produce the effects that would be required to produce PL or PM gains and improved school achievement. Regardless of which explanation(s) is/are correct, continued use of PL and PM models for assessment and intervention in clinical or educational settings is questionable.

Aptitude by Treatment-Interaction Models

Three prominent treatment-by-aptitude-interaction (ATI) models are used by many clinicians and educators today. All specify relationships between cognitive processes and the methodology used to teach cognitive and academic skills.

The modality-matching model generally focuses on three kinds of information-processing strengths and weaknesses: (a) auditory, (b) visual, or (c) kinesthetic. Children experiencing difficulty in acquiring academic skills are assessed to determine strengths and weaknesses over these processes, and then an instructional method that utilizes the child’s strengths is prescribed. The same procedures are used in the second and third models, cognitive style and neuropsychological. In the latter model additional inference(s) is/are made about underlying brain functioning. The actual prescriptions for instructional methodology are highly similar across the three ATI models (Reschly & Gresham, 1989).

The ATI models see aptitudes as relatively unchanging, even to the point in the neuropsychological variation of cautioning against “teaching dead tissue” (Hartlage & Reynolds, 1981). Focusing interventions on deficit areas, as is the case in the PL and PM models, is seen as inefficient and perhaps futile. Aptitudes typically are assessed with conventional standardized measures of cognitive functions including individually administered measures of general intellectual functioning. Two-well standardized general intelligence measures have been developed with primary attention to diagnos-
ing elements of cognitive style (Naglieri & Das, 1997; Kaufman & Kaufman, 1983). The ATI models are used primarily to design instruction with an underlying assumption that matching process strengths will generalize to higher learner performance in instructional settings. Neuropsychological concepts also have appeared in 1980s definitions of SLD. The phrases “presumed to be due to central nervous system dysfunction” and “presumed neurological origin” appeared recently in two separate definitions formulated by learning-disabilities advocacy groups (Reschly & Gresham, 1989). These phrases and neuropsychological diagnostic criteria have not appeared in federal or state SLD definitions or diagnostic criteria (Mercer et al., 1990).

Modality matching, cognitive style determination, and neuropsychological assessment and intervention depend heavily on the existence of aptitude by treatment interactions (ATI). The ATI notion has enormous intuitive appeal. Reynolds (1992) described this process as, “Instruction... formatted around the child’s best developed processes, avoiding those that are poorly developed or inept” (p. 10). According to a survey by Arter and Jenkins (1977), 99 percent of teachers believe that there are differences in how children process information, and that instruction will be more effective if instructional materials and methods are matched to modality or neuropsychological strengths.

The case for ATI was stated by Cronbach (1957) in a highly influential and widely-cited article that appeared in the American Psychologist, the largest circulating psychology journal in the world. Cronbach asserted that aptitudes could be measured accurately and that, “For any potential problem, there is some best group of treatments to use and best allocation of persons to treatments” (p. 680). The allocation process meant matching individuals' aptitude strengths to treatments that utilize those strengths through differential stimulus properties or instructional methods.

Although the potential list of aptitudes that might be used in matching is nearly unlimited, the different instructional methodologies are much more limited. The typical matching procedure involves aptitudes such as auditory, visual, and kinesthetic processes, cognitive styles such as simultaneous and successive (Das, Kirby, & Jarman, 1979) or sequential and simultaneous (Kaufman, Goldsmith, & Kaufman, 1984), or neuropsychological constructs such as right hemisphere and left hemisphere functioning. The different instructional methodologies prescribed for children with these aptitude strengths are highly similar (Reschly & Gresham, 1989). Phonic methods

**Table 8.1. Aptitude by Treatment Interaction**

<table>
<thead>
<tr>
<th>Instructional Method</th>
<th>Aptitude</th>
<th>Match Method to Strength</th>
<th>Presumed</th>
<th>Maximum Benefit</th>
<th>Actual-No Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonic</td>
<td>Left Hemisphere</td>
<td>Auditory-Vocal</td>
<td>Successive</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Right Hemisphere</td>
<td>Visual-Motor</td>
<td>Simultaneous</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Match Method to Strength</td>
<td>Mismatch</td>
<td>Presumed</td>
<td>Minimal Benefit</td>
<td>Actual-No Effect</td>
</tr>
<tr>
<td></td>
<td>Presumed</td>
<td></td>
<td>Presumed</td>
<td>Minimal Effect</td>
<td>Actual-No Effect</td>
</tr>
<tr>
<td></td>
<td>Maximum Benefit</td>
<td></td>
<td>Maximum</td>
<td>Benefit</td>
<td>Actual-No Effect</td>
</tr>
<tr>
<td></td>
<td>Actual-No Effect</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sight</td>
<td>Mismatch</td>
<td>Match Method to Strength</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Presumed</td>
<td></td>
<td>Presumed</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Minimal</td>
<td></td>
<td>Maximum</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Effect</td>
<td></td>
<td>Benefit</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Actual-No Effect</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:** Assumption: Matching aptitude with treatment (instruction) produces maximum benefits. Empirical basis: Weak for underlying process or neuropsychological strengths and weaknesses.
of teaching reading and overall emphasis on auditory cues typically are prescribed for children believed to have strengths in auditory processing, sequential or successive cognitive styles, or left hemisphere functions. Similarly, whole-word methods of teaching reading via visual cues are stressed for children with strengths in visual processing, simultaneous cognitive styles, or right hemisphere functioning.

For these models to have treatment validity, there must be an interaction between the presumed aptitude strength and instructional methodology. For example, children with right hemisphere strengths must learn more efficiently when instructional materials and methods are selected and presented in ways to utilize that strength, and conversely, such students must do less well if instructional methodology is not matched to strengths (see Table 8.1).

Unfortunately, the research-to-date does not support the existence of significant treatment by aptitude interactions. Some of the difficulties with ATI research were summarized by Cronbach (1975) in another American Psychologist article in which he expressed doubt about ever being able to use matching in clinical or educational settings. Based on 18 years of largely unsuccessful ATI research, Cronbach concluded that, "Once we attend to interactions, we enter a hall of mirrors that extends to infinity" (p. 119). The major problems in the ATI research were: (a) non-existent or very weak interactions; that is, matching strengths had, at best, small and inconsequential effects; (b) results over studies were enormously inconsistent; and (c) there were higher order interactions, that is, complex three- and four-way interactions, that would be impossible to apply to practical clinical and educational problems.

The current research on modality matching and neuropsychological assessment and prescriptions fits the pattern described by Cronbach (1975). Matching presumed strengths with instructional methodologies does not lead to demonstrable differential gains in academic achievement, regardless of whether the aptitude strengths are conceptualized as modality preferences (Kavale & Forness, 1987, 1990; Kavale, 1990), cognitive styles (Ayers & Cooley, 1986; Ayers, Cooley, & Severson, 1988), or neuropsychological functions (Reschly & Gresham, 1989; Teeter, 1987, 1989). Despite the negative evidence, the modality matching, cognitive style, and neuropsychological-functioning approaches to assessment and intervention continue to be used widely in a variety of settings by psychologists and educators.

The hall of mirrors that Cronbach described in 1975 continues to confound efforts to apply an inherently sensible idea, that is, selecting and implementing instructional methodology that utilizes an individual's cognitive-processing strengths. Reasons similar to those that may account for the negative-treatment validity evidence on the PL and PM models are relevant to the ATI models. The problem(s) may reside with the basic theory, the measures of aptitudes, or the presently available interventions. Reschly and Gresham (1989) noted deficiencies in all of these areas. The models often are predicated on rather primitive theories of neurological functioning or information processing. The determination of strengths usually involves intense analyses of profiles of scores on different subtests or measures from different standardized batteries. The profile-difference scores that are fundamental to determination of strengths and weaknesses typically have low reliabilities and other psychometric deficiencies (Macmann & Barnett, 1994a, 1994b; McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992; McDermott, Fantuzzo, & Glutting, 1992). Perhaps most important, the interventions are limited and not very powerful. Regardless of which explanation(s) is/are correct, and whether any of these deficiencies can be overcome, current use of any of the ATI models in diagnosis and treatment of learning problems is highly questionable.

Dynamic Assessment/Change Models

Dynamic assessment and mediated learning intervention models presume that learning abilities, or aptitudes, are modifiable rather than stable. Specifically, within this paradigm, what had been previously conceived of as largely genetically determined and stable within the human organism is believed to be open to change with human intervention. Through specific mediations, actual modification of cognitive structures and motivational factors are believed to influence the manner in which the individual is able to approach new situations; thus, the acquisition of knowledge. The function of assessment in this model is to identify efficient and inefficient cognitive and motivational parameters of the individual; the parameters accessible to mediation; and the kind and
intensity of mediated interventions needed to produce change in accessible parameters. Determining accessibility of cognitive parameters and the measurement of their changes occur as mediated intervention is provided. In contrast to conventional forms of assessment that typically produce classifications, labels, and predictions based on the belief in stable individual differences, dynamic assessments are designed to identify the targets and methods of intervention for enhancement of individual functioning across classroom, home, and community settings.

The notion of testing the ability to learn while observing learning-in-progress emerged near the early part of the century (Dearborn 1921; De Weerdt, 1927; Penrose, 1934). The idea that observing the results of deliberate stimulation of learning would yield important data and information for actually developing aptitude has its early roots in the work of Vygotsky (1934/1962) and Rey (1934). Vygotsky’s (1962, 1978) cultural-historical theory of human mental development, the genetic epistemology of Piaget (1952), Luria’s (1966a, 1966b) neuropsychological investigations of brain-behavior relationships, Schwartz’s (1977, 1983) psychosocial-neuropsychological model of self-regulation, and Feuerstein’s (1970; Feuerstein, Jensen, Hoffman, & Rand, 1985) theory of mediate-learning experiences and structural cognitive modifiability (Feuerstein, Rand, Jensen, Kaniel, & Tzuriel, 1987; Jensen & Feuerstein, 1987) were all part of this evolution of ideas (Jensen, Robinson-Zaňartu, & Jensen, 1992).

The Range of Dynamic Models

A number of models of “dynamic assessment” have been developed, each of which attends to the assessment of intra-individual differences. However, there are significant differences between the models, which range from a modified testing-the-limits approach (Carlson & Wiedl, 1976, 1978, 1979) to the structural cognitive modifiability theory of Feuerstein and his colleagues (Feuerstein, 1970; Feuerstein, Haywood, Rand, Hoffman, & Jensen, 1985) and Jensen (1990; 1992). Underlying theoretical assumptions, measurement, examiner-examinee interactions, goals for change, number, and types of parameters targeted for intervention, and assumptions regarding transfer effects vary widely across these models, and may be conceptualized on a continuum (see Table 8.2).

On one end of the range are those that maintain psychometric standardization and insert training or direct instruction in problem solving between test trials (Budoff, 1987a, 1987b; Budoff & Friedman, 1964; Carlson & Wiedl, 1978, 1979). Campione & Brown (1987), strongly influenced by the work of Vygotsky (1978) and neo-Vygotskians in Russia, observed the effects of instruction on targeted tasks, focusing on “readiness to learn.” Their work seems to presume that one cannot actually influence the readiness, but only locate it and determine who will profit most from instruction. The assessed measure of gain, which they refer to as dynamic assessment, is presumed to have greater predictive utility than the initial unaided level of performance. They remark that although more clinically based procedures may in fact yield richer
information, their choice was an approach that could yield strong quantitative data.

Feuerstein’s (1970; Feuerstein, Haywood, Rand, Hoffman, & Jensen, 1985) structural cognitive modifiability outlined a far more clinical model. Here, using a series of nonacademic, or de-contextualized tools, designed to tap cognitive skills in various modalities, the examiner used both specific interactive behaviors known as mediation and some 30 specific cognitive functions to elicit and attempt to re-form cognitive habits. The functions identified as basic to that individual’s enhanced cognitive functioning, and the specific mediations found effective in the enhancement process, then became the targets of intervention, assuming these new skills would then transfer to new tasks. Feuerstein’s work has permeated the educational communities across North America, South America (e.g., Venezuela, Chile), Africa (South Africa) and Europe with training and applications of his Learning Potential Assessment Device (LPAD) and companion Instrumental Enrichment (IE) intervention program. These have been used not only with the low-performing population for which it was originally designed, but with bilingual and gifted children as well. Feuerstein’s work has been perhaps the most controversial of these (change) models because of its radical departure from long-standing concepts of stable individual differences, assumptions about upper limits of individual potential, and reliance on psychometric measurement. Jensen (1990; 1992) extended this change model, addressing not only cognitive functions, but knowledge-structure development, as well. We will refer especially to the latter change models in the remaining discussions of the nature of dynamic assessment and mediated learning interventions.

Goals of Enhanced Cognitive Functioning

Interventions designed to accompany dynamic assessments presume to modify the nature of the individual’s cognitive functioning or learning processes over time. Initially, mediated learning interventions depart from the use of (contextualized) tasks with specific academic context such as reading or arithmetic in the attempt to target underlying cognitive skills without the interference of motivational barriers. They rely instead on using problem-solving tasks which lend themselves to interaction with the examiner, and are designed to require specific, often progressively complex or abstract cognitive skills. Through specific “mediational” interactions, fragile areas of cognitive functioning (e.g., comparative behavior, systematic exploratory behavior; use of two or more sources of information) are gradually modified and new habits or skills formed. These new skills are then gradually tested and applied in settings of increasingly distant transfer, with application to the curriculum an example of far transfer.

A major goal in the mediation or training of cognitive skills is their transfer to new situations, thus enhancing the ease and flexibility with which learners approach new information and problem-solving. It is in the arena of near transfer to other problem-solving situations, including presumed measures of intelligence, that a fair amount of evidence supports the efficacy of mediated learning to date (Babad & Budoff, 1974; Budoff, 1987a; Carlson & Wiedl, 1978, 1980; Feuerstein, Rand, Hoffman, & Miller, 1980; Johnson, 1996; Klauer, 1989; Lidz & Peña, 1996). For instance, Thickpenny & Howie (1990) evaluated the effects of teaching thinking skills to deaf adolescents and found significant gains on two subtests of the WISC-R as well as on the Matching Familiar Figures test. Campione & Brown (1987) summarized three sets of studies of mediated learning, in which they found that near transfer was substantial for “low ability” students. Johnson (1996) reported significant increases in Full-scale IQ scores on the WISC-R following a semester of mediated learning instruction in a pilot study with low-functioning children.

In one of the most rigorous examinations of these issues, Jensen & Singer (1987) measured the effects of Feuerstein’s IE Program, using the full three-year program with 234 experimental and 164 control low-functioning adolescents. They provided clear evidence for the acquisition and near transfer of new cognitive functions (Jensen & Singer, 1987). In addition, Jensen (1990) found that a factor analysis of these functions loaded into four categories; the first three, which he termed reception, transformation, and communication, closely paralleled Feuerstein’s clinical grouping of input, elaboration, and output, thus contributing to the validity of Feuerstein’s grouping and processes. The fourth, termed cognitive control, was associated with the role of impulsivity and control over the tempo of the mental act. However, and of great importance, far transfer to significant academic improvement was not found (Jensen & Singer, 1987). Shortly thereafter, Jensen (1990,
1991) postulated that a review of the research must lead us to conclude that cognitive enrichment programs by themselves have not been effective in producing better academic outcomes. This research led to his later postulations regarding the development of cognitive and knowledge structures (Jensen, 1992; Jensen, Robinson-Zañartu, & Jensen, 1992).

**Goals of Enhancing School Achievement**

Although researchers working within a dynamic assessment/mediated learning paradigm have found significant changes in the manifest levels of cognitive functioning of both school age and adult learners, evidence suggests that far transfer from enhanced problem-solving to enhanced academic performance does not automatically transfer to content areas, but must be deliberately taught. Jensen (1992) has defined this problem as the attempt to “procedurralize” knowledge, or actually infuse new cognitive processes into content or knowledge areas. Beginning in the late 1980s, other researchers who had worked with dynamic models of cognitive modifiability (e.g., Greenberg, 1990; Harth, 1982; Haywood, Towery-Woolsey, Arbitman-Smith, & Aldrudge, 1988; Perkins, 1987; Salema & Valente, 1990) were proposing that instructional models must be developed that actually infused and deliberately worked on transfer to the school curricula.

Some researchers have targeted specific cognitive skill development within specific content areas. Salema & Valente (1990), for instance, examined effects of systematic teaching of thinking skills and metacognition in developing composition skills in Portuguese among low achievers. They reported a significant difference in experimental and control groups in learning to write compositions. Consistent with the far-transfer question, that shift did not transfer into other subject matter. Krieglar & Kaplan (1990) used an abbreviated form of Feuerstein’s Instrumental Enrichment (IE) Program to try to demonstrate a bridge to reading achievement, assuming that inattention, hyperactivity, and impulsivity were intervening variables that interfered with reading performance and could be mediated. They created cognitive links to a specific reading task and found significant differences between experimental and control groups on (a) teacher ratings of attention, (b) reading accuracy, and (c) the Porteus Maze test. Perhaps the most broadly applied of these models to date is Greenberg’s (1990, 1992, 1994; Greenberg, Coleman, & Rankin, 1993) Cognitive Enrichment Network (COGNET) Educational Model. COGNET gives teachers a set of cognitive strategies and mediating skills to infuse within and across the curriculum. Models, known as mini-lesson plans, are adapted for each teacher’s curricular needs. In addition, the program encourages and models the incorporation of cooperative learning strategies concurrent with the mediated learning strategies. Greenberg’s (1994) research on the effects of the program in four different types of schools reported that high-risk students in the COGNET schools made greater gains overall than comparison groups on standardized tests of basic skills.

**The Dynamic Assessment and Mediation of Cognitive and Knowledge Structures**

Mediated learning and dynamic assessment as described in Modifiability Enhancement Theory (MET), the change model proposed by Jensen (1992; Jensen, Robinson-Zañartu, & Jensen, 1992), is useful for exploring the integration of thinking skills into academic areas, as it provides: (a) a description of the relationship and rationale of the content-process link in learning, (b) a process by which assessment is linked with intervention, and (c) incorporation of the contextualizing culture of the child (Robinson-Zañartu & Cook-Morales, 1992). Further, MET posits that the proceduralization of knowledge structures is a highly specific process; thus, automatic transfer cannot be assumed, but must be facilitated.

Knowledge structures are proposed in MET to be formed via a process called proceduralization. In proceduralization, the factual knowledge base of any given content area is woven together with cognitive structures, motivational factors, and personality attributes to form a set of processes that enable efficient collection, transformation, and communication of information within that content area. Cognitive functions, such as comparative behavior, planning, and strategies for inferential thinking, are theorized to convert information in an analytical manner, generating an expected outcome from given input. Associators, such as the experience of disequilibrium, interiorization, formation of mental representations, and imaginative hypothetical thinking, are theorized to convert
Table 8.3. Continuum of Assumptions and parameters of Dynamic Models of Assessment

<table>
<thead>
<tr>
<th>Adapted theoretical foundation</th>
<th>Comprehensive theoretical foundation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single cognitive skill mediated</td>
<td>Mediation of multiple cognitive functions, personality and motivational factors</td>
</tr>
<tr>
<td>Standardized administration</td>
<td>Nonstandard administration guided by attempts to modify cognitive structure</td>
</tr>
<tr>
<td>Assumption of far transfer</td>
<td>Mediation for far transfer</td>
</tr>
<tr>
<td>Psychometric measurement</td>
<td>Measurement of process of change</td>
</tr>
<tr>
<td>Task-related changes sought</td>
<td>Structural change sought</td>
</tr>
<tr>
<td>Upper limits presumed</td>
<td>Upper limits not presumed</td>
</tr>
</tbody>
</table>

Note: Carlson & Wiedl >>>Campione & Brown >>>Budoff>>>Feuerstein>>>Jensen

information by an associative process that may generate a variety of potential outcomes. Together, the functions and associators form cognitive structures, contributing intellective capacity to human functioning. Motivational factors, such as a need for mastery, a desire for novelty, and the presence of aspirations, determine the inclination of the individual to engage in mental acts and, in turn, support those mental acts. Personality attributes such as self-confidence, frustration tolerance, and optimism, are seen in MET to determine aspects of the manner and style of the individual’s cognitive and knowledge-structure development (Jensen, 1992). The variables in this model have been found to be sensitive to change, and thus able to contribute to enhancement of functioning. Approximately half were carefully investigated by Feuerstein and his colleagues (Feuerstein, Haywood, Rand, Hoffman, & Jensen, 1985; Jensen & Feuerstein, 1987). The additional variables are currently under investigation at Delphi Health & Science (Jensen, 1991).

Stability versus Change Models

The basic underlying assumptions and characteristics of dynamic versus static assessments are framed in the paradigms of stability versus change. Table 8.3 presents a synopsis of those differences.

Stability models (e.g., measurement of I.Q.) presume stable individual differences that can be measured, yielding a valid indication of current functioning, as well as a prediction of future performance. Thus, stability models often limit the expectations set out for the individual measured, and are consistent with such practices such as labeling, classifying, and placement in situations such as classrooms, in which the environment is often modified to accommodate those expectations (e.g., simplified curricula). Stability models are characterized by an orientation toward products (e.g., numerical predictors), and assume that naturally occurring individual differences in ability exist. Their influence has been particularly strong in the field of psychometrics, where the search for stable individual differences yielded methods for the identification and classification of individuals based on their performance on standardized and normed tests (Jensen, 1992). The efficacy of this paradigm has been called into question for some time across subfields of psychology in which empirical findings point to context as a critical variable in understanding human functioning (Basic Behavioral Science Task Force of the National Advisory Mental Health Council, 1996; Bowlby, 1960; Harlow & Harlow, 1966; Kaufman & Rosenblum, 1967; Rogoff & Chavajay, 1995; Sackett, 1967).

Change models, as applied to human learning and modifiability, operate from the assumption that context is a critical variable that must be applied to the evaluation of human functioning. Although these models vary in their attention to such factors as age, etiology, and severity of impairment, they share the assumptions that the human organism is an open system and, therefore, that assessment should target the possibilities for cognitive and motivational enhancement (Jensen, 1992). They posit that human nature is cultural, and that learning involves the processing of contextually (e.g., culturally) meaningful symbols. Further, they propose that learning is a dynamic and open process in which active modification can be applied to the enhancement of functioning (Jensen, 1992; Robinson-Zaïartu & Cook-Morales, 1992).
Dynamic Attention to Process versus Static Attention to Product

Dynamic assessment represents a significant departure from the static product-oriented model in which the only process involved is the posing of questions or situations, and recording of responses by the examiner, who remains deliberately separate from the process. In dynamic assessment the focus of attention is on the examinee’s learning process, which is evaluated under conditions of learning. That is, the examiner not only observes but intervenes in the assessment process based on those observations, attempting to modify cognitive or learning approaches and then observe the results of the interventions. The targets of interventions are drawn from a repertoire of cognitive functions and motivational factors.

Dynamic assessment as outlined in both Jensen’s and Feuerstein’s models holds that learning begins with the primary caregiver, whose context is their primary culture and language. Thus, that caregiver is the first mediator of learning: he or she introduces the child to the elements of the environment with intentionality, establishes a reciprocal relationship, helps frame and shape behaviors associated with learning, and gives meaning and motivation to the interactions and emergent learning. These behaviors comprise essential characteristics of mediation. These behavioral features of mediation, in which the examiner engenders intentionality-reciprocity, the feeling of competence, regulation of behavior or cognitive pacing, transcendent value to the immediate experience, and a sense of meaning of change as it occurs, characterize the nature of the examiner-examinee interaction in a dynamic assessment.

Issues of Measurement

Issues of measurement, validity, and reliability within change models are significant concerns of theoreticians operating from within as well as outside of the dynamic assessment paradigm. Although outcomes on certain dimensions of acquisition or even proceduralization may be measured against outcomes of alternative methods, using a variety of criteria (e.g., curriculum based measurement; authentic/portfolio assessment), the necessary level of individualization of each dynamic assessment precludes the use of some conventional measures of validity and reliability. The shift in basic assumptions will require a shift from the psychometric approach to mathematical constructs that are designed to measure change rather than stability. The focus is not to compare products of one individual with others, but to measure processes of change within that individual. Thus, they must depend on understanding those changes: what is presumed to undergo change, how that is accomplished, and when they are said to have occurred.

Jensen (1992) has proposed and is currently researching a mathematical measure of performance efficiency which utilizes computer-assisted touch-screen technology. This technology is able to record learner responses instantly and continually, and produces a within-subject measure of level and change of efficiency. The learning curve representing the changing value of efficiency provides data from which reliability and validity of those functional changes can be ascertained. He suggests that although individual and group statistics could be computed for this measure, its primary significance lies in its clinical application to the mediation process while the effort is being made to develop new modes of functioning.

Transfer of Effects

Transfer of the effects of dynamic assessment and mediated learning intervention have been discussed in some detail above. A substantial body of research supports the assumption that this model leads to enhanced cognitive functioning and problem-solving skills within training contexts and on tasks similar to those used in training. Some evidence also suggests that teachers perceive their students as better problem-solvers following mediated learning interventions (Greenberg, 1990; Kriegler & Kaplan, 1990). However, far transfer has yet to be established. Goals of education are currently being re-examined for their relevance to student needs in the 21st century. In the United States, mandates for educational reform now stress that the goals of education should reach beyond academic achievement. Significant attention is being directed toward enhanced thinking skills (Carnegie Council on Adolescent Development, 1989; National Educational Goals Report, 1991) as well as producing outcomes that enable the learner to profit from instruction (Ysseldyke & Thurlow, 1992).
While research with these models demonstrates enhanced problem-solving skills in near transfer, it does not yet support the assumption of transfer to broad academic enhancement, although a number of studies now indicate that when bridged to specific content, enhanced cognitive skills can enhance content acquisition (Greenberg, 1990; Kriegler & Kaplan, 1990; Salema & Valente, 1990). Promising preliminary studies and new theoretical directions may change these conclusions about far transfer in the future. Greenberg's (1990) COGNET model, for instance, has coupled mediated and cooperative learning strategies to produce demonstrable academic gains across multiple-content areas. Because of his attention to the multiple-dimensions variables in the enhancement of aptitude, Jensen's theoretical work on proceduralization of knowledge seems particularly worthy of rigorous scientific examination.

Applications to Decisions

Three types of applications to decision-making emerge from dynamic assessment: identification of misclassified students, preventive or developmental teaching methodologies, and strategic child interventions. The first identifies students misclassified on traditional intelligence measures as low in ability. Students regarded as misclassified are those who perform at levels on dynamic measures comparable to students in regular classes.

The second and third applications to decision making rely on the articulation of cognitive functions and motivational factors that are believed to enhance learning. The second application infuses mediated learning into the teaching methodology of the regular classroom. Here, cognitive functions implicit in the content of the curriculum are identified, and the functions are mediated within the existing academic content as a part of the lesson plan.

The third application, mediated intervention, specifies individual goals for enhancement determined through dynamic assessment. These targets will be prerequisite functions or factors designed to prepare the child to grasp and use the classroom-based functions. For instance, if the teacher has targeted the cognitive function of categorizing in conjunction with learning grouping in math, the prerequisite function of comparative behavior may need to be targeted to prepare the child. A broad variety of other functions, such as inhibition of impulsive responding, conservation of constancies, precision and accuracy in data gathering, or awareness of a problem would be examined during the dynamic assessment, and appropriate goals and sequences established for the child.

Treatment Validity

Treatment validity for dynamic assessment and mediated learning intervention models is enmeshed with the issues of near and far transfer discussed above. Most models of dynamic assessment have claimed that their aims are to enhance general cognitive functioning or problem-solving skills (e.g., Feuerstein, Haywood, Rand, Hoffman, & Jensen, 1985; Greenberg, 1990, 1992). This criterion has been demonstrated on near-transfer tasks such as non-verbal measures of intelligence and teacher reports of new student approaches to problem solving (Johnson, 1996; Kriegler & Kaplan, 1990). In addition, attempts to deliberately pair learning or thinking skills with academic content have produced evidence of positive outcomes in reading accuracy, writing skills, and teacher ratings of attention (Greenberg, 1990, 1994; Salema & Valente, 1990). However, far transfer of enhanced cognitive functioning to either general cognitive functioning or enhanced broad academic performance has not yet been adequately demonstrated. Researchers are currently involved in new directions that apply the cognitive-skills enhancement directly to curriculum. These directions may lead to evidence of these models becoming more useful in the enhancement of student achievement.

SUMMARY

Reflecting the trends of the past decade, aptitude is examined from a perspective that includes any measure designed to predict individual differences in later learning. Specifically, attention is directed to three models of cognitive- and processing-aptitudes measures presumed to improve diagnosis, placement, and/or treatment of children and youth with learning problems: (1) psycholinguistic (PL) and perceptual-motor (PM) models, (2) aptitude by treatment interaction (ATI) models, and (3) change models of dynamic assessment (DA) and mediated learning intervention. All three models share the assumptions that a repertoire of cognitive pro-
cesses underlie the learning process, that examination of intra- rather than inter-individual differences should yield important data in the design of training or instruction, and that enhanced cognitive functioning and academic achievement should be ultimate goals in interventions based on these models.

Significant differences between the models lead to differing conclusions about their usefulness. The ATI models presume that aptitudes are stable individual traits, relatively insensitive to intervention. PL and PM models presume that deficient functions can be identified and changed with interventions. Consistent with stability models, ATI, PL, and PM models rely on cognitive responses (or products) as the primary unit of observation, and conventional standardized assessment procedures. In contrast, DA models assume that cognitive functions are not only modifiable, but interrelated, so that their enhancement would lead to the enhancement of overall cognitive skills. Consistent with change models, they place the unit of observation on the thinking processes used to produce a variety of responses, rather than the products. Procedures in DA are dynamic and interactive, guided by the thinking processes of the student and, therefore, are necessarily nonstandard.

Each model differs on presumed transfer of effects, leading to differences in the decisions associated with the models. ATI models assume that stable strengths should be matched with treatment or teaching to that strength, thus producing enhanced functioning. The PM and PL models assume that deficient functions can be trained, thus leading to enhanced functioning. Decisions related to these stability models usually relate to the kind of instruction that should be used, or to classifying a student for placement based on the assumption of certain cognitive or processing deficits. Dynamic change models assume that transfer can be achieved through the intervention processes identified during and integral to the assessment. Decisions relate to teaching methodology aimed at enhancing overall functioning. Interventions are interactive behaviors (mediations) applied by teachers, parents, or specialists to the targeted cognitive functions in the context of the curriculum.

When the standard of treatment validity based on enhanced academic outcomes is applied, two models currently fall short of expectations; the third is insufficiently researched to draw firm conclusions. Although the reasons for negative treatment validity evidence may be weaknesses in the basic theory, the measure of aptitude, or the available interventions, ATI research shows very weak interactions at best. In the case of PM, interventions appear to have limited effects on PM processes. PL interventions appear to have some minimal effect on PL processes, but no demonstrated effect on achievement. DA interventions appear to have a modest effect on cognitive enhancement and problem solving that is restricted to similar tasks (such as other non-verbal problem-solving tasks). However, no compelling evidence of transfer to broad areas of achievement exists to date. Recent theoretical and research developments in this area appear promising, and should be followed. Experiments that deliberately pair learning skills and academic content may produce better evidence of transfer of cognitive training effects. With increased attention in education to the role of thinking skills, assessment and intervention approaches that can provide evidence of enhanced thinking and enhanced achievement may become increasingly important.

REFERENCES


Evaluation of Aptitudes


Suggested Readings


