CHAPTER 1

PSYCHOLOGICAL ASSESSMENT BY SCHOOL PSYCHOLOGISTS: OPPORTUNITIES AND CHALLENGES OF A CHANGING LANDSCAPE

Jack A. Naglieri

The reliability and validity of information obtained from any psychological test is dependent on the scope and psychometric attributes of the instrument used. As in all areas of science, what psychologists discover depends on the quality of the instruments used and the information they provide as well as skillful interpretation of the test results. Better conceptualized instruments yield more accurate and informative data than do weaker instruments. Instruments that uncover more useful information about the individual being examined are more valid and ultimately better inform both researchers and clinicians. The tools school psychologists choose for diagnostic decision making substantially influence the reliability and validity of the information they obtain and the decisions they make. Simply put, the better the tool is, the more valid and reliable the decisions; the more useful the information obtained is, the better the services provided. In this chapter, some important issues regarding quality and effectiveness of the tools used in school psychology are discussed.

The purpose of this chapter is to discuss some important issues in school-based psychological and educational assessment. To capture the essence of the major changes occurring in the schools, the chapter is organized into three sections. The first section involves the role of intelligence tests in determining learning disability eligibility. Next, some changes in achievement testing are reviewed. Third, evaluation of social–emotional status is examined. Each of these areas has been influenced by a combination of federal legislation and changes in school psychological practice, as described by the National Association of School Psychologists (2010). The goal of this chapter is not to summarize all the changes that have recently occurred or to predict the outcomes of these changes but rather to summarize a few important issues related to the current state of the field and the apparent strengths and weaknesses of the various options.

INTELLIGENCE AND SPECIFIC LEARNING DISABILITIES

Controversy is not new to the construct of intelligence and its measurement (see Jensen, 1998). Arguments have raged about the nature of intelligence—is it one factor or multiple factors, are intelligence tests biased or not, what are the best ways to interpret test results, do children with specific disabilities have distinctive ability profiles, and do intelligence test scores have relevance beyond diagnostic classification (e.g., implications for instruction and treatment)? In recent years, the most important questions have centered on the utility of intelligence tests for evaluation and treatment of children suspected of having a specific learning disability (SLD). More important, although the construct of general intelligence has considerable empirical support (see Jensen, 1998, for a review), especially when measured by tests such as the Wechsler scales and the Stanford–Binet, the value of traditional intelligence tests for evaluation of children with SLD is less clear.

There is little doubt that the psychometric characteristics of the Wechsler and Binet tests, the oldest

DOI: 10.1037/XXXXX.XXX
APA Handbook of Testing and Assessment in Psychology: Vol. 3. Testing and Assessment in School Psychology and Education,
K. F. Geisinger (Editor-in-Chief)

intelligence tests, have advanced considerably over the past 30 years (see O'Donnell, 2009, and Roid & Tippin, 2009, for summaries). The hallmark of their advancement has been improved psychometric qualities, including improved reliability, more representative normative samples, more attractive physical materials, and computer-assisted scoring and interpretive analysis. These improvements have provided clear advantages to traditional intelligence tests over their predecessors. Despite excellent psychometric qualities, the limitations of these traditional tests have been noted by many, particularly those related to the evaluation and classification of children with SLD.

One of the most important and hotly debated limitations, particularly relevant for school psychologists, is the diagnostic value and stability of Wechsler subtest profiles. The interpretation of subtest profiles is widely accepted by many practitioners and was encouraged in many influential textbooks (e.g., Kaufman, 1979). Over time, however, a series of compelling articles have been published that have questioned the stability of subtest profiles and strongly suggested that Wechsler subtest variability is ineffective for diagnosis (see McDermott, Fantuzzo, & Glutting, 1990). It is quite clear that traditional intelligence tests that were developed in the early 1900s as measures of general ability may not meet more modern purposes, especially for evaluation of SLDs. To clarify the role intelligence tests play in the evaluation of SLD, the definitions of this disorder and the limitations and strengths of the tests that are used should be understood. This chapter, therefore, provides a brief summary of the Individuals With Disabilities Education Improvement Act of 2004 (IDEA), assessment issues related to learning disabilities in school psychology, and some summative data on current test profiles.

Learning Disabilities Defined

In the schools, IDEA (2004) and related state laws define SLD. For those in the medical profession and many psychologists in independent practice, the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text revision; *DSM–IV–TR;* American Psychiatric Association, 2000) is often used. Both of these definitions involve evaluation of a child's cognitive abilities. The *DSM–IV–TR* bases diagnosis on an inconsistency between assessed ability and achievement (i.e., reading, math, or written expression or a nonspecified academic area) when that difference is not better accounted for by other life conditions, such as inadequate education, cultural or ethnic differences, impaired vision or hearing, or mental retardation. The *DSM–IV–TR* definition is based on documenting achievement scores on individually administered, standardized tests in reading, mathematics, or written expression that are substantially below that which would be expected for peers of comparable age, schooling, and level of intelligence. The size of the discrepancy should be at least 1 standard deviation if the intelligence test score might have been adversely influenced by an associated disorder in cognitive processing, a mental disorder, or the ethnic or cultural background of the individual and 2 standard deviations if not. More important, the learning disorder should significantly interfere with the student's reading, math, or writing (which can be quantified with a variety of achievement tests) or daily living (which is often more difficult to quantify). The *DSM–IV–TR* also recognizes or assumes that problems with cognitive processing (e.g., deficits in visual perception, linguistic processes, attention, or memory) may have preceded or be associated with the learning disorder (i.e., the underlying cause of the disorder).

A SLD as defined in IDEA (2004) has similarities to and differences from the definition used in the *DSM–IV–TR.* The similarities include academic failure not explained by inadequate education, cultural or ethnic differences, or impaired vision or hearing, mental retardation, or other disability. The differences between IDEA and the *DSM–IV–TR* include (a) the age range for which the definition applies, (b) the disability being described as a specific disability, and (c) the definition of the disability as a disorder in basic psychological processes. The IDEA definition is as follows:

> Specific learning disability means a disorder in one or more of the basic psychological processes involved in understanding or in using language, spoken or written, which may manifest itself

> in an imperfect ability to listen, think, speak, read, write, spell, or to do mathematical calculations. The term includes such conditions as perceptual handicaps, brain injury, minimal brain dysfunction, dyslexia, and developmental aphasia. The term does not include children who have problems that are primarily the result of visual, hearing, or motor disabilities, or mental retardation, emotional disturbance, or of environmental, cultural, or economic disadvantage. (pp. 11–12)

The measurement of intelligence plays a key role in both approaches to SLD determination, and these differences have important implications for the use of intelligence. Both the *DSM–IV–TR* and IDEA (2004) definitions involve a comparison of ability with achievement, the so-called ability–achievement discrepancy model. This approach has been widely criticized for some time and is no longer considered effective (see Fletcher, Denton, & Francis, 2005; Meyer, 2000; Stanovich, 1994) and is no longer required under IDEA (Kavale, Kaufman, Naglieri, & Hale, 2005).

To go beyond the ability–achievement discrepancy, practitioners were encouraged to examine subtest profiles, expecting that this information could aid in eligibility determination and instructional planning. Kaufman (1979) was among the first to recognize the limitation of global ability scores and suggested that useful information about a child could be obtained by a careful, psychometrically defined examination of subtest scores. Over time, the idea of going beyond the Full Scale IQ and the difference between that global score and achievement has gained favor, but using subtest analysis has not. More recently, greater emphasis has been placed on theoretically guided interpretations as described by Naglieri (1999); Flanagan, Ortiz, Alfonso, and Mascolo (2002); and Hale and Fiorello (2004). Before these methods are described, an examination of profiles for intelligence test scales rather than subtests is provided.

Because intelligence tests play such an important role in SLD eligibility determination, it is important to ask the question, "Do intelligence tests yield scale profiles that are distinctive for children with SLDs?" Naglieri (1999, 2000) suggested that subtest profile analysis should be replaced by scale profile analysis so that diagnostic reliability could be increased and, more important, so that each scale should be clearly related to some theoretical ability construct. To examine this method of profile analysis, Naglieri and Goldstein (2011) provided an examination of intelligence test profiles for adolescents and adults with SLD on the basis of information provided in the respective test manuals or book chapters of Naglieri and Goldstein (2009). They found that traditional intelligence tests did not yield a pattern of scores on scales encompassing these tests that was distinct to any one type of disability. This chapter describes a broader analysis of scores based on samples of children ages 5 to 18 years.

The research on intelligence test scale profiles is summarized next with the goal of examining mean score patterns of the scales for children with reading failure. This review helps to determine whether ability tests show particular patterns for children with a SLD in reading decoding. This information could have important implications for understanding the cognitive characteristics of that clinical group, which might allow for possible diagnostic and intervention considerations (Naglieri, 1999). To compile data from various intelligence tests, several different sources were used. Reports in the technical manuals were used for the Wechsler Intelligence Scale for Children—Fourth Edition (Wechsler, 2003), Stanford–Binet—Fifth Edition (Roid, 2003), Differential Ability Scales—Second Edition (Elliott, 2007), Kaufman Assessment Battery for Children—Second Edition (Kaufman & Kaufman, 2004), and Cognitive Assessment System (CAS; Naglieri & Das, 1997). (The CAS data also included findings from Naglieri, Otero, DeLauder, & Matto, 2007.) The findings, however, must be taken with recognition that the samples were not matched across the various studies, the accuracy of the diagnosis may not have been verified, and some of the sample sizes were small. Notwithstanding these limitations, the findings provide important insights into the extent to which these various tests can be used for assessing adolescents and adults suspected of having a specific learning disorder.

The results of this analysis are provided in Figure 1.1, which includes the standard scores obtained on these various intelligence tests for students with a specific reading disability. The comparison of scale profiles across the various ability tests suggests that some tests are more sensitive to the cognitive characteristics of individuals with specific reading disabilities than others. The Differential Ability Scales—Second Edition, Stanford–Binet—Fifth Edition, and Kaufman Assessment Battery for Children—Second Edition showed relatively little variability among the scales; the differences between the lowest and the highest scale within each test were 3.2, 3.8, and 3.8, respectively. That is, the pattern of scores on the separate scales making up these tests did not suggest that a specific cognitive disorder was uncovered. The scales on the Wechsler Intelligence Scale for Children—Fourth Edition showed more variability (range = 7.4), followed by the Woodcock–Johnson III Tests of Achievement (Woodcock, McGrew, & Mather, 2007; range = 10) and the CAS (range = 10.3). More important, the

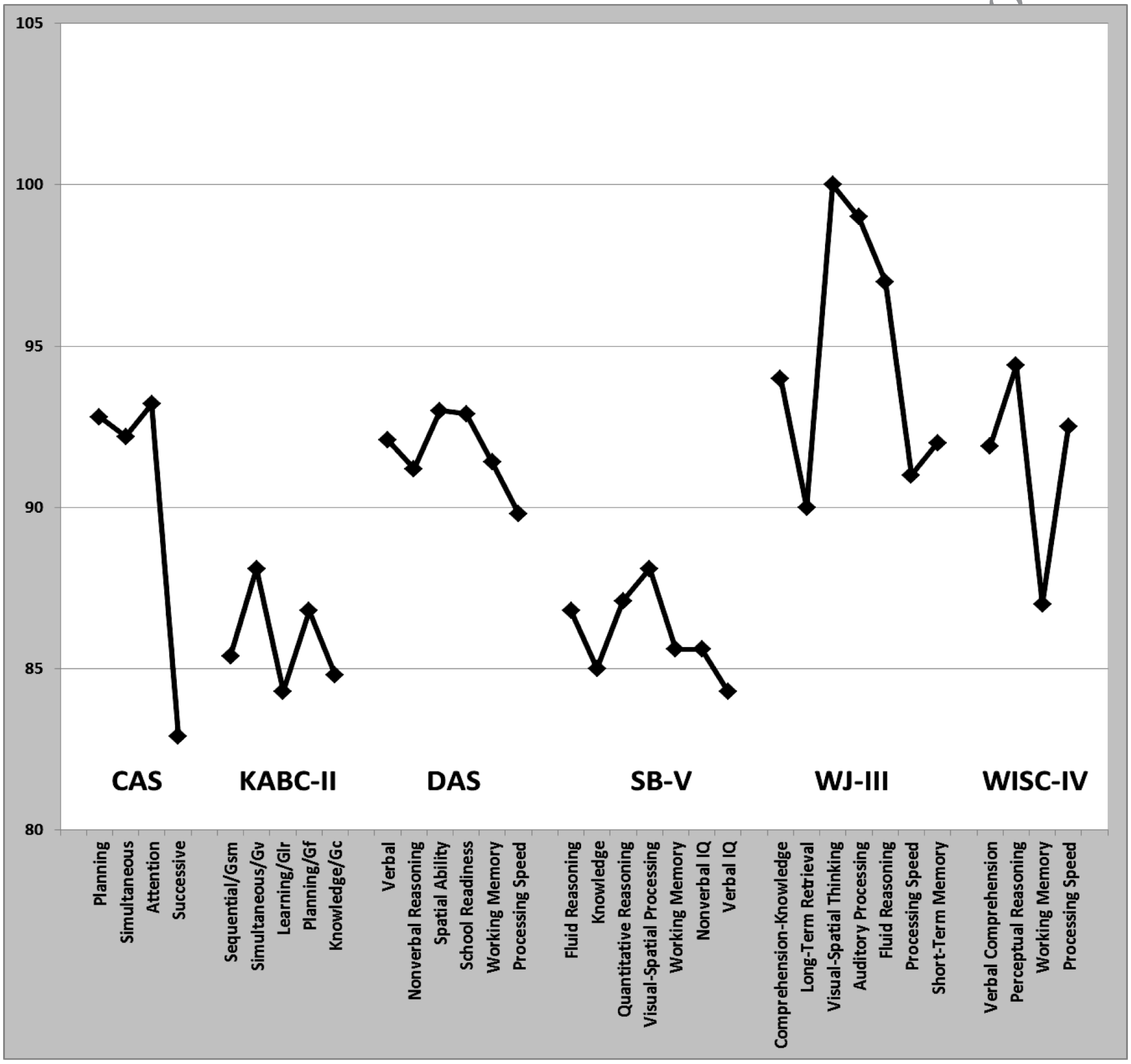

FIGURE 1.1. Mean scores earned by samples of students with reading decoding disorders by ability test. CAS = Cognitive Assessment System; KABC–II = Kaufman Assessment Battery for Children—Second Edition; DAS = Differential Ability Scales—Second Edition; SB–V = Stanford–Binet—Fifth Edition; WJ–III = Woodcock–Johnson III Tests of Cognitive Abilities; WISC–IV = Wechsler Intelligence Scale for Children—Fourth Edition.

lowest score (90) on the Woodcock–Johnson III was for Long Term Retrieval, which measures associative memory (Wendling, Mather, & Shrank, 2009). The Wechsler Intelligence Scale for Children—Fourth Edition profile suggests that the sample was low in Working Memory and the remaining scales were in the low end of the average range, such as for the CAS Planning, Simultaneous, and Attention Scales. Interestingly, the Working Memory tests on the Wechsler Intelligence Scale for Children—Fourth Edition require repetition of numbers in the order provided by the examiner (Digit Span Forward) or in the reverse order (Digit Span Backward) and recitation of numbers in ascending sequential order and letters in alphabetical order (Letter–Number Sequencing), both of which require sequencing. Schofield and Ashman (1986) showed that Digit Span Forward and Digit Span Backward correlated significantly with measures of successive processing as measured in the CAS.

The CAS showed the most variability (range = 10.3) even though three of the four scale means were within 1 point of each other. The exception was successive processing ability, on which the sample earned a very low score (82.9). The CAS profile for the sample with SLD suggested that this group had a specific academic (reading decoding) and a specific cognitive weakness (successive), meaning that as a group, these individuals had difficulty working with stimuli that are arranged in serial order, as in the sequence of sounds that make words, the sequence of letters to spell words, and the sequence of groups of sounds and letters that make words. Taken as a whole, these findings suggest that the tool with which practitioners choose to evaluate children suspected of having a SLD may or may not uncover a disorder in one or more of the basic psychological processes required in the IDEA (2004) definition.

Next Steps

The evaluation of children with a SLD is among the most complex and contentious issues facing the field of school psychology. Because IDEA (2004) specifies that children with SLD have a disorder in one or more of the basic psychological processes, cognitive processes must be measured (Kavale et al., 2005). A comprehensive evaluation of the basic psychological processes unites the statutory and regulatory components of IDEA and ensures that the methods used for identification more closely reflect the definition. Any defensible eligibility system would demand continuity between the statutory and regulatory definitions, and for this reason alone SLD determination requires the documentation of a basic psychological processing disorder (Hale, Kaufman, Naglieri, & Kavale, 2006). Moreover, the tools used for this assessment must meet the technical criteria included in IDEA, and well-validated measures of cognitive and neuropsychological measures are available that can be used to document SLD (Hale & Fiorello, 2004; Kaufman & Kaufman, 2001; Naglieri & Otero, 2011). To use a cognitive processing approach to SLD identification, three main components are needed. First, the child must have significant intraindividual differences among the basic psychological processes, with the lowest processing score substantially below average. Second, average processing scores and some specific area of achievement need to differ significantly. Third, consistency between poor processing scores and a specific academic deficit or deficits is essential (Hale & Fiorello, 2004; Naglieri, 1999, 2011). These systematic requirements are collectively referred to as a *discrepancy–consistency model* by Naglieri (1999, 2011) and as the *concordance–discordance model* by Hale and Fiorello (2004).

Naglieri (1999, 2011) described the discrepancy–consistency model for the identification of SLDs on the basis of finding a cognitive processing disorder (see Figure 1.2). The method involves a systematic examination of variability of basic psychological processes and academic achievement test scores. Determining whether cognitive processing scores differ significantly is accomplished using the method originally proposed by Davis (1959), popularized by Kaufman (1979), and modified by Silverstein (1993). This so-called ipsative method determines when the child's scores are reliably different from the child's average score. It is important to note that in the discrepancy–consistency model described by Naglieri (1999), the ipsative approach is applied to the scales that represent four neuropsychologically defined constructs, not subtests from a

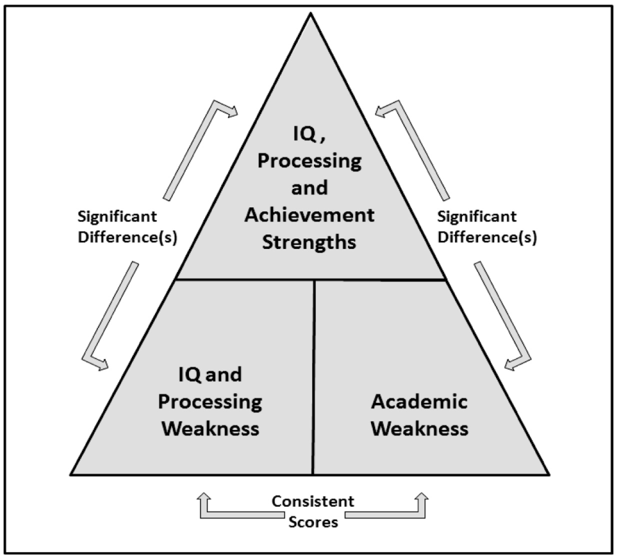
IQ ,
Processing
and
Achievement
Strengths
Significant Difference(s)
Significant Difference(s)
IQ and Processing Weakness
Academic Weakness
Consistent Scores

FIGURE 1.2. **Naglieri's (1999, 2011) discrepancy–consistency model for determination of specific learning disability.**

larger test of ability. This distinction is important because the criticisms of the ipsative method (McDermott et al., 1990) have centered on subtests, not scale-level analysis. In contrast, good evidence for the utility of using scales (from CAS) for diagnosis has been reported (see Canivez & Gaboury, 2010; Huang, Bardos, & D'Amato, 2010), and Huang et al. (2010) concluded that their study "substantiate[d] the usefulness of profiles analysis on composite scores as a critical element in LD [learning disability] determination" (p. 19).

Naglieri (1999) and Flanagan and Kaufman (2004) recognized that because a low score in relation to the child's scale mean could still be within the population's average range, adding the requirement that the weakness in a processing test score also be well below the average range is important. For example, Naglieri (2000) found that those students who had a low processing score relative to their personal mean and the normative group were likely to have significantly lower achievement scores and were more likely to have been identified as having an academic disability. That study was described by Carroll (2000) as one that illustrated a more successful profile methodology. Davison and Kuang (2000) suggested that "adding information about the absolute level of the lowest score improves identification over what can be achieved using ipsative profile pattern information alone" (p. 462). More important, this method has been shown to have implications for instruction for children with SLD (Naglieri & Gottling, 1995, 1997; Naglieri & Johnson, 2000) and attention deficit/hyperactivity disorder (Iseman & Naglieri, 2011) and to be tied to many instructional methods used in the classroom (Naglieri & Pickering, 2010).

Hale and Fiorello's (2004) proposed method, the concordance–discordance model, is based on the cognitive hypothesis testing methodology that relies on multiple assessment tools and data sources to maximize validity of assessment findings. Hale and Fiorello used cognitive and neuropsychological assessment data for both diagnostic and intervention purposes. When cognitive hypothesis testing results suggest that a child may have a SLD, differences among the scores is determined using the standard error of difference (Anastasi & Urbina, 1997) to test differences among the three components of the model: cognitive assets, cognitive deficits, and achievement deficits in standardized test scores. This approach has been advocated for use in school psychology by Hale and colleagues (Hale & Fiorello, 2004; Hale et al., 2006), who also cautioned that the method not be rigidly applied. They argued that practitioners follow the literature to ensure that the apparent cognitive strength is not typically related to the deficit achievement area and that the apparent cognitive weakness could explain the achievement deficit. This method ensures that children identified as having a SLD meet both IDEA (2004) requirements and, more important, has been shown to be relevant to intervention (Fiorello, Hale, & Snyder, 2006; Hale & Fiorello, 2004).

These two methods for identifying children with SLDs provide a means of uniting the definition found in IDEA (2004) with well-standardized tests that practitioners use on a regular basis. As the field of SLD evolves within the context of federal law and federal and state regulations, the applicability of these methods will become more apparent. Although initial research on the effectiveness of these methods for both eligibility determination and remediation of academic deficiencies is encouraging, additional studies are warranted.

ASSESSMENT OF ACHIEVEMENT

Achievement tests used by school psychologists are comprehensive, well-developed, and psychometrically refined tools. For example, tests such as the Kaufman Test of Educational Achievement—Second Edition (Kaufman & Kaufman, 2005; see Lichtenberger & Sotelo-Dynega, 2009), the Wechsler Individual Achievement Test—Second Edition (Wechsler, 2005; see Choate, 2009), and the Woodcock–Johnson III Tests of Achievement (Woodcock et al., 2007; see Mather & Wendling, 2009) are high-quality instruments (see Naglieri & Goldstein, 2009, for descriptions of these and other tests of academic skills). These individually administered tests offer many content-dependent subtests with ample coverage of various aspects of academic achievement, excellent normative samples, and strong psychometric documentation. All of these tests provide age-corrected standard scores that calibrate a student's standing relative to his or her respective standardization groups. Some, for example, the Kaufman Test of Educational Achievement—Second Edition, offer the added advantage of item-level analysis so that the student's score can be more completely described on the basis of which academic skills have been acquired or are in need of instruction. Additionally, some offer excellent psychometric scores that can be used to monitor progress over time (e.g., Wide Range Achievement Test—Fourth Edition; see Roid & Bos, 2009).

Traditional measures of academic skills have been challenged by proponents of the curriculum-based measurement (CBM) field, including two tests representing CBM methodology. Traditional and CBM assessments differ in that CBM measures are used more as universal screening tools for identifying poor readers in early elementary grades and for monitoring academic progress when evaluating the effectiveness of instructional methods. In school psychology, these very brief alternative measures are used for evaluating reading for universal screening and monitoring student progress and as part of the evaluation process for SLD eligibility determination, sometimes in lieu of a comprehensive assessment of academic skills (e.g., Koehler-Hak & Bardos, 2009). This alternative approach to academic assessment is perhaps best illustrated by brief fluency tests and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002).

The approach to testing and monitoring of progress exemplified by the DIBELS differs from that of the traditional achievement tests mentioned earlier as well as from that of group-administered measures of achievement (e.g., the Stanford Achievement Test—10th Edition (Pearson, 2006) that can used for universal screening. Tests such as the Stanford Achievement Test—10th Edition and many individually administered achievement tests (a) are nationally normed on a representative sample of students, (b) cover many different aspects of reading and math curriculum, (c) are based on appropriate learning standards, and (d) yield age-corrected standard scores. These tests can be used to identify students at risk of academic failure on the basis of a comparison to national norms as well as ranking within the classroom, school, or school district. The greatest difference between a test such as the DIBELS and more traditional achievement tests is the brevity of CBM assessments and the shift toward measures that come from the CBM field.

CBM procedures are intended to give educators tests that are reliable, valid, inexpensive, and efficient estimates of student achievement. Researchers have generally found consistency in the relationship between CBM scores and standardized measures across samples and various achievement tests as well as acceptable levels of reliability and validity (Reschly, Busch, Betts, Deno, & Long, 2009). For this reason, these brief tests (e.g., correct words read per minute) have been used to identify children at risk of reading failure and to assess student progress over time. The main differences between tests from the CBM field and traditional achievement tests rest on the CBM assumption that a brief measure of achievement is as effective as a comprehensive, standardized measure of current and future academic performance. So instead of measuring reading comprehension, for example, a 1-minute CBM reading fluency test is used because it correlates moderately with reading comprehension as measured by tests such as the Stanford Achievement Test—10th Edition or the Wechsler Individual Achievement Test—Second Edition. Another important difference is that the CBM measures do not yield scores that are calibrated against a national norm and are not corrected for age effects.

Those who advocate for the use of CBM place emphasis on the goals of identifying children at risk for academic failure and monitoring academic progress over time to determine instructional effectiveness. The psychometric methods used, however, raise several important concerns that have been largely ignored by CBM advocates. These issues include the publication and use of tests without technical manuals that explicate the psychometric quality of the scores the tests yield (e.g., reliability and validity) and, perhaps most important, norms. The use of raw scores as measures of current status and as a means of calibrating current standing and response to intervention is another important difference between CBM measures and traditional normed measures of achievement. In this chapter, I ill focus on issues related to the use of raw scores from the DIBELS Oral Reading Fluency (ORF) test. The first topic concerns the use of raw scores to identify which students may be at risk of academic failure; the second concerns the use of raw scores to monitor progress over time for an individual child; and the third concerns the use of raw scores for the purpose of examining changes in groups of students as a function of some intervention.

Identifying At-Risk Students

To illustrate some of the problems with using raw scores instead of age-corrected standard scores, I show a simple examination. Figure 1.3 shows raw scores that are used as benchmarks (Koehler-Hak & Bardos, 2009) for making decisions about students' academic standing in a classroom. These values are approximately associated with the 40th percentile for the DIBELS ORF. This figure was developed by finding which raw scores were associated with the 40th-percentile scores during the fall, winter, and spring of second grade according to Table 7 in the DIBELS Technical Report Number 9 (Good, Wallin, Simmons, Kame'enui, & Kaminski, 2002). According to the technical report, the transformation of raw scores to "system-wide percentile ranks . . . [was] based on all participating students in the DIBELS Web data system as of May 20, 2002" (Good et al., 2002, p. 2). The test authors used raw scores from the 2001 to 2002 academic year to obtain percentile scores for children tested in the beginning, middle, and end of the year. Good et al.'s (2002) description of the sample used to obtain the raw-score-to-percentile-score conversion is very limited, and they do not indicate whether their sample is representative of the U.S. population. Although this sample description does not meet commonly accepted standards for reporting reference groups for commonly used standardized achievement tests, I used the conversion tables to approximate normal maturation rates in ORF scores. The results provided in Figure 1.3 show that DIBELS ORF raw scores can increase dramatically over the course of a school year while the percentile score associated with the score remains the same. This implies that even though a student can read more words per minute, his or her relative standing has not improved. This situation leaves the user in a quandary: How exactly should examiners interpret raw scores on the DIBELS?

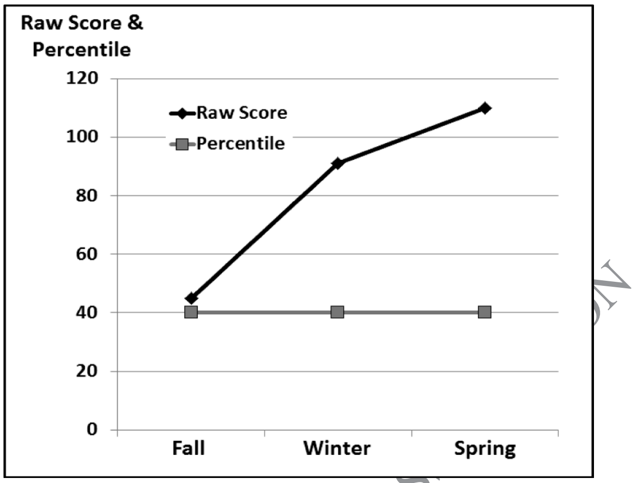

FIGURE 1.3. Relationships between Dynamic Indicators of Basic Early Literacy Skills raw scores and percentile ranks over three points in time.

Some professionals (e.g., Shinn, Tindal, & Stein, 1988) have advocated for the use of local norms to determine whether a child's academic needs are being met in the classroom or whether a referral for special services is appropriate. The apparent expectation is that local norms can help school psychologists make sound data-based decisions and more accurately identify students at risk of academic failure. As an example, I present constructed local norms for DIBELS ORF scores for nine schools (N = 620) and the results across schools and in comparison to the contrast group provided by Good et al. (2002). The data used for this illustration came from a medium-sized city in the mid-South region of the United States. Local norms were constructed by transforming raw scores to z scores and then to standard scores of an IQ metric (M = 100, SD = 15) on the basis of raw-score DIBELS ORF means and standard deviations for each school. In addition to local norms, I converted DIBELS raw scores to standard scores (M = 100, SD = 15) via the percentile ranks provided by Good et al. (2002). That is, raw scores were converted to percentile ranks on the basis of conversion Table 7 in Good et al. Next, I converted percentiles to standard scores (M = 100, SD = 15) using the statistical function NORMINV from Microsoft Excel. This procedure provided a means of comparing local norms with those of a national comparison group (assuming, however, that there is no evidence that this group represents the U.S. population). The findings are quite revealing.

As seen in Table 1.1, the nine schools' mean ORF scores varied considerably, as did minority representation and percentages of students on free or reduced lunch programs. The mean number of words per minute was highest for the school with the least percentage of students receiving free or reduced lunch and the smallest number of minority students. The raw scores corresponding to standard scores are provided in Table 1.2 and show that the standard score a child would earn for the same raw score varies considerably across the nine schools. For example, a raw score of 20 words per minute yields a standard score of 101 for a child in School 1 but a standard score of 86 for a student in School 9. This considerable difference would ensure inequity of assessment and faulty interpretations within the same school district. The problem is that those students in schools in which the raw score mean is lowest will earn scores that are average, implying that no deficiency was found. Even more concerning is that the students who earn a raw score of 20 on the basis of the local norm are actually well below the national reference group, which earned a standard score of 84. In fact, the students in Schools 1, 2, and 3 earned local standard scores of 101, 98, and 98, respectively, but when compared with the national reference group would have earned a standard score of 84 (more than 1 standard deviation below the mean).

The only logical conclusion drawn from this analysis is that local norms mislead the user into thinking that students (as well as teachers and curricula) are doing well when in fact they may be well below what would be considered normal or expected on a national basis. In this illustration, the schools with the lowest scores were those with the highest percentage of minority children. Because these students earn high scores when local norms

TABLE 1.1

Demographic Characteristics of Nine Schools Used to Create Local Norms for DIBELS Measure of Oral Reading Fluency

	Schools								
Characteristic	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
M	19.17	23.47	23.70	26.55	27.61	33.61	35.30	43.08	60.59
SD	20.37	22.88	22.97	25.30	29.08	30.15	26.35	34.14	44.04
N	63	72	90	77	57	83	61	84	96
% Black	83	91	16	21	12	56	87	51	22
% Hispanic	2	3	42	29	54	8	4	4	1
% White	14	6	40	48	27	34	8	39	73
% Other	1	0	2	2	7	2	1	6	4
% Free or reduced lunch	99	99	99	65	99	56	99	56	25

Note. DIBELS = Dynamic Indicators of Basic Early Literacy Skills.

TABLE 1.2

Calibration of Standard Scores Using Local Norms by School and Using DIBELS Reference Group for Fall of Second Grade

	Standard scores for each school (*M* = 100, *SD* = 15)									National
Raw score	1	2	3	4	5	6	7	8	9	reference
60	130	124	124	120	117	113	114	107	100	102
58	129	123	122	119	116	112	113	107	99	101
56	127	121	121	117	115	111	112	106	98	100
54	126	120	120	116	114	110	111	105	98	100
52	124	119	118	115	113	109	110	104	97	99
50	123	117	117	114	112	108	108	103	96	98
48	121	116	116	113	111	107	107	102	96	98
46	120	115	115	112	109	106	106	101	95	97
44	118	113	113	110	108	105	105	100	94	96
42	117	112	112	109	107	104	104	100	94	95
40	115	111	111	108	106	103	103	99	93	94
38	114	110	109	107	105	102	102	98	92	93
36	112	108	108	106	104	101	100	97	92	93
34	111	107	107	104	103	100	99	96	91	92
32	109	106	105	103	102	99	98	95	90	91
30	108	104	104	102	101	98	97	94	90	90
28	106	103	103	101	100	97	96	93	89	89
26	105	102	102	100	99	96	95	92	88	87
24	104	100	100	98	98	95	94	92	88	86
22	102	99	99	97	97	94	92	91	87	85
20	101	98	98	96	96	93	91	90	86	84
18	99	96	96	95	95	92	90	89	85	82
16	98	95	95	94	94	91	89	88	85	82
14	96	94	94	93	93	90	88	87	84	80
12	95	92	92	91	92	89	87	86	83	78
10	93	91	91	90	91	88	86	85	83	75
8	92	90	90	89	90	87	84	85	82	74
6	90	89	88	88	89	86	83	84	81	72
4	89	88	87	87	88	85	82	83	81	69
2	87	87	86	85	87	84	81	82	80	65

Note. DIBELS = Dynamic Indicators of Basic Early Literacy Skills.

are used, fewer would be identified as being in need of instruction and fewer would be provided good instruction as a result. This approach would, therefore, result in considerable educational inequality.

Monitoring Progress

An essential goal of the CBM approach is the examination of changes over time. Research studies that seek to evaluate the effects of educational interventions need to consider the potential confound resulting from the issue of natural maturation in numbers of words a student can read per minute. That is, evaluation research often involves comparison of pretest and posttest scores, but with raw scores, such as those for reading fluency, it is unclear how much of the pretest–posttest change is associated with the intervention and how much is attributable to normal growth and learning. The value of the normalizing raw score to percentile rank conversion is that age-related changes are controlled because the student's score is calibrated in relation to a similarly aged comparison group. However, to avoid the problems associated with analyzing percentile ranks (see Anastasi & Urbina, 1997), converting percentiles to standard scores retains rank and improves the psychometric qualities of the scores. To better understand the influence raw scores can have on evaluation of pretest–posttest treatment, I examined scores analysis of a data set containing DIBELS ORF scores in an intervention study.

The data used included 352 second-grade boys and girls enrolled in 12 public elementary schools located in the southern Atlantic region of the United States. The experimental (*N* = 136) and the control (*N* = 216) groups consisted of students from six schools whose reading skills were tested at the start of school and in the middle of the school year. The students in the experimental group participated in an online reading program called Ramps to Reading (see Naglieri & Pickering, 2010, for a description), and the control group was not exposed to Ramps to Reading at all. The demographic characteristics of the schools and students in the experimental and control groups were similar, but the schools that made up the experimental group were represented by high percentages of individuals receiving free and reduced lunch and African Americans. All schools in the experimental group met criteria for Title I funding. What is most important, however, is that in this study both ORF raw scores and standard scores (obtained by converting the percentile scores to standard scores having a mean of 100 and a standard deviation of 15) were obtained for the groups. Additionally, each group was further divided on the basis of having initial DIBELS ORF scores that were described as low risk, some risk, or at risk. The results were quite informative.

Table 1.3 provides the means, standard deviations, sample sizes, and effect sizes for the various groups using ORF raw scores and standard scores. For the experimental group, one sees effect sizes based on raw scores of 1.5, 2.8, and 1.7 for the low-risk, some-risk, and at-risk groups. These values are very large, as are the effect sizes for the control group (1.1, 2.2, and 1.2). Because growth in raw scores of words read per minute over the course of time is considerable, as previously discussed (see Figure 1.2), these effect sizes can be considered to be inflated. Examination of the standard scores, which calibrate standing relative to a comparison group, suggest a far different result. The control group's pretest–posttest differences were essentially zero, but the experimental group showed small to medium effect-size changes. Thus, the method of calibration of the raw scores had a direct impact on the interpretation of the intervention's effectiveness. Using scores that take into account developmental changes that occur over time inflated the effect sizes for both groups, and only when age-related changes were controlled did a more realistic finding result.

What Now?

The current state of achievement testing in school psychology can be described as having too much variance. The psychometric quality of measures used today ranges from marginal to excellent. As with assessment of other constructs, disorders, and

TABLE 1.3

Effectiveness of a Reading Intervention on the Basis of Comparisons of Raw Scores and Standard Scores

		Raw scores					Standard scores				
		Pretest		Posttest			Pretest		Posttest		
Group	*N*	*M*	*SD*	*M*	*SD*	Effect size	*M*	*SD*	*M*	*SD*	Effect size
Experimental											
Low risk	61	67.7	19.6	98.4	21.8	1.5	104.5	6.8	106.8	7.7	0.3
Some risk	37	33.8	4.7	60.3	12.7	2.8	91.5	2.3	93.4	4.4	0.6
At risk	38	16.1	6.8	33.2	12.3	1.7	80.4	5.7	83.4	5.3	0.5
Control											
Low risk	145	71.4	19.8	93	20.7	1.1	105.9	7	105	7.4	−0.1
Some risk	43	33.5	5.2	53.3	11.8	2.2	91.3	2.5	91.1	4.1	−0.1
At risk	28	15.4	8.7	27.5	11.9	1.2	79.3	7.6	80.6	5.8	0.2

abilities, for example, autism spectrum disorders (see Naglieri & Chambers, 2009), the options range considerably, and practitioners have to choose wisely between tools to obtain scores they can use with confidence. As far as the information provided in this section is concerned, practitioners must be particularly cautious when using very short measures of skills, such as using reading fluency as a predictor of reading, and when using raw scores for (a) evaluating current status and (b) evaluating changes over time. The best option remains using well-normed tests that assess academic skills directly, especially those that provide strong psychometric quality and norms for calibrating growth. Put simply, the use of raw scores for the calibration of academic skills and progress monitoring is not good science.

SOCIAL–EMOTIONAL STATUS

Evaluation of emotional status has been dominated by projective tests and rating scales. As with the assessment of achievement and intelligence, the evaluation of emotional well-being has also been evolving in the areas of both individual and universal assessment. This evolution has been driven by efforts to focus on positive attributes (so-called social–emotional strengths related to resilience) instead of, or in addition to, emotional or behavioral disorders and psychopathology. Emphasis on social–emotional strengths that are related to resilience and particularly on universal screening has come from governmental agencies, professional organizations, and practitioners in fields such as psychology, sociology, and education. For example, in 2003 the President's New Freedom Commission on Mental Health urged that early mental health screening and assessment services be routinely conducted and that school district personnel ensure the mental health care of children. In 2010, the National Association of School Psychologists published its model for comprehensive and integrated school psychological services, which addressed the delivery of school psychological services within the context of educational programs and educational settings. The model states that school psychologists should have knowledge of principles and research related to resilience and risk factors that are important for learning and mental health. Additionally, the National Association of School Psychologists' position contended that school psychologists should be involved in universal screening programs to identify students in need of support services to ensure learning and promote social–emotional skills and resilience. Clearly, the field is moving toward assessment of social–emotional strengths as well as psychopathology.

The emphasis on assessment and interventions for social–emotional competence is important for several reasons. First, at any given time about 20% of children and adolescents are estimated to have a diagnosable emotional or behavioral disorder that interferes with learning (Doll, 1996). Second, emerging research has suggested that social–emotional competence underlies school success (Payton et al., 2008). Third, state departments of education have adopted or are in the process of developing social–emotional learning standards that could lead to (a) universal screening of social–emotional skills and (b) social–emotional skills instruction within the regular education curriculum. This approach, as with any assessment and intervention approach, requires reliable and valid tools for assessing and monitoring social–emotional competencies (see Goldstein & Brookes, 2005).

Progress has been made in recent years as evidenced by the availability of published rating scales to measure protective factors that measure children's social–emotional strengths related to resilience. Sometimes social–emotional strengths have been integrated into scales that also include problem behaviors related to emotional or behavioral disturbance. For example, Bracken and Keith (2004) included specific scales related to serious emotional disturbance and social maladjustment as well as both clinical and adaptive (e.g., social skills) scales using items that are designed to identify children and adolescents in need of behavioral, educational, or psychiatric treatments. Similarly, the Behavior Assessment System for Children, Second Edition (Reynolds & Kamphaus, 2004), measures adaptive and maladaptive behavior. Using all positively worded items for assessment of social–emotional strengths and behavioral needs, LeBuffe and Naglieri (2003) published the Devereux Early Childhood

Assessment—Clinical Form. These three scales illustrate how measures of social–emotional problems and strengths can be combined into one rating scale. The authors of other scales, however, conceptualized evaluation of mental health using a different approach—assessment of social–emotional factors related to resilience.

The Resiliency Scales for Children and Adolescents (Prince-Embury, 2005) measure areas of perceived strength and vulnerability related to psychological resilience along three dimensions (sense of mastery, sense of relatedness, and emotional reactivity). Using a similar approach, researchers at the Devereux Center for Resilient Children have published a comprehensive system made up of several measures of factors related to resilience that vary across ages and purposes. For example, the Devereux Early Childhood Assessment for Infants and Toddlers (Mackrain, LeBuffe, & Powell, 2007) and the Devereux Early Childhood Assessment (LeBuffe & Naglieri, 1999) are designed to measure social–emotional strengths of young children. The Devereux Student Strengths Assessment (LeBuffe, Shapiro, & Naglieri, 2009) was developed for children from kindergarten through eighth grade; each of these is a thorough measure with many items. In contrast, the Devereux Student Strengths Assessment—Mini (Naglieri, LeBuffe, & Shapiro, 2011) is an eight-item scale of social–emotional strengths for universal screening. The availability of carefully developed measures of protective factors related to resilience offers the opportunity to examine validity questions related to these new instruments, especially as they may be used for universal screening.

The availability of new scales built on the concept of social–emotional strengths using a perspective described as strengths based is clearly an important development in the assessment of mental health. An evolution has also occurred in the assessment of psychological and behavioral disorders, especially as it relates to the use of raw scores and comparison groups, as discussed earlier in this chapter for achievement tests. More specifically, in some contexts, for example, identification of specific psychological disorders such as autism, researchers are using a specific reference group for calibration of scores instead of using a nationally representative reference group.

Naglieri and Chambers (2009) summarized the characteristics of rating scales used to assess behaviors associated with autism and examined the psychometric qualities that such measures possess. They concluded that the methods used to develop rating scales differed considerably in their approaches to instrument development. For example, some of the scales are very short (e.g., 15 items), and others contain many items (e.g., about 90 items). Some authors provided only raw scores, which makes interpretation difficult, and only two scales provided standard scores (*T* scores). Although some rating scales provide derived scores, the samples on which they were based were the particular group the scale was intended to identify. Raters obtained a score that tells how similar the individual being assessed is to those the scale is intended to identify, for example, those with an autism spectrum disorder (ASD). Of all the scales Naglieri and Chambers summarized, only one used a national comparison sample; all the others used samples of individuals who had or were referred for autism. The question of the utility of a comparison group consisting of children referred for or having the disorder of interest needs to be addressed. I consider this issue next using data from a recent project involving the Autism Spectrum Rating Scale (Goldstein & Naglieri, 2009).

An understanding of the differences between using a nationally representative sample and a sample of children identified as having autism as a reference group is best examined empirically. To do so, Goldstein and Naglieri (2009) constructed a raw-score-to-standard-score (*T*-scores) conversion table on the basis of a sample of children with ASD (N = 243) who were diagnosed with autism (n = 137), Asperger syndrome (n = 80), or pervasive developmental disorder—not otherwise specified (n = 26). This sample was made up of individuals with a single primary diagnosis made by a qualified professional (e.g., psychiatrist, psychologist) according to the *DSM–IV–TR* (APA, 2000) or the International Statistical Classification of Diseases and Related Health Problems, 10th Revision (World Health Organization, 2007) using appropriate methods (e.g., record review, rating scales, observation, and interview). The sample, representative of the U.S. population, included boys and girls from each of the four geographic regions of the United States and four racial–ethnic groups (Asian, Black, White–not Hispanic, and Hispanic origin) ages 6 to 18 years. The sample size was 1,828. (See Goldstein & Naglieri, 2009, for more details about the normative sample of the Autism Spectrum Rating Scale and those identified with ASD.)

Table 1.4 provides a raw-score-to-*T*-score conversion table based on the descriptive statistics for the ASD (*N* = 243, *M* = 129.1, *SD* = 46.9) and national (*N* = 1,828, *M* = 53.1, *SD* = 36.1) reference groups. It is clear from an examination of this table that a raw score of 130 yielded very different scores for the two samples. A raw score of 130 yielded a *T* score of 50 for the ASD comparison group and a *T* score of 71 for the national comparison group. A raw score of 80 yielded a *T* score of 40 (1 standard deviation below the mean) for the ASD group and a *T* score of 57 (nearly 1 standard deviation above the mean) for the national comparison group. These results illustrate how different conclusions may be reached when the same rating scale is calibrated against two different samples.

TABLE 1.4

Comparison of *T* Scores Based on a Sample of Individuals With Autism (*N* = 243) and a National Comparison Group (*N* = 1,828)

Raw score	ASD comparison	National comparison
170	59	82
165	58	81
160	57	80
155	56	78
150	54	77
145	53	75
140	52	74
135	51	73
130	50	71
125	49	70
120	48	69
115	47	67
110	46	66
105	45	64
100	44	63
95	43	62
90	42	60
85	41	59
80	40	57
75	38	56
70	37	55
65	36	53
60	35	52
55	34	51
50	33	49
45	32	48
40	31	46
35	30	45
30	29	44
25	28	42

Note. ASD = autism spectrum disorder.

CONCLUSIONS

The field of assessment in school psychology, as in other areas of psychology, is changing. This chapter has focused on three main issues related to measurement, test development, and norming of scores. These issues are important at both theoretical and practical levels. Theoretically, the need for intelligence tests to be firmly grounded in a theory of intelligence, preferably one that is multidimensional, is increasingly apparent. These separate cognitive abilities need to be well examined and, insofar as identification of special populations is concerned, different ability profiles should be related to different academic performance patterns. In the field of skills assessment, in which tests are structured according to academic content rather than some underlying theoretical concept, it is clear that the validity of CBM measures warrants considerable research, particularly in regard to the validity of test score interpretation and normative versus true academic growth. Finally, in the area of emotional status, assessment of social–emotional strengths offers important advantages to traditional methods based on behavioral manifestations of psychopathology. The validity of this change in perspective also warrants more research. In summary, practitioners and researchers alike need to be mindful of the need to take a scientific perspective on the strengths and weaknesses of these various approaches to assessment, ask the important reliability and validity questions, and follow the research to make good decisions about which tests to use and for what purposes.

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

Anastasi, A., & Urbina, S. (1997). *Psychological testing.* Upper Saddle River, NJ: Prentice Hall.

Bracken, B. A., & Keith, L. (2004). *Clinical Assessment of Behavior.* Lutz, FL: Psychological Assessment Resources.

Canivez, G. L., & Gaboury, A. R. (2010, August). *Cognitive assessment system construct and diagnostic utility in assessing ADHD.* Paper presented at the 118th Annual Convention of the American Psychological Association, San Diego, California.

Carroll, J. B. (2000). Commentary on profile analysis. *School Psychology Quarterly, 15*, 449–456. doi:10.1037/h0088800

Davis, F. B. (1959). Interpretation of differences among averages and individual test scores. *Journal of Educational Psychology, 50*, 162–170. doi:10.1037/h0044024

Davison, M. L., & Kuang, H. (2000). Profile patterns: Research and professional interpretation. *School Psychology Quarterly, 15*, 457–464. doi:10.1037/h0088801

Doll, B. (1996). Prevalence of psychiatric disorders in children and youth: An agenda for advocacy by school psychology. *School Psychology Quarterly, 11*, 20–467. doi:10.1037/h0088919

Fiorello, C. A., Hale, J. B., & Snyder, L. E. (2006). Cognitive hypothesis testing and response to intervention for children with reading disabilities. *Psychology in the Schools, 43*, 835–853. doi:10.1002/pits.20192

Flanagan, D. P., & Kaufman, A. S. (2004). *Essentials of WISC-IV assessment.* Hoboken, NJ: Wiley.

Flanagan, D. P., Ortiz, S. O., Alfonso, V. C., & Mascolo, J. (2002). *The achievement test desk reference (ATDR): Comprehensive assessment and learning disabilities.* Boston, MA: Allyn & Bacon.

Fletcher, J. M., Denton, C., & Francis, D. J. (2005). Validity of alternative approaches for the identification of learning disabilities: Operationalizing unexpected underachievement. *Journal of Learning Disabilities, 38*, 545–552. doi:10.1177/00222194050380061101

Goldstein, S., & Brooks, R. (Eds.). (2005). *Handbook of resilience in children.* New York, NY: Kluwer/Academic Press. doi:10.1007/b107978

Goldstein, S., & Naglieri, J. A. (2009). *Autism Spectrum Rating Scale.* Toronto, Ontario, Canada: Multi-Health Systems.

Hale, J. B., & Fiorello, C. A. (2004). *School neuropsychology: A practitioner's handbook.* New York, NY: Guilford Press.

Hale, J. B., Kaufman, A. S., Naglieri, J. A., & Kavale, K. A. (2006). Implementation of IDEA: Using RTI and cognitive assessment methods. *Psychology in the Schools, 43*, 753–770. doi:10.1002/pits.20186

Huang, L. V., Bardos, A. N., & D'Amato, R. C. (2010). Identifying students with learning disabilities: Composite profile analysis using the Cognitive Assessment System. *Journal of Psychoeducational Assessment, 28*, 19–30. doi:10.1177/0734282909333057

Individuals With Disabilities Education Improvement Act of 2004, P.L. 108–446, 20 U.S.C. § 1400 *et seq.*

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Kaufman, A. S. (1979). *Intelligent testing with the WISC-R.* New York, NY: Wiley.

Kaufman, A. S., & Kaufman, N. L. (Eds.). (2001). *Learning disabilities: Psychological assessment and evaluation.* Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511526794

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children* (2nd ed.). San Antonio, TX: Pearson.

Kavale, K. A., Kaufman, A. S., Naglieri, J. A., & Hale, J. B. (2005). Changing procedures for identifying learning disabilities: The danger of poorly supported ideas. *School Psychologist, 59*, 16–25.

Koehler-Hak, K. M., & Bardos, A. N. (2009). Dynamic Indicators of Basic Early Literacy Skills (DIBELS): General outcomes measurement for prevention and remediation of early reading problems. In J. A. Naglieri & S. Goldstein (Eds.), *A practitioner's guide to assessment of intelligence and achievement* (pp. 389–416). New York, NY: Wiley.

LeBuffe, P. A., & Naglieri, J. A. (1999). *Devereux Early Childhood Assessment.* Lewisville, NC: Kaplan Press.

LeBuffe, P. A., & Naglieri, J. A. (2003). *Devereux Early Childhood Assessment—Clinical form.* Lewisville, NC: Kaplan Press.

LeBuffe, P. A., Shapiro, V. B., & Naglieri, J. A. (2009). *Devereux Student Strengths Assessment.* Lewisville, NC: Kaplan Press.

Mackrain, M., LeBuffe, P., & Powell, G. (2007). *Devereux Early Childhood Assessment for Infants and Toddlers.* Lewisville, NC: Kaplan Press.

McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment, 8*, 290–302. doi:10.1177/073428299000800307

Meyer, M. S. (2000). The ability–achievement discrepancy: Does it contribute to an understanding of learning disabilities? *Educational Psychology Review, 12*, 315–337. doi:10.1023/A:1009070006373

Naglieri, J. A. (1999). *Essentials of CAS assessment.* New York, NY: Wiley.

Naglieri, J. A. (2000). Can profile analysis of ability test scores work? An illustration using the PASS theory and CAS with an unselected cohort. *School Psychology Quarterly, 15*, 419–433. doi:10.1037/h0088798

Naglieri, J. A. (2011). The discrepancy/consistency approach to SLD identification using the PASS theory. In D. P. Flanagan & V. C. Alfonso (Eds.), *Essentials of specific learning disability identification* (pp. 145–172). Hoboken, NJ: Wiley.

Naglieri, J. A., & Chambers, K. (2009). Psychometric issues and current scales for assessing autism spectrum disorders. In S. Goldstein, J. A. Naglieri, & S. Ozonoff (Eds.), *Assessment of autism spectrum disorders* (pp. 55–90). New York, NY: Springer.

Naglieri, J. A., & Das, J. P. (1997). *Cognitive Assessment System.* Itasca, IL: Riverside.

Naglieri, J. A., & Goldstein, S. (2009). *Assessment of intelligence and achievement: A practitioner's guide.* New York, NY: Wiley.

Naglieri, J. A., & Goldstein, S. (2011). Assessment of cognitive and neuropsychological processes. In S. Goldstein & J. A. Naglieri (Eds.), *Understanding and managing learning disabilities and ADHD in late adolescence and adulthood* (2nd ed., pp. 137–160). New York, NY: Wiley.

Naglieri, J. A., & Gottling, S. H. (1995). A cognitive education approach to math instruction for the learning disabled: An individual study. *Psychological Reports, 76*, 1343–1354. doi:10.2466/pr0.1995.76.3c.1343

Naglieri, J. A., & Gottling, S. H. (1997). Mathematics instruction and PASS cognitive processes: An intervention study. *Journal of Learning Disabilities, 30*, 513–520. doi:10.1177/002221949703000507

Naglieri, J. A., & Johnson, D. (2000). Effectiveness of a cognitive strategy intervention to improve math calculation based on the PASS theory. *Journal of Learning Disabilities, 33*, 591–597. doi:10.1177/002221940003300607

Naglieri, J. A., LeBuffe, P. A., & Shapiro, V. (2011). *Devereux Student Strengths Assessment—Mini.* Lewisville, NC: Kaplan Press.

Naglieri, J. A., & Otero, T. (2011). Cognitive Assessment System: Redefining intelligence from a neuropsychological perspective. In A. Davis (Ed.), *Handbook of pediatric neuropsychology* (pp. 320–333). New York, NY: Springer.

Naglieri, J. A., Otero, T., DeLauder, B., & Matto, H. (2007). Bilingual Hispanic children's performance on the English and Spanish versions of the Cognitive Assessment System. *School Psychology Quarterly, 22*, 432–448. doi:10.1037/1045-3830.22.3.432

Naglieri, J. A., & Pickering, E. (2010). *Helping children learn: Intervention handouts for use in school and at home* (2nd ed.). Baltimore, MD: Brookes.

National Association of School Psychologists. (2010). National Association of School Psychologists model for comprehensive and integrated school psychological services. *School Psychology Review, 39*, 320–333.

New Freedom Commission on Mental Health. (2003). *Achieving the promise: Transforming mental health care in America: Final report* (No. SMA03-3832). Rockville, MD: U.S. Department of Health and Human Services.

O'Donnell, L. (2009). The Wechsler Intelligence Scale for Children. In J. A. Naglieri & S. Goldstein (Eds.), *A practitioner's guide to assessment of intelligence and achievement* (4th ed., pp. 153–190). New York, NY: Wiley.

Payton, J., Weissberg, R. P., Durlak, J. A., Dymnicki, A. B., Taylor, R. D., Schellinger, K. B., & Pachan, M. (2008). *The positive impact of social and emotional learning for kindergarten to eighth grade students: Findings from three scientific reviews.* Chicago, IL: Collaborative for Academic, Social, and Emotional Learning.

Prince-Embury, S. (2005). *Resiliency Scales for Children and Adolescents.* San Antonio, TX: Pearson Education.

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47*, 427–469. doi:10.1016/j.jsp.2009.07.001

Roid, G. (2003). *Stanford-Binet* (5th ed.). Itasca, IL: Riverside.

Roid, G., & Bos, J. (2009). Achievement assessment and progress monitoring with the Wide Range Achievement Test. In J. A. Naglieri & S. Goldstein (Eds.), *A practitioner's guide to assessment of intelligence and achievement* (4th ed., pp. 537–572). New York, NY: Wiley.

Roid, G. H., & Tippin, S. M. (2009). Assessment of intellectual strengths and weaknesses with the Stanford-Binet Intelligence Scales—Fifth Edition (SG5). In J. A. Naglieri & S. Goldstein (Eds.), *A practitioner's guide to assessment of intelligence and achievement* (pp. 127–152). New York, NY: Wiley.

Schofield, N. J., & Ashman, A. F. (1986). The relationship between Digit Span and cognitive processing across ability groups. *Intelligence, 10*, 59–73. doi:10.1016/0160-2896(86)90027-9

Shinn, M. R., Tindal, G. A., & Stein, S. (1988). Curriculum-based measurement and the identification of mildly handicapped students: A research review. *Professional School Psychology, 3*, 69–85. doi:10.1037/h0090531

Silverstein, A. B. (1993). Type I, Type II, and other types of errors in pattern analysis. *Psychological Assessment, 5*, 72–74. doi:10.1037/1040-3590.5.1.72

Stanovich, K. E. (1994). Are discrepancy-based definitions of dyslexia empirically defensible? In K. P. van den Bos, L. S. Siegel, D. J. Bakker, & D. L. Share (Eds.), *Current directions in dyslexia research* (pp. 15–30). Lisse, the Netherlands: Swets & Zeitlinger.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children* (4th ed.). San Antonio, TX: Psychological Corporation.

World Health Organization. (2007). *International statistical classification of diseases and related health problems, 10th revision.* Geneva, Switzerland: Author.

Wendling, B. J., Mather, N., & Shrank, F. A. (2009). Woodcock-Johnson III Tests of Cognitive Abilities. In J. A. Naglieri & S. Goldstein (Eds.), *A practitioner's guide to assessment of intelligence and achievement* (pp. 191–229). New York, NY: Wiley.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2007). *Woodcock-Johnson III Tests of Cognitive Abilities.* Itasca, IL: Riverside.