INTRODUCTION AND OVERVIEW

A decade has passed since the Differential Ability Scales (DAS; Elliott, 1990a) was introduced to the stable of tests for assessing cognitive abilities in children. For many clinicians, weaned on a diet of Wechsler and Stanford-Binet scales, the DAS remains unfamiliar ground. However, with its strong psychometric properties, emphasis on profiling distinct cognitive abilities, and high appeal to young children, the DAS has much to offer. In this chapter, we wish to argue that the DAS is an extremely valuable addition to a clinician's set of tools, both for assessing the learning and developmental problems with which children frequently present and for translating those assessment results into practical recommendations.

Theoretical Perspectives

The DAS was developed based on an eclectic blend of theories about cognitive development ranging from those of Spearman (1923) to those of J. P. Das (Das, Kirby, & Jarman, 1979). There are several distinguishing theoretical characteristics about the test.

It is no accident the DAS is named the Differential Ability Scales and not
the “Differential Intelligence Scales.” Elliott (1990b) rejects the possibility that any test can measure “intelligence,” saying that the term has been so vaguely and variously defined as to have become practically meaningless. He quotes Spearman (1927), who described intelligence as “a word with so many meanings that finally it has none” (p. 14). Instead, Elliott sets the goal of the DAS so as to measure a range of “cognitive abilities.” Cognitive ability refers to a more specific and narrower domain of human cognition than the term intelligence. For all age levels from 3 years, 6 months and upward, six subtests, measuring a number of cognitive abilities, form an overall composite score reflecting Spearman’s g. Psychometric g is defined by Elliott (1990c) as “the general ability of an individual to perform mental processing that involves conceptualization and transformation of information” (p. 20). From this definition comes the term used to describe the most general DAS composite score: General Conceptual Ability (GCA).

However, the raison d’être of the DAS is not just to be another measure of children’s overall cognitive ability or intelligence. Instead, Elliott (1990c) intends it to be an assessment tool that creates a profile of a child’s strengths and weaknesses across a range of distinct cognitive abilities. To accomplish this, a major focus in the development of the DAS was to create a battery of subtests with sufficiently high specificity and reliability as to allow the clinician to interpret each as measuring something unique and distinct from the others. While the DAS provides meaningful information about more general factors, the intended focus for the clinician’s interpretive efforts is at the subtest or lower-order composite level. This shift, to analysis at the subtest level and away from focusing on general ability, reflects the theoretical tradition of Thurstone (1938) and his concept of a number of specific “primary mental abilities.” This tradition has been continued in the more recent work of Cattell and Horn (e.g., Cattell, 1971; Horn & Cattell, 1966; Horn, 1985; Horn & Noll, 1997) and Carroll (1993, 1997). The hope is that the creation of a reliable profile of a child’s cognitive strengths and weaknesses will lead to a better understanding of his or her learning difficulties, and ultimately to practical recommendations to classroom teachers about remediation (Elliott, 1990c).

The structure of the DAS assumes a hierarchical organization of cognitive ability (Elliott, 1990b). Subtests, or specific measures of distinct abilities, make up the base of this structure. However, since all ability measures are intercorrelated, these subtests will tend to group together at a second, higher, level in what are referred to as clusters. These clusters, in turn, are interrelated, which yields an estimate of psychometric g that is at the apex of this hierarchical structure. Factor analyses of DAS data show that this hierarchy becomes more differentiated as a child develops (Elliott, 1990c).

While the development and structure of the DAS is not formally linked to the Horn-Cattell model, the ability clusters proposed in the DAS are interpretable in terms of the Horn-Cattell factors. For preschool-age children,
abilities are seen as clustering into verbal and nonverbal factors, which are similar to Gc and Gv in the Horn-Cattell theory. For school-age children, a third cluster of cognitive abilities becomes differentiated, which reflects fluid reasoning (Gf); at its heart is the ability to integrate complex verbal and spatial information. The function of this third cluster is very similar to that of "planning" in the Luria and J. P. Das tradition (Das et al., 1979).

Finally, the DAS is influenced by research from cognitive psychology and neuropsychology (McCarthy & Warrington, 1990) in its approach to memory abilities. This research suggests that visual and verbal short-term memory are quite distinct and doubly dissociated. Because of this, the DAS does not represent short-term memory as a separate and unitary cluster such as is attempted by the Short-term Memory Area Composite on the Stanford-Binet Intelligence Scale—Fourth Edition (Thorndike, Hagen, & Sattler, 1986). Instead, auditory and visual short-term memory subtests on the DAS are partitioned as separate measures that do not cluster together.

**NATURE OF THE TEST**

The DAS is an individually administered measure of cognitive abilities, standardized for children between 2 years, 6 months and 17 years, 11 months. It is based on the British Ability Scales (BAS; Elliott, Murray, & Pearson, 1979, Elliott, 1997a). The DAS is essentially two test batteries. The first is geared to preschoolers from age 2 years, 6 months to 5 years, 11 months. The second is designed for school-age children from age 6 years, 0 months to 17 years, 11 months. Although these are the nominal age ranges of the two batteries, they were completely co-normed across the age range 5 years, 0 months through 7 years, 11 months, and may be used equivalently in that age range. Thus, the school-age battery may be used appropriately to assess relatively gifted 5-year-olds. Most important, for the majority of clinicians, the preschool battery may be used to assess average and lower functioning 6- and 7-year-olds for whom the school-age materials are developmentally less appropriate.

The preschool and school-age batteries are similar in structure. Both feature a set of core subtests designed to measure general cognitive ability. The six core subtests in each battery were selected because of their relatively high correlation with \( g \); they thus measure more complex mental processing. Both the preschool and school-age batteries also contain a set of diagnostic subtests. These are intended to measure more specific or distinct skills such as short-term memory or speed of information processing. They have a lower correlation with \( g \), and measure less complex mental processing.

The preschool battery of the DAS is further divided into an upper and lower preschool level. The lower level is specifically designed for children from 2 years, 6 months to 3 years, 5 months. The upper preschool level is
used for assessing children from 3 years, 6 months to 7 years, 11 months. Table 3.1 provides information on the core subtests for the preschool battery while Table 3.2 provides information on the diagnostic subtests that can be used with preschoolers and early school-age children.

The core subtests of the upper preschool level consist of all of the subtests listed in Table 3.1, with the exception of Block Building. These six core subtests combine to provide an estimate of overall cognitive ability (GCA). They are also divided into verbal and nonverbal clusters or factors as illustrated in Table 3.1. The Verbal Ability cluster is designed to measure acquired verbal concepts and knowledge. The Nonverbal Ability cluster is designed to measure complex nonverbal (primarily spatial) reasoning abilities. The sixth core subtest, Early Number Concepts, is not grouped with either cluster. Originally hypothesized as a part of the Verbal Ability cluster, factor analyses revealed Early Number Concepts loading significantly on both the verbal and nonverbal factors (Elliott, 1990c). Because of the subtest's high correlation with g and the important clinical content it covers, it was decided to retain Early Number Concepts as part of the core subtests measuring overall cognitive ability. All of the diagnostic subtests can be used with children at the

| TABLE 3.1 |
| Preschool Subtests of the Differential Ability Scales |

<table>
<thead>
<tr>
<th>Subtest</th>
<th>What the child does</th>
<th>Cluster</th>
<th>Usual age range</th>
<th>Extended/Out-of-level age range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block Building</td>
<td>Uses wooden blocks to copy designs from a model.</td>
<td>None</td>
<td>2:6-3:5</td>
<td>3:6-4:11</td>
</tr>
<tr>
<td></td>
<td>2. Follows verbal directions with toys.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Picture Similarities</td>
<td>Matches a target picture to the most similar of four options.</td>
<td>Nonverbal</td>
<td>2:6-5:11</td>
<td>6:0-7:11</td>
</tr>
<tr>
<td>Naming Vocabulary</td>
<td>Names pictures.</td>
<td>Verbal</td>
<td>2:6-5:11</td>
<td>6:0-8:11</td>
</tr>
<tr>
<td>Pattern Construction</td>
<td>Builds designs from pictures or models using squares or cubes with colored patterns.</td>
<td>Nonverbal</td>
<td>3:6-17:11</td>
<td>3:0-3:5</td>
</tr>
<tr>
<td></td>
<td>2. Answers questions about numerical and mathematical concepts.</td>
<td></td>
<td></td>
<td>6:0-7:11</td>
</tr>
<tr>
<td>Copying</td>
<td>Copies line drawings with a pencil.</td>
<td>Nonverbal</td>
<td>3:6-5:11</td>
<td>6:0-7:11</td>
</tr>
</tbody>
</table>
### TABLE 3.2
Diagnostic Subtests of the Differential Ability Scales

<table>
<thead>
<tr>
<th>Subtest</th>
<th>What the child does</th>
<th>Usual age range</th>
<th>Out-of-level age range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall of Objects—Immediate</td>
<td>Looks at a card with 20 pictured objects for 60 or 20 seconds. The card is then taken away and the child names as many objects as he or she can remember. This is repeated over 3 trials.</td>
<td>4:0–17:11</td>
<td>None</td>
</tr>
<tr>
<td>Recall of Objects—Delayed</td>
<td>Between 10 and 30 minutes after the Recall of Objects—Immediate subtest, the child is asked to name as many of the pictured objects as he or she can still remember.</td>
<td>4:0–17:11</td>
<td>None</td>
</tr>
<tr>
<td>Matching Letter-Like Forms</td>
<td>Finds an exact match to an abstract drawing from six pictured options. The five incorrect options are reversals or rotations of the target.</td>
<td>4:6–5:11</td>
<td>4:0–4:5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6:0–7:11</td>
<td></td>
</tr>
<tr>
<td>Recall of Digits</td>
<td>Repeats back a sequence of numbers.</td>
<td>3:0–17:11</td>
<td>2:6–2:11</td>
</tr>
<tr>
<td>Recognition of Pictures</td>
<td>Looks at a picture of one or several objects for 5 or 10 seconds. Then picks out those objects from a second picture that includes additional distractors.</td>
<td>3:0–7:11</td>
<td>2:6–2:11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8:0–17:11</td>
<td></td>
</tr>
<tr>
<td>Speed of Information Processing</td>
<td>1. Looks at a row of circles, each with a different number of boxes in it. Child marks the circle with the most boxes.</td>
<td>6:0–17:11</td>
<td>5:0–5:11</td>
</tr>
<tr>
<td></td>
<td>2. Looks at a row of numbers. Child marks the highest number.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The lower preschool level of the DAS consists of four core subtests: Block Building, Verbal Comprehension, Picture Similarities, and Naming Vocabulary. These subtests combine to produce an overall composite, or GCA, score. The lower preschool level is not differentiated into clusters, but it does yield a Special Nonverbal Composite based on the child's performance on the upper preschool level, primarily to provide more specific information on visual and verbal short-term memory abilities. The Matching Letter-Like Forms subtest provides information on a child's visual discrimination skills and may be relevant to the child's readiness to acquire reading and writing skills (Elliott, 1990c). The Speed of Information Processing subtest can be used to assess this skill only with 5-year-olds showing average and above abilities in this area.
Block Building and Picture Similarities subtests. The Recall of Digits and Recognition of Pictures diagnostic subtests can be used to explore the short-term memory abilities of most 2-year-olds and all 3-year-olds at the lower preschool level.

The school-age battery of the DAS consists of the core subtests outlined in Table 3.3 and the diagnostic subtests described in Table 3.2. The six core subtests combine to provide an estimate of overall cognitive ability (GCA) and are further divided into three clusters, as outlined in Table 3.3. The Verbal Ability cluster is designed to measure complex verbal mental processing. The Spatial Ability cluster measures complex visual-spatial processing. The Nonverbal Ability cluster consists of reasoning subtests that involve minimal verbal instructions and minimal verbal responses from the child. However, verbal mediation strategies are generally essential for solving these problems. Thus the Nonverbal Reasoning subtests require an integration of ver-

<table>
<thead>
<tr>
<th>Subtest</th>
<th>What the child does</th>
<th>Cluster</th>
<th>Usual age range</th>
<th>Extended/Out-of-level age range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall of Designs</td>
<td>Looks at an abstract line drawing for 5 seconds. Drawing is removed and the child draws it from memory with a pencil.</td>
<td>Spatial</td>
<td>6:0–17:11</td>
<td>5:0–5:11</td>
</tr>
<tr>
<td>Word Definitions</td>
<td>Provides a verbal definition of a word.</td>
<td>Verbal</td>
<td>6:0–17:11</td>
<td>5:0–5:11</td>
</tr>
<tr>
<td>Pattern Construction</td>
<td>Builds designs from pictures or models using squares or cubes with colored patterns.</td>
<td>Spatial</td>
<td>3:6–17:11</td>
<td>3:0–3:5</td>
</tr>
<tr>
<td>Matrices</td>
<td>Selects from four or six options the one that would best complete an abstract visual pattern.</td>
<td>Nonverbal</td>
<td>6:0–17:11</td>
<td>5:0–5:11</td>
</tr>
<tr>
<td>Similarities</td>
<td>Describes verbally how three related words are similar.</td>
<td>Verbal</td>
<td>6:0–17:11</td>
<td>5:0–5:11</td>
</tr>
<tr>
<td>Sequential &amp; Quantitative Reasoning</td>
<td>1. Completes a series of abstract figures by drawing in the missing figure. 2. Determines the relationship shared between two pairs of numbers and then uses it to provide the missing number of an incomplete pair.</td>
<td>Nonverbal</td>
<td>6:0–17:11</td>
<td>5:0–5:11</td>
</tr>
</tbody>
</table>
TABLE 3.4
Achievement Tests of the Differential Ability Scales

<table>
<thead>
<tr>
<th>Subtest</th>
<th>What the child does</th>
<th>Usual age range</th>
<th>Out-of-level age range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Number Skills</td>
<td>Solves math problems on a worksheet.</td>
<td>6:0–17:11</td>
<td>None</td>
</tr>
<tr>
<td>Spelling</td>
<td>Spells words dictated by the clinician.</td>
<td>6:0–17:11</td>
<td>None</td>
</tr>
<tr>
<td>Word Reading</td>
<td>Pronounces words presented on a card.</td>
<td>6:0–17:11</td>
<td>5:0–5:11</td>
</tr>
</tbody>
</table>

Elliott (1997b) considers that the three DAS clusters measure the following factors in the Horn-Cattell tradition: Verbal cluster (Crystallized, G_c); Spatial cluster (Visual Processing, G_v); and Non-verbal Reasoning cluster (Fluid, G_f).

The school-age battery also features three measures of academic achievement, as outlined in Table 3.4. The achievement subtests were co-normed with the cognitive subtests, thus allowing for analyses of ability/achievement discrepancies. The DAS manual (Elliott, 1990a) provides support for use of both simple-difference and regression models to explore whether assessment results suggest the possibility of a learning disability.

TECHNICAL INFORMATION

Standardization Sample

The DAS was standardized between 1987 and 1989 on a sample of 3475 children selected to match the 1988 U.S. Census on the stratification variables of age, gender, ethnicity, geographic location, and parent education level. The preschool sample was further stratified to match the census data according to the proportion of children who had attended a preschool educational program. Detailed census records were used in order that, at each age level of the test, the standardization sample matched the population at that particular age level in terms of the joint distributions of ethnicity, geographic location, and parent education level.

This is an unusually stringent constraint in sampling, and it is far more difficult to achieve than representative distributions on each stratification variable taken singly. This, and other procedural refinements in sampling, makes the DAS unique in its sampling accuracy (Elliott, 1997b). A sample of 175 children was used for each 6-month age interval at the preschool range, while a sample of 200 children was used at each age level of the school-age range (Elliott, 1990c).
Reliability

Internal consistency reliability is high at the level of the GCA and cluster scores (Elliott, 1990c). GCA score reliability ranges from an average of .90 at the lower preschool level to .94 at the upper preschool level, to .95 for the school-age battery. Cluster score reliability ranges from .88 to .92. Internal consistency at the subtest level is relatively strong, with some exceptions, with reliability coefficients ranging from .70 to .92. Of the 26 possible internal reliability coefficients at the subtest level, 17 were greater than .8 and 4 were greater than .9.

Test-retest reliability data suggest GCA and cluster scores are very stable (Elliott, 1990c). GCA test-retest scores correlate between .89 and .93 across the three age groups sampled, while the range is from .79 to .90 for cluster scores. At the subtest level, test-retest reliabilities range from .38 to .93 for preschoolers and from .47 to .97 for school-age children. The mean subtest reliability coefficient is .78 (Kamphaus, 1993). The practice effect on the GCA is roughly 3 points for preschoolers and 6 points for school-age children. Four DAS subtests, which require a significant amount of clinician judgment to score, were examined for interrater reliability. Coefficients were found to be .9 or above.

DAS scores were also evaluated in terms of specificity, or how much of the score variance is reliable and unique to that measure. The higher the specificity of a measure, the more confident the clinician can be about interpreting the score as measuring something unique in comparison to what is measured by the other tests in the battery. McGrew and Murphy (1995) argue that specificity is high when it accounts for 25% or more of the test's total variance and when it is greater than the error variance. Each DAS subtest meets these criteria for high specificity, with values ranging from .30 to .82 (Elliott, 1990c). Elliott (1997b) presents data showing that the DAS has significantly greater specificity than other popular measures of cognitive ability for children.

Validity

Construct validity for the DAS is supported by confirmatory and exploratory factor analyses supporting a 1-factor model at the lower preschool level, a 2-factor (verbal/nonverbal) model at the upper preschool level, and a 3-factor (verbal/nonverbal/spatial) model for school-age children (Elliott, 1990c). Keith's (1990) independent hierarchical confirmatory factor analyses reported consistent results that Elliott (1997b) found were essentially in agreement with the DAS data analyses given in the test handbook (Elliott, 1990c). Elliott (1997b) also reports joint factor analysis of the DAS and the WISC-R (Wechsler, 1974) that support a verbal/nonverbal/spatial-factor model for school-age children.
Elliott (1990c) also provides evidence supporting the convergent and discriminant validity of the DAS cluster scores. The Verbal Ability cluster score consistently correlates much higher with the verbal composite score than it does with the nonverbal composite score of other cognitive ability tests for children. Similarly, the Nonverbal Reasoning and Spatial Ability cluster scores correlate significantly lower with the verbal composite score of other tests than does the DAS Verbal Ability cluster score.

Evidence for the concurrent validity of the DAS is provided by studies (Wechsler, 1991; Elliott, 1990c) showing consistently high correlations between the GCA and the composite scores of other cognitive batteries such as the Wechsler Intelligence Scale for Children—Third Edition (WISC-III; Wechsler, 1991), the Wechsler Preschool and Primary Scale of Intelligence—Revised (WPPSI-R; Wechsler, 1989), and the Stanford-Binet Intelligence Scale—Fourth Edition (Thorndike et al., 1986). High correlations were also found between the DAS achievement tests and other group or individual achievement tests as well as with actual student grades (Elliott, 1990c).

Bias

Extensive effort was put into ensuring the fairness of the DAS across cultures. Test items were first reviewed for possible bias by a panel representing women and several ethnic groups, and, based on its recommendations, a number of items were changed or dropped from the test. To aid statistical analyses of bias, an additional 600 Hispanic and African-American children were tested along with the standardization sample in order that each test item could be analyzed for differential item difficulty across cultures. The children in this bias oversample also assisted in ensuring that test scoring rules reflected possible culture-specific responses from minority children. Finally, the analyses showed that there was no unfair bias against minority children in the ability of the DAS GCA score to predict school achievement (Elliott, 1990c).

Further analyses of these data have examined construct bias in the DAS (Keith, Quirk, Schartzer, & Elliott, 1999). This was accomplished by conducting hierarchical, multisample confirmatory factor analysis of the DAS standardization data, including data from the bias oversample. Results showed that the DAS measured the same constructs for all three ethnic groups (Black, White, and Hispanic) across the entire age range of 2½ through 17 years. Thus it was concluded that the DAS shows no construct bias across groups.

Administration

There are several notable differences in administering the DAS compared to most other cognitive batteries for children. The vast majority of DAS subtests
reject the traditional model of establishing basal and ceiling levels, which
presumes that the child would have passed all items before the basal and
failed all items after the ceiling. Instead, most DAS subtests are adminis-
tered using an adaptive or tailored testing model. The goal is to have the child work
primarily on items that are of moderate difficulty and reduce the number of
items administered that are either too hard or too easy for that child. To do
this, most DAS subtests designate a set of items that are typically appropri-
ate for most children at a given age level. The child completes all of the items
in that set, reaching a decision point. If the child experiences a reasonable
combination of success and failure within the item set (which is defined
specifically for each subtest), the clinician can now proceed to the next sub-
test. If the child experiences very little success on the initial item set, the
clinician follows up by dropping back and administering the items from the
next-easiest item set. If the child reaches the initial decision point and ex-
periences almost complete success, the clinician continues on with more
difficult items until reaching a decision point where the child has begun to
experience some degree of difficulty, usually defined as failing three items
overall. Each subtest also has a "mercy rule," or alternative stopping point, so
that administration of a subtest can be stopped partway through an item set
if a child has failed so many items in succession that it is clear that further
items will be too hard for the child.

In order to ensure that the child understands the nature of the task, sev-
eral subtests feature "teaching items." Teaching, using explicit instructions
provided in the manual, is provided to a child who fails the first items of the
subtest. Since feedback is provided to the child about failures, the manual
encourages the examiner to provide positive feedback when a child passes a
teaching item.

The average administration time for the school-age battery of the DAS is
estimated to be 45 minutes for the core subtests, another 20 minutes for the
achievement battery, and up to 20 additional minutes if all of the diagnostic
subtests are used. At the upper preschool level, the core subtests are esti-
mated to take about 40 minutes, and administration takes an additional
25 minutes if all of the diagnostic subtests are used. At the lower preschool
level, the core subtests take about 25 minutes, and administration takes an
additional 10 minutes if the examiner chooses to use both of the diagnostic
subtests.

Scoring

Test Scores

The DAS requires one more step in the process of calculating test scores by
hand than most other cognitive batteries for children, because of its use of
item sets. Raw scores from each subtest must first be converted into an "abil-
ity score." This conversion is accomplished directly on the Record Form through a Rasch scaling procedure (Elliott, 1990c), and it essentially provides a measure of the child's ability based on both the number of items passed and the relative difficulty of those items. A child will obtain a higher ability score for passing 5 difficult items than for passing 5 very easy items. The ability score itself is usually not clinically interpreted but from it a normative score for the subtest can be obtained from the tables in the manual.

The GCA, cluster scores, Special Nonverbal Composite, and achievement test scores are all provided in the form of standard scores with a mean of 100 and a standard deviation (SD) of 15. Core and diagnostic subtests produce T scores with a mean of 50 and an SD of 10. Grade-based percentile scores are also provided for the achievement battery.

An appealing feature of the DAS is the flexibility given to the clinician in subtest selection by the provision of extended norms for subtests outside their usual age range. All of the upper preschool core subtests were standardized on 6- and 7-year-olds, while all of the school-age core subtests were standardized on 5-year-olds. Some of the preschool subtests are too easy for higher functioning 6-year-olds, while some of the school-age subtests are too hard for lower functioning 5-year-olds. However, it is possible to use the upper preschool core subtests with average and lower functioning 6- and 7-year-olds and obtain a T score value. Similarly, it is also possible to use the school-age subtests with average and higher functioning 5-year-olds.

In addition, a number of subtests were standardized over an even wider age range than the overlapping preschool and school-age ranges. For example, Recognition of Pictures, a preschool diagnostic subtest, was normed across the entire DAS age range of 2 years, 6 months through 17 years, 11 months. Although its usual age range is 3 years, 6 months through 7 years, 11 months, it can be administered out-of-level with 2½- to 3½-year-olds of average to high ability and to 8- to 17-year-olds of average to low ability. This makes the subtest a valuable additional measure of visual short-term memory for school-age children.

Extended GCA norms are also provided to calculate an estimate of overall cognitive ability for extremely low functioning children. Thus, it is possible to use the Lower Preschool level core subtests to obtain an extended GCA score for children up to 7 years old. The Upper Preschool level core subtests can be used to calculate an extended GCA score for children up to 14 years old. As will be elaborated, these extended norms are extremely helpful in accurate assessment of children with severe intellectual disabilities.

**Interpretation**

The DAS uses a classification system for GCA and cluster scores which is descriptive of the child's functioning rather than utilizing diagnostic-sounding terminology. Table 3.5 outlines this system.
Another particularly strong feature of the DAS is the excellent support provided to the clinician in the process of interpreting test results by the test handbook (Elliott, 1990c). The clinician is guided through a systematic plan of attack for interpreting DAS test scores as well as given a clear rationale for this approach. Encouragement is given to first checking for significant differences between the cluster scores themselves and also between the cluster scores and the GCA. Detailed ipsative analyses of subtest relative strengths and weaknesses is not encouraged unless a significant difference is found between subtest scores within a cluster. This approach should significantly reduce the risk of Type I error during subtest analyses. A further section of the test handbook provides guidance on clinical interpretation of subtest profiles, with tables outlining underlying processes involved in each subtest.

### CLINICAL SIGNIFICANCE

#### Psychometric Considerations

The DAS is an extremely valuable addition to the group of tests available for assessing cognitive ability in children. A number of features stand out that make the DAS more than just another battery. Reviewers are very pleased with the test's very strong psychometric characteristics, including the high reliability of the various DAS scales and the careful attention paid to ensuring a representative standardization sample as well the fairness of the test (Aylward, 1992; Braden, 1992; Kamphaus, 1993; Flanagan & Alfonso, 1995; Anastasi & Urbina, 1997).

Two important features of the DAS design, co-norming and high specificity, make it particularly useful for clinicians:

- **Co-norming**: For children with learning disabilities, the co-normed achievement battery allows for reliable analyses of ability/achieve-
3. Assessment with the Differential Ability Scales

ment discrepancies using the regression model preferred by many clinicians (Reynolds, 1990; Braden, 1992). The co-norming of the achievement battery means the percentage of children showing a particular discrepancy between GCA and an achievement subtest is known rather than estimated.

- **Specificity:** The assessment of children with possible learning disabilities forms a major portion of the work of school psychological services. Therefore, it is critical for a test battery to be able to detect significant strengths and weaknesses in scores, thereby aiding the clinician in understanding the specific learning processes with which the child is struggling. The DAS has the highest level of specificity in subtests and lower-order composites (Verbal, Nonverbal Reasoning, and Spatial) of any published battery (Elliott, 1997b). Such high levels of specificity will lead not only to greater accuracy but also to a higher incidence in the detection of specific, significantly high and low scores in subtests and clusters. Common cognitive profile patterns in learning disabled children are discussed more fully in the next section.

**Special Populations**

The DAS places less emphasis on verbal responses than most cognitive batteries for children. That, combined with its special composite score made up of only nonverbal subtests, offers advantages in the assessment of children with hearing impairments or speech dyspraxias, or those for whom English is a second language.

From our clinical experience, one of the greatest advantages of the DAS is its high appeal for preschoolers. For a test to provide a reliable and valid assessment when used with young children, it must not only have good psychometric characteristics but also look interesting enough so that they want to participate in the assessment and keep going when the work starts becoming difficult. The DAS does this much better than any of its current competitors because the test includes many items that involve little toys and activities that children do with their hands. Many of us who work with preschoolers lamented the changes made to the Stanford-Binet when the Fourth Edition was released (Thorndike et al., 1986); gone were most of the great little toys and activities from the Stanford-Binet L-M (Terman & Merrill, 1960), replaced in the new version by a much greater requirement for the child to focus on looking at pictures. The DAS restores the Stanford-Binet L-M approach to working with preschoolers.

Another important advantage of the DAS for preschoolers is that administration time for the core subtests is significantly shorter than that typically needed for competitors like the Stanford-Binet or WPPSI-R. An experienced examiner can complete the core subtests of the DAS and obtain a reliable
and valid estimate of overall cognitive ability in about 30 to 40 minutes. Our experience is that the Stanford-Binet Fourth Edition or WPPSI-R takes about 15 to 30 minutes longer than the DAS to obtain a similar estimate. This is a critical advantage in working with a population not known for having lengthy attention spans for intense diagnostic work.

Also, our clinical experience convinces us that the DAS is the test of choice for the intellectual assessment of children with developmental delays and mental retardation. The DAS offers great flexibility in choosing an individually tailored assessment battery for children with developmental disabilities. As well, it offers much greater sensitivity to the differential diagnosis of severe and profound mental retardation.

On first glance one might think, from evidence presented in the DAS Handbook (Elliott, 1990c), that the DAS may not be sensitive to identifying intellectually gifted children. A sample of 62 students previously identified as gifted, chosen from the standardization sample, showed a mean GCA of only 118.1. Braden (1992) speculates that this may be due to regression to the mean, a "Flynn effect" from using updated norms to evaluate these children, or variation in the definition of "giftedness" used to designate the students. It is certainly always the case that when children are selected according to certain criteria (say, Wechsler IQ plus achievement test scores), there will be regression to the mean when other criteria are applied (such as DAS GCA scores). The more extreme the scores, the greater the phenomenon. This seems the most likely explanation of the effect that Braden observed. There is no evidence that ceiling effects in subtests result in any lack of sensitivity in identifying intellectually gifted children. Even for the oldest DAS age group (17 years, 6 months through 17 years, 11 months), maximum scores on all core subtests yield a GCA of 156, indicating ample score range for assessing older gifted students. A further advantage of the DAS in assessing intellectually gifted children is the nature of the composites. Only those subtests that are the best measures of "g" are used to estimate composite scores (clusters and GCA). The exclusion of subtests that have low g saturations is unique to the DAS, making the GCA a particularly well-focused measure of g.

COMMON PATTERNS AND INTERPRETATION HYPOTHESES

Profile Analyses

As previously discussed, the DAS was created with the intention of providing a profile of a child's cognitive strengths and weaknesses, with the hope that this profile would lead to a better understanding and more effective remediation of possible learning difficulties. A major strength of the DAS is the comprehensive and lucid description of a method for the systematic interpreta-
tion of the DAS profile (Elliott, 1990c). This intent follows in the traditional wisdom passed on to most student clinicians in the first days of their cognitive assessment courses. Most of us were taught that, in interpreting intelligence tests with children, it is important to look beyond the overall composite scores to the unique information about the child that can be garnered from careful examination of the patterns of performance on the various sub-tests of the measure. Major texts on assessment with children emphasize this approach (Sattler, 1992; Kamphaus, 1993; Kaufman, 1994). Kaufman states that the composite scores tell us about the “what” of a child’s abilities while the subtests bring to light the “how.” It will be argued that the DAS lower-order composites (Verbal, Nonverbal Reasoning, and Spatial) are also particularly important in illuminating the “how.”

The analysis of subtest profile patterns to better understand a child’s learning strengths and weaknesses is controversial, and McDermott, Glutting, and colleagues (e.g., McDermott & Glutting, 1997; Glutting, McDermott, Konold, Snelbaker, & Watkins, 1998; Glutting, McDermott, Watkins, Kush, & Konold, 1997) have been very active in questioning such procedures. At the beginning of the last decade, McDermott, Fantuzzo and Glutting (1990) made a general statement advising “that psychologists just say ‘no’ to subtest analysis.” This was on the basis of a critique of practice using the WISC-R, but the recommendation came to be perceived as generalized to all cognitive tests. One of McDermott, Fantuzzo, and Glutting’s concerns centers on the relatively lower reliability and stability of subtest scores in comparison with composites. McDermott et al. argue that, because subtests typically have lower reliability and stability than composites, it is likely that the pattern of strengths and weaknesses among subtests that appears one day might not be there the next. Another concern relates to the use of ipsative scores in profile interpretation. Ipsative scores are produced by subtracting the child’s average normative score across various subtests from each individual subtest score, thereby removing the mean elevation of scores (i.e., variance associated with g). We emphasize that although they have applied these critiques to the DAS, the critiques do not constitute a major threat to DAS profile interpretations for the following reasons:

1. Interpretation of high and low subtest and composite scores is only recommended when differences between scores are statistically significant. This takes account of the reliability of the measures being compared. Relatively lower reliability results in larger differences being required for significance. Moreover, the DAS method adjusts significant differences for multiple comparisons. Because several comparisons are being made, differences required for significance are greater than for a simple comparison of two scores. This conservative approach is designed to ensure that only reliable differences, not attributable to measurement error, are reported.
2. McDermott and Glutting's negative conclusions about the value of profile analysis using ipsative scores do not apply to the DAS, which uses direct comparisons of normative subtest and composite scores. The only time scores are "ipsatized" is when the mean standardized T-score for the core subtests is subtracted from individual subtest scores for the purpose of evaluating whether that subtest score is, overall, significantly high or low. The ipsatized score (or the difference score, to put it another way) is never reported: Once a subtest score is identified as significantly high or low, the unadjusted T-score itself is reported. Also note that the composite scores are never ipsatized in the DAS procedure.

Glutting et al. (1997) also make the point that the interpretation of a profile should be done with reference to base rate information. Base rate refers to the frequency with which a particular profile is found in the population. Glutting et al. correctly assert that statistically significant differences (those that are reliable, i.e., unlikely to have arisen because of measurement error) can be quite common and ordinary, even though very "significant." To address this problem, Holland and McDermott (1996), using hierarchical multistage cluster analyses of the DAS standardization sample, identified seven core profile types in this sample that are representative of the child population. Five of these profiles, shown by a total of 71 percent of the sample, were flat in terms of their scores on the DAS core subtests (although there is some variation among the diagnostic subtests). The differences between the groups were defined mainly by variation in general ability. In other words, the profiles varied in the altitude rather than in the pattern of their scores. The remaining two core profile types were defined by (a) 16 percent of students who have relatively high Verbal versus low Spatial subtest scores, and (b) 13 percent of students who have relatively high Spatial versus low Verbal subtest scores. The Verbal and Spatial cluster scores were 10 points different in both cases. These two profile types will be discussed more fully, later in this section.

Attempts to show the utility of profile analyses with the DAS have concentrated mostly on children with learning disabilities, and in particular with children with reading disabilities. Elliott (1990c) examined DAS profiles for children from the standardization sample with previously identified learning disabilities. He reported that children with learning disabilities tend to score lower on the diagnostic subtests relative to their performance on the core subtests. The greatest differences are shown on the Recall of Objects, Recognition of Pictures, and Speed of Information Processing subtests. Similarly, Kercher and Sandoval (1991) reported finding a characteristic DAS subtest profile in children with reading disabilities. Their study found children with reading disabilities scoring significantly lower than typical children on the Recall of Digits and Recall of Objects diagnostic subtests. They argue that
3. Assessment with the Differential Ability Scales

this likely reflects a particular difficulty for these children with auditory and sequential processing and perhaps with verbal labeling as well. By way of contrast, a study by Shapiro, Buckhalt, and Herod (1995) reported a relatively flat profile of core and diagnostic subtests in a sample of 83 children with learning disabilities. Of course, studies such as these, examining mean profiles for such groups as “Learning Disabled” (LD), which are likely to be heterogeneous in profiles and causal factors, hide any possible major profile differences among subgroups.

To address this problem, cluster analyses have often been used to place individuals in subgroups according to their profile similarities. Using this method, Elliott (1990c) examined the profiles of 136 children from the standardization sample identified with reading disabilities because of the significant discrepancy between GCA and their performance on the Word Reading subtest. Four distinct profile types were identified among these children. Elliott argues that this supports the possibility of using subtest profile analyses to make clinically helpful distinctions between children with different types of learning disabilities.

McIntosh and Gridley (1993) examined DAS subtest profiles of 83 children with various learning disabilities. Cluster analyses indicated six distinct profile patterns, four of which showed little variation among subtest scores and were dominated primarily by their overall cognitive ability. The other two clusters showed more variation in subtest performance. Students in one cluster were characterized by difficulties with processing visual-spatial information, together with low scores on three diagnostic subtests: Recall of Objects, Recognition of Pictures, and Speed of Information Processing. The key characteristic of children in the other cluster was a relative weakness in the Nonverbal Reasoning cluster. This suggests that a major difficulty for these children may be with fluid reasoning or with the integration of visual and verbal information, important in successful performance of many complex tasks.

In another study with preschool children, McIntosh (1999) explored the ability of the DAS to discriminate between typical preschoolers and those identified as being “at-risk” for developmental or behavioral problems. The children in the at-risk group showed their greatest differences in performance compared to the typical children on the Picture Similarities core subtest and the Recall of Digits diagnostic subtest. McIntosh argued that poor performance on these two subtests is a particularly good screener for children at-risk for developmental problems.

Combined, these studies suggest that DAS profile analyses are useful in better understanding the unique difficulties experienced by children with particular learning disabilities. These results demonstrate that the diagnostic subtests and the cluster scores, as opposed to the overall GCA, provide the key information about these unique difficulties. In these studies, a variety of DAS cluster and subtest profiles are reported. It seems common, both
in studies of a normal population and in studies of children with disabilities, to find some groups with flat cluster and/or subtest profiles. Looking only at cluster scores, some studies find groups with relatively high Verbal versus Spatial scores, relatively high Spatial versus Verbal scores, and relatively low Nonverbal Reasoning versus Verbal and Spatial scores. In most studies, there is considerable variability among the diagnostic subtests.

Profiles of Samples of Dyslexic and Learning Disabled Children

Introduction

The following study addresses most of the issues raised by critics of profile interpretation, and also aims to identify a number of distinct profiles of DAS cluster scores among groups of normal and exceptional children. This study is principally based on an examination of patterns of cluster scores, measuring Verbal, Nonverbal Reasoning, and Spatial abilities.

Method

Subjects. Three major sources of data were used:

1. DAS Standardization Sample. This consists of 2400 children aged 6 years, 0 months through 17 years, 11 months. A total of 353 poor readers were identified in this total sample. Poor readers are defined as those with DAS Word Reading standard scores below 85. These poor readers were further subdivided into two subsamples:

   • Poor Readers with No Significant Discrepancy: 86 poor readers whose observed Word Reading score was not significantly lower than that predicted from their GCA
   • Poor Readers with Significant Discrepancy: 267 poor readers whose observed Word Reading score was significantly lower than that predicted from their GCA

This sample provides data constituting a baseline against which results from the other two samples may be evaluated.

2. Dyslexic Sample. This sample consists of 160 children identified as dyslexic by psychologists of the Dyslexia Institute in Great Britain. From a research perspective, this sample has the major advantage that the DAS had not been used in the original diagnostic process to identify these individuals as dyslexic. No information is available as to how much time elapsed between their initial identification and the subsequent DAS assessment. It seems likely that many would have received a considerable period of intervention for their reading difficul-

---

1 The data for this sample are used by kind permission of Dr. Martin Turner, Head of Psychology, Dyslexia Institute, Staines, England.
ties before their DAS assessment. The sample was divided into two subsamples, as follows:

- Dyslexics with DAS Word Reading standard scores below 85
- Dyslexics with DAS Word Reading scores between 85 and 100

3. Learning Disabled (LD) Sample. This sample consists of 53 children identified as learning disabled, with the WISC-III used as the initial assessment battery. Once again, this sample has the major advantage that the DAS had not been used in the original diagnostic process to identify these individuals as learning disabled. The sample was reevaluated on the DAS three years after their initial assessment. Dumont, Cruse, Price, and Whelley (1996) report full details of the sample, the procedure, and the initial findings.

Definition of Subgroups. The subgroups were defined according to the possible combinations of high and low scores that may be found among the three DAS clusters, and also including subgroups with flat cluster profiles. Children in all three samples were placed into subgroups based upon the presence or absence of significant discrepancies between cluster scores that were significant at the 95% confidence level, adjusted for multiple comparisons. The differences were obtained from Tables B.4 and B.5 in the DAS Introductory and technical handbook (Elliott, 1990c), and are similar to the differences indicated on the DAS Record Form.

Even among poor readers with a significant discrepancy between GCA and Word Reading (or, more properly, between observed Word Reading and Word Reading predicted from the GCA), it would be expected that there would be a proportion of children with flat cognitive test profiles. Poor reading has many causes, and there is no reason to believe that children who have difficulty with reading because of lack of exposure to teaching through absences from school, poor teaching, or poor motivation, should have anything other than normal (i.e., flat) cognitive profiles. Other poor readers may have verbal or spatial disabilities, both of which are amply reported in the literature (Snow, Burns, & Griffin, 1998; Rourke, Del Dotto, Rourke, & Casey, 1990). Finally, we may find some individuals whose Nonverbal Reasoning ability is lower than both their Verbal and Spatial abilities. Such a group had been identified by McIntosh and Gridley (1993). The second author has also received many questions and comments during the past several years from psychologists who have observed LD children showing this profile pattern. Finally, there may be some individuals who show the reverse pattern, with Nonverbal Reasoning (NVR) ability higher than both Verbal and Spatial, although no research studies have identified such a subgroup. The subgroups are, therefore, as follows:

---

2The data for this sample are used by kind permission of Dr. Ron Dumont, Director, M.A. and Psy.D. Programs in School Psychology, Fairleigh Dickinson University, Teaneck, NJ.
• **Flat Cluster Profile.** No significant differences among the three DAS cluster scores.

• **Low Spatial, High Verbal.** Verbal cluster significantly higher than Spatial cluster. Possible nonverbal learning disability.

• **Low Verbal, High Spatial.** Verbal cluster significantly lower than Spatial cluster. Pattern typically reported for poor readers. (See, e.g., Snow et al., 1998; British Psychological Society, 1999).

• **High NVR.** Nonverbal Reasoning cluster higher than both the Verbal and Spatial scores, and significantly higher than at least one of them. Interpreted as signifying good ability to process complex auditory-visual information.

• **Low NVR.** Nonverbal Reasoning cluster lower than both the Verbal and Spatial scores, and significantly lower than at least one of them. Interpreted as indicating difficulty in processing complex auditory-visual information.

### Results

Table 3.6 shows the frequency and percentages of children in the standardization sample who had each profile. Because the DAS standardization sample was chosen to stringently match U.S. census data, we can assume that these percentages provide a good estimate of the rates of each profile in the U.S. school-age population. Fifty percent of the total standardization sample had a flat cluster profile. However, 66.3% of the poor readers who had

<table>
<thead>
<tr>
<th>Type of profile</th>
<th>Poor readers with no discrepancy</th>
<th>Poor readers with discrepancy</th>
<th>Total DAS standardization sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Cluster Profile</td>
<td>57</td>
<td>121</td>
<td>1203</td>
</tr>
<tr>
<td></td>
<td>66.3%</td>
<td>45.3%</td>
<td>50.1%</td>
</tr>
<tr>
<td>Low Spatial, High Verbal</td>
<td>6</td>
<td>16</td>
<td>257</td>
</tr>
<tr>
<td></td>
<td>7.0%</td>
<td>6.0%</td>
<td>10.7%</td>
</tr>
<tr>
<td>Low Verbal, High Spatial</td>
<td>8</td>
<td>63</td>
<td>239</td>
</tr>
<tr>
<td></td>
<td>9.3%</td>
<td>23.6%</td>
<td>10.0%</td>
</tr>
<tr>
<td>High NVR</td>
<td>8</td>
<td>28</td>
<td>355</td>
</tr>
<tr>
<td></td>
<td>9.3%</td>
<td>10.3%</td>
<td>14.8%</td>
</tr>
<tr>
<td>Low NVR</td>
<td>7</td>
<td>39</td>
<td>346</td>
</tr>
<tr>
<td></td>
<td>8.1%</td>
<td>14.6%</td>
<td>14.4%</td>
</tr>
<tr>
<td>Column Totals</td>
<td>86</td>
<td>267</td>
<td>2400</td>
</tr>
</tbody>
</table>

Note: The subsamples in the first two columns form 14.7% of the total standardization sample.
no discrepancy between observed and predicted Word Reading scores had a flat cluster profile. The range of GCA scores in this particular group is quite restricted, ranging from 54 to 84. The group likely contains many children functioning in the mild to moderate range of mental retardation who would be expected to be more likely to show a flat cluster profile.

The larger group of poor readers who had a significant discrepancy between observed and predicted Word Reading scores showed a larger variance in GCA scores, ranging from 46 to 118. Compared with the total standardization sample, a slightly smaller percentage (45.3%) of these poor readers showed flat profiles. Also, about a quarter (23.6%) of this subgroup, as might be expected, had significantly lower Verbal than Spatial ability. In the total sample, 10% of children had Low Spatial and High Spatial scores, and 14% to 15% showed Low and High NVR profiles.

Table 3.7 shows the results for the dyslexic and learning disabled samples. They are remarkably similar for the two samples, despite the data being gathered in different countries and in different settings. Chi-square tests showed that there is no significant difference between the two dyslexic groups (Word Reading below 85 and Word Reading 85–100; \( \chi^2 = 5.13; df = 4; \text{N.S.} \)), nor is there any significant difference between the combined dyslexic groups and Dumont et al.'s (1996) LD sample (\( \chi^2 = 1.337; df = 4; \text{N.S.} \)). About one third of these samples had flat cluster profiles, fewer than in the standardization sample but still a substantial proportion. Both the dyslexics with Word Reading below 85 group and the Dumont sample had 11% to 12% in the Low Verbal, High Spatial subgroup. This is about half the frequency of Low Verbal children in the comparable subgroup from the standardization sample who were poor

<table>
<thead>
<tr>
<th>Type of profile</th>
<th>Dyslexic with WORD READING below 85</th>
<th>Dyslexic with WORD READING 85–100</th>
<th>Dumont et al. (1996) LD sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Cluster Profile</td>
<td>28</td>
<td>28</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>34.5%</td>
<td>35.5%</td>
<td>37.7%</td>
</tr>
<tr>
<td>Low Spatial, High Verbal</td>
<td>4</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>4.9%</td>
<td>15.2%</td>
<td>9.4%</td>
</tr>
<tr>
<td>Low Verbal, High Spatial</td>
<td>10</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>12.3%</td>
<td>8.9%</td>
<td>11.3%</td>
</tr>
<tr>
<td>High NVR</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>6.2%</td>
<td>5.1%</td>
<td>1.9%</td>
</tr>
<tr>
<td>Low NVR</td>
<td>34</td>
<td>28</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>42.0%</td>
<td>35.4%</td>
<td>39.6%</td>
</tr>
<tr>
<td>Column Totals</td>
<td>81</td>
<td>79</td>
<td>53</td>
</tr>
</tbody>
</table>
readers with ability-achievement discrepancies. One wonders if Low Verbal children tend not to be identified as dyslexic or learning disabled. It seems possible that many such children may be found in inner-city and poor socio-economic environments. They may thereby get special education services from other sources or designations (e.g., Title I funding in the United States). Such children may often be considered to be "garden-variety" poor readers, to use Stanovich's (1988) term, rather than dyslexic or LD. Further research is needed to clarify these issues. Turning to the Low Spatial, High Verbal subgroup, both samples showed a similar proportion, compared to the total DAS standardization sample, of students with this profile. Dumont et al.'s sample had 9.4% and the combined dyslexic sample had 10% with this profile. It is possible that a number of children with this profile have a Nonverbal learning disability (Rourke et al., 1990).

There is a highly significant difference between the frequencies for each profile for the combined dyslexic and LD samples, on the one hand, and the standardization sample, on the other. Comparison of the frequencies for each profile for the combined dyslexic/LD sample and the Poor Readers with Discrepancy, taken from the standardization sample, yields a chi-square of 48.48 ($df = 4; p < .001$). Similarly, comparison of the combined dyslexic/LD sample and total standardization sample yields a chi-square of 94.71 ($df = 4; p < .001$). The differences that account for the highest chi-square values are for children with the Low NVR profile.

Very few dyslexic or LD children had a High NVR profile—considerably fewer than the proportion in the total DAS sample. However, more than one third of the dyslexic and LD samples fell into the Low NVR subgroup. The results from the Dumont et al. (1996) LD sample and the dyslexic sample are remarkably similar, providing mutual cross-validation of these findings. The mean profile for the combined dyslexic and LD children who are in this subgroup ($N = 83$) is shown in Figure 3.1. The differences between the mean scores are dramatic; the Nonverbal Reasoning mean is lower than both Verbal and Spatial means by more than one standard deviation.

**Discussion**

Why should children with reading disabilities score poorly on the two DAS subtests measuring Nonverbal Reasoning? The answer seems most plausibly to lie in the nature of the tasks of reading and "nonverbal" reasoning. Reading requires a high level of visual/verbal integration in order to convert visual printed codes into sounds and words. For fluent reading, and for recognition of common words or letter strings, an individual needs information in the auditory/verbal and visual processing systems to be effectively integrated. Similarly, to perform well on the DAS Nonverbal Reasoning tasks (or, indeed, on any good measures of fluid reasoning) one needs good integration of the visual and verbal processing systems. These tasks are presented visually, hence the term "nonverbal" that describes them. But to solve the problems effec-
tively, the use of internal language to label and to mediate the solution of the problems is generally essential. Even if an individual has excellent verbal and spatial abilities, if the two brain processing systems specialized for those abilities do not "talk" to each other effectively, this may have an adverse effect on performance both in reasoning and in reading acquisition.

Readers may wonder why these striking findings, on two independent samples, have not been reported previously for other test batteries. The short and simple answer (since there is insufficient space to elaborate on it) is that all other psychometric batteries used with children, with one exception, do not have separate measures of Verbal ability (Gc), Spatial ability (Gv), and Nonverbal Reasoning ability (Gf).

**Case Example**

The case of Mike (age 7 years, 11 months) is typical of a dyslexic or reading-disabled student with a low NVR profile. Mike was referred for assessment

---

3The one exception is the Woodcock-Johnson Tests of Cognitive Ability—Revised (WI-R; Woodcock & Johnson, 1989). A lack of research evidence on such difficulties with fluid reasoning tasks for dyslexic and learning disabled children may be due to one of two possible reasons: (a) a lack of research with substantial samples, or (b) a problem with the subtests that purport to measure Gv and Gf. For example, the correlation between the two WI-R subtests measuring Gv (Visual Closure and Picture Recognition) is very low: 0.22 at age 6 years, 0.30 at age 9 years, and 0.29 at age 13 years. Such low correlations beg the question of whether the composite formed from such a weak pairing measures anything meaningful. In comparison, the correlations between the two DAS measures of Gv (Recall of Designs and Pattern Construction) are 0.56, 0.54, and 0.61 for the same age groups.
because, despite being self-evidently bright verbally and very capable in visual-motor tasks, he had had persistent problems since starting school in learning to read fluently and to spell accurately. His scores on the DAS are shown in Table 3.8. There are no significant differences between the two Verbal, the two Nonverbal Reasoning, or the two Spatial subtests. However, his Nonverbal Reasoning (NVR) cluster score is significantly lower than his Spatial cluster score, and his NVR score is lower than his Verbal score (this difference falling just short of statistical significance). Nevertheless, his NVR score is well below both the Verbal and Spatial scores, making his profile fit the low NVR subgroup. His NVR and Spatial scores are also significantly different from his GCA score, and are consequently marked “L” and “H,” respectively. As discussed earlier, it seems probable that while Mike is fluent verbally and has good Spatial skills, he has problems in auditory-visual integration that arguably have influenced his acquisition of reading skills.

Turning to the diagnostic subtests, Mike’s score on Recall of Digits is

<table>
<thead>
<tr>
<th>TABLE 3.8</th>
<th>DAS Subtest, Cluster, and Achievement Scores for Mike</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DAS subtest or cluster</strong></td>
<td><strong>Score</strong></td>
</tr>
<tr>
<td><strong>Core Subtests</strong></td>
<td><strong>(T-Score)</strong></td>
</tr>
<tr>
<td>Word Definitions</td>
<td>53</td>
</tr>
<tr>
<td>Similarities</td>
<td>49</td>
</tr>
<tr>
<td>Matrices</td>
<td>40</td>
</tr>
<tr>
<td>Sequential &amp; Quantitative Reasoning</td>
<td>47</td>
</tr>
<tr>
<td>Recall of Designs</td>
<td>59</td>
</tr>
<tr>
<td>Pattern Construction</td>
<td>61</td>
</tr>
<tr>
<td><strong>Diagnostic Subtests</strong></td>
<td></td>
</tr>
<tr>
<td>Recall of Digits</td>
<td>40 (L)</td>
</tr>
<tr>
<td>Recall of Objects (Immediate)</td>
<td>50</td>
</tr>
<tr>
<td>Speed of Information Processing</td>
<td>61 (H)</td>
</tr>
<tr>
<td><strong>Clusters</strong></td>
<td><strong>(Standard Score)</strong></td>
</tr>
<tr>
<td>Verbal</td>
<td>101</td>
</tr>
<tr>
<td>Nonverbal Reasoning</td>
<td>88 (L)</td>
</tr>
<tr>
<td>Spatial</td>
<td>116 (H)</td>
</tr>
<tr>
<td>GCA</td>
<td>102</td>
</tr>
<tr>
<td><strong>Achievement Tests</strong></td>
<td><strong>(Standard Score)</strong></td>
</tr>
<tr>
<td>Word Reading</td>
<td>79 (L)</td>
</tr>
<tr>
<td>Spelling</td>
<td>87 (L)</td>
</tr>
<tr>
<td>Basic Number Skills</td>
<td>91 (L)</td>
</tr>
</tbody>
</table>

Note: In the Score column, “L” denotes a statistically significant low score and “H” a statistically significant high score. These are explained more fully in the text.
significantly low, in comparison with his mean T-score, derived from the six core subtests. His score on Speed of Information Processing is significantly high. He therefore appears to have a significant weakness in auditory short-term memory, in addition to his relative weakness in auditory-visual integration. His parents and teacher also commented that Mike often quickly forgets verbally given instructions. They put this down to inattention, but an alternative hypothesis is a relative weakness in verbal short-term processing. On the other hand, his speed of visual information processing is relatively high, supporting his good Spatial ability.

On the three DAS achievement tests, Mike's obtained scores on Word Reading, Spelling, and Basic Number Skills are all significantly lower than the scores predicted from his GCA score (Word Reading and Spelling have predicted scores of 101; Basic Number Skills has a predicted score of 102). The difference of 22 points between Mike's observed and predicted scores on Word Reading is found in fewer than 5% of children.

So, what would appropriate intervention recommendations be for a boy like Mike? Previous research (e.g., Bryant, 1968) had suggested that poor auditory-visual integration was not a cause of poor reading acquisition. In spite of this, for many years, teachers of dyslexic children have actively advocated multisensory teaching methods. Teachers appear to have long held to the view that dyslexic children have difficulty integrating visual and verbal information. The reader will recall that it is hypothesized that a relative weakness in this ability underlies the Low NVR profile found in the samples of dyslexic and LD children, reported earlier. Thus, it was recommended that a multisensory teaching method should be used with Mike. His poor auditory short-term memory should also be taken into account by (a) minimizing the length of verbal instructions; (b) using repetition where necessary; and (c) using his strengths with visual short-term memory to compensate for his relatively poor auditory short-term memory, perhaps by keeping a notebook of things to remember. Useful references to multisensory teaching approaches are given by Thomson and Watkins (1998), Augur and Briggs (1992), and Walker and Brooks (1993). The 40% of children in this study who were found to be reading disabled may well benefit from such an approach.

ADVANTAGES AND DISADVANTAGES

Use with Children with Developmental Disabilities

The clinical experience of the first author leads us to argue that the DAS is the test of choice for the intellectual assessment of children with developmental delays or mental retardation. One of the reasons for this assertion is the flexibility the DAS offers to tailor a battery of test activities to fit the child's developmental level. For example, the extended GCA norms allow the clinician to use the upper preschool level core subtests with children up to 13 years,
11 months. For school-age children with severe or profound mental retardation, this provides an invaluable alternative to the WISC-III. It permits a valid estimate of overall cognitive ability while using a set of test activities that are most appropriate for the child's developmental level. The upper preschool core subtests offer the clinician the chance to see children with severe disabilities working on activities in which they will experience a fair degree of success balanced with their failures. Similarly, it is a real advantage that the upper preschool subtests are fully normed through 7 years, enabling them to be used in assessing the vast majority of 6- and 7-year-olds likely to be referred for developmental delays. This allows these appealing and developmentally appropriate test activities to be used to obtain comprehensive, accurately normed assessment results with these children.

Another excellent feature for working with children with developmental difficulties is the provision of an alternative method of administration of the Pattern Construction subtest, using nontimed norms. Normative comparison can be made for children based only on their ability to complete the designs rather than their speed in doing so. For many children with milder forms of cerebral palsy, our experience is that they are capable of copying the patterns but somewhat awkward and slower in doing so. The nontimed norms allow a way of teasing out the problem-solving skills from the fine motor difficulties.

A critical feature for the clinician working with children with mental retardation is that the DAS is superior to other intellectual batteries in its ability to differentiate moderate, severe, and profound intellectual disabilities. As Table 3.9 shows, the DAS provides a lower floor in terms of overall test score than the WPPSI-R or Stanford-Binet across the entire preschool age range if the clinician uses the extended norms options available. For school-age children, the extended norms option allows for estimates of intellectual disability down to 5 standard deviations below the mean, to the border of the severe to profound intellectual disability range. The Stanford-Binet and WISC-III permit estimates down to only 4 standard deviations below the mean, or just

<table>
<thead>
<tr>
<th>Table 3.9</th>
<th>Minimum Possible Composite Test Scores on the DAS, WPPSI-R, and Stanford-Binet across Preschool Age Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age Level</td>
<td>Test Score ↓</td>
</tr>
<tr>
<td></td>
<td>2-0</td>
</tr>
<tr>
<td>DAS GCA 1</td>
<td>—</td>
</tr>
<tr>
<td>WPPSI-R</td>
<td>—</td>
</tr>
<tr>
<td>Full Scale IQ</td>
<td>—</td>
</tr>
<tr>
<td>Stanford-Binet Test Composite 2</td>
<td>80</td>
</tr>
</tbody>
</table>

1 DAS and WPPSI-R scores are based on raw scores of zero for each subtest.
2 Stanford-Binet minimum possible test composite score based on analyses of lowest possible score, assuming a raw score of 1 on at least one subtest.
3. Assessment with the Differential Ability Scales

to the border of the moderate to severe intellectually disabled range. Discussion of the low floor of the DAS may strike some like the scene in the movie *This Is Spinal Tap* where the band boasts about being the only group with "amplifiers that go up to 11"; however, there are definite real-world advantages. Being able to distinguish between moderate and severe mental retardation is of significant help in providing practical educational programming recommendations to maximize a student's learning potential. Also, many preschool services demand documentation of a developmental delay of at least 3 standard deviations below the mean, and for many age ranges in the preschool years the DAS is the only test that can do this.

While GCA and cluster scores boast low floor levels for preschoolers, Flanagan and Alfonso (1995) caution that several of the subtests have inadequate floors. They define a floor as inadequate if a raw score of 1 does not produce a standard score more than 2 standard deviations below the mean. By this definition, all four core subtests at the lower preschool level, which starts at 2 years, 6 months, have adequate floors for age 3 years, 3 months and older. Inadequate floors are shown by Block Building up to 3 years, 2 months; Picture Similarities up to 2 years, 11 months; and Naming Vocabulary up to 2 years, 8 months. Verbal Comprehension has an adequate floor at all ages. Flanagan and Alfonso point out that the low floor on the DAS for children from 2 years, 6 months to 3 years, 6 months would come from an assessment where the child had actually succeeded at doing very little. Given this, the Bayley Scales of Infant Development—Second Edition (Bayley, 1993) would likely be the better assessment tool for children in this age range with very significant delays, to obtain a more comprehensive assessment of their developmental skills.

The three new subtests added to make up the upper preschool level (Pattern Construction, Early Number Concepts, and Copying) have inadequate floors from 3 years, 6 months through 3 years, 11 months. However, in practice this is not a major problem; if a clinician were to find that a child below 4 years had raw scores of 0 or 1 on these three subtests, then Block Building should always be administered and the child assessed using the four core subtests, which have good floors at the lower preschool level for 3½- to 4-year-olds.

**Use with School-Age Children**

So far we have argued for the advantages of the DAS in the assessment of preschoolers and children with developmental disabilities. Clinicians often wonder, however, what would recommend the DAS over the WISC-III or Stanford-Binet for more typical school-age children. The correlation between the GCA and the Full Scale IQ of the WISC-III is .92 (Wechsler, 1991). Further, Dumont et al. (1996) found with a sample of children with learning disabilities that the DAS GCA produced, for the vast majority of the children, the same intelligence level classification as that provided by the WISC-III 3 years
earlier, when confidence intervals were taken into account. This evidence supports the notion that the DAS and WISC-III estimate overall cognitive ability equally well in school-age children. However, the DAS has two major advantages. First, it can provide this estimate in significantly less time than the WISC-III. In our experience, it takes about 45 to 60 minutes to complete the six core school-age subtests on the DAS with a school-age child as compared to about 60 to 90 minutes to complete the 10 to 13 WISC-III subtests. The potential time saving can be very valuable to the clinician who is using the intellectual battery primarily to obtain a good estimate of overall cognitive ability but who also needs time to assess other important areas about the child, such as emotional functioning. The second major advantage for the DAS is the Nonverbal Reasoning composite. It was demonstrated earlier that this is a critical element in the assessment of school-age children with learning disabilities. The NVR element is absent from the WISC-III and all other test batteries except the Woodcock-Johnson—Revised (WJ-R), in which it can be argued that both the Gf and Gv factors are measured less effectively than in the DAS.

**Limited Sampling of Verbal Skills**

For some assessment situations, a disadvantage of the DAS is that it provides a more limited sampling of expressive language than most other assessment batteries. At the preschool level, the child only has to speak to provide single-word answers to the Naming Vocabulary subtest and to count out loud the number of squares on Early Number Concepts. Especially with 3-year-olds and up, it is vital for the clinician to get some idea of how well the children can express themselves in phrases and sentences to complete a thorough developmental assessment. In the first author's clinic, he often supplements the DAS with the Verbal Reasoning subtests from the Stanford-Binet Fourth Edition. In many settings, if the clinician suspects that the child has a specific language disorder, further in-depth language assessment by a speech and language pathologist would be appropriate.

The school-age battery is somewhat better with the verbal core subtests requiring the child to give a verbal definition of a word and describe how three words are similar. Still, there is no equivalent of the WISC-III's Comprehension subtest, which often allows a sensitive clinician insight into a child's social skills and common sense knowledge. For most of us, our training as clinicians emphasizes that administering an intellectual assessment battery to a child is about more than just obtaining an estimate of overall cognitive ability; it is also a structured clinical interview with the potential for insight into the child's personality. The more limited opportunity for verbal expression on the DAS may detract from this purpose somewhat.

This more limited sampling of verbal content was quite deliberate (Elliott, 1997b), and, as we have previously discussed, it does offer advantages in the
assessments of children with hearing impairments, oral-motor difficulties, or those who are from different cultures. However, this advantage does come with a price: a more limited developmental assessment of language for many other children.

**Difficulties in Mastering Administration and Scoring**

Many have commented (Aylward, 1992; Braden, 1992) that the DAS is a difficult test for the clinician to learn to administer and score. The use of item sets requires a mental shift for many examiners who have been indoctrinated to follow simple basal and ceiling rules. To complicate matters further, most tests require the examiner to keep in mind two possible situations for discontinuing the subtest, either at the decision point or if enough consecutive failures occur. A good example of this complexity comes on the Pattern Construction subtest with 4-year-olds. Novice DAS clinicians must learn the rules so that, for most 4-year-olds, they will administer items 1 to 7 to reach the decision point. There, they will stop if the child has failed to obtain the maximum time bonuses on more than two items. However, if the child is experiencing great difficulties, the clinician must also stop if the child has obtained four first trial failures on five consecutive items. So the clinician must simultaneously keep in mind two complicated stopping rules.

There is a definite reward for the clinician able to master these rules. Many of us have suffered along with the preschoolers struggling on the WPPSI-R’s Block Design subtest. Its discontinuance requirement is that the child fail both trials of three consecutive items. This process, of needing to fail six times in a row, can often lead a child to feel quite discouraged and less eager to try anything else. The option to stop Pattern Construction on the DAS at item 7, while many 4-year-olds are still experiencing a good deal of success, is a real boon to rapport. In general, while the DAS stopping rules are more complex than other tests, once they are mastered, they typically result in more efficient administration and a more pleasant experience for the child.

Another issue with respect to the complexity of learning to score the DAS by hand is the double transformation of raw scores required to obtain a subtest T score. First, the raw score must be matched to the appropriate item set and converted to an ability score. Next, the ability score is then taken to the norms tables for transformation into a T score. We are all aware of the literature indicating the surprising number of clerical errors made in calculating test scores (Alfonso & Pratt, 1997). This double transformation may make the possibility of clerical error twice as likely in hand-scoring the DAS, and extra caution is required by the clinician to prevent these types of errors. Fortunately, the recent availability of scoring software for the DAS (Elliott, Dumont, Whelley, & Bradley, 1998) will no doubt help to alleviate the problem. With the software, the clinician just enters the subtest ability scores, with no further need to look up information in tables.
Comprehensiveness of the Achievement Battery

One disadvantage of the DAS achievement battery is the lack of a measure of reading comprehension. The Word Reading subtest looks at basic word recognition skills, but does not really assess the critical function of how well children are able to understand written information in the manner they are likely to encounter it in the real world. There is no compelling reason for the DAS to replicate the comprehensiveness of academic content coverage provided by the Wechsler Individual Achievement Test (WIAT; The Psychological Corporation, 1992), but the addition of a measure of reading comprehension would have been welcomed by many clinicians.

RECOMMENDATIONS

To be effective, assessment must be about more than generating a set of test scores. Assessment must lead to a better understanding of our clients and offer insight into how to solve the problems with which they present. As the following case study illustrates, the DAS can be very useful for the clinician who strives to determine a child's relative cognitive strengths and weaknesses and then to translate this profile into an effective treatment plan.

The Case of Jean-Guy: Background Information

Jean-Guy is a 6-year-old boy with Williams Syndrome seen for a review psychological assessment at the request of his school. On three earlier assessments, when he was 14, 23, and 38 months old respectively, the Bayley Scales of Infant Development—Second Edition had been used. Each time, Jean-Guy had shown a very significant delay in his overall cognitive development with his composite score on the Bayley each time falling below the minimum of 50. Thus, while the results suggested that Jean-Guy had an overall intellectual disability (or mental retardation), it was not possible to determine from the Bayley results whether his intellectual disability was in the moderate or severe range.

Jean-Guy was now in his second year of kindergarten. His work with a speech and language pathologist had led to many gains in his expressive language skills. His parents were quite heartened by these gains, and had asked his teachers when they expected that Jean-Guy might catch up to the other children. His teachers referred Jean-Guy to the school psychologist to try to determine his long-term prognosis and learning needs. The school psychologist attempted unsuccessfully to assess Jean-Guy using the WISC-III. Jean-Guy found the WISC items extremely difficult, and after a few attempts at the first three subtests began to refuse to try. The school psychologist felt that the results suggested that Jean-Guy had a very significant intellectual dis-
ability, but his parents found it hard to believe that opinion, since they felt that Jean-Guy had been bored by the WISC items and had not shown what he was truly capable of doing. The school psychologist suggested a referral to a clinic specializing in children with intellectual disabilities.

**Tests Used**

The clinic psychologist elected to assess Jean-Guy's cognitive abilities with the DAS. She chose to work with him using the upper preschool level subtests, since she felt that the preschool activities would give Jean-Guy a better balance of success and failure than the School Age Battery.

Along with the DAS, the clinic psychologist administered the Vineland Adaptive Behavior Scales (Sparrow, Balla, & Cicchetti, 1984) by interviewing Jean-Guy's parents. The Vineland is a structured interview that gathers information on adaptive behavior skills in areas such as communication, self-help, and social skills.

**Assessment Results and Interpretation**

Jean-Guy was very cooperative with trying all of the DAS preschool subtests. He was able to stay focused on his work and genuinely seemed to be enjoying himself. His parents observed the DAS being administered and commented when it was over that they felt this time that Jean-Guy had been able to show both what he was good at and what was difficult for him.

Table 3.10 shows Jean-Guy's scores on the DAS and the Vineland. Jean-Guy's GCA score of 51 suggested that he was currently functioning in the moderate range of mental retardation. However, there was a very significant difference between his Verbal cluster score (70) and his Nonverbal cluster score (43). This pattern of relative strengths with verbal reasoning accompanied by very significant difficulty with visual or nonverbal problem solving is seen frequently in children with Williams Syndrome (Jarrold, Baddely, & Hewes, 1998).

Jean-Guy's performance on the Naming Vocabulary subtest was significantly higher than his average overall subtest performance. He was able to name many of the pictures shown to him. However, on the pictures that he did not name, Jean-Guy tended to echo back the psychologist's prompt rather than making an attempt or saying that he didn't know the answer. Jean-Guy had much more difficulty with the Verbal Comprehension subtest. His T score of 24 was significantly below his Naming Vocabulary score, and was very similar to his performance on the Nonverbal subtests. Jean-Guy had been able to follow directions when they involved picking out the right toy in a group, but he began making mistakes as soon as the verbal directions involved two steps or positional concepts such as "on" or "under." Similarly, his
TABLE 3.10
Assessment Results for Jean-Guy

<table>
<thead>
<tr>
<th>Subtest</th>
<th>T score (mean = 50, SD = 10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal Comprehension</td>
<td>24</td>
</tr>
<tr>
<td>Picture Similarities</td>
<td>20</td>
</tr>
<tr>
<td>Naming Vocabulary</td>
<td>40</td>
</tr>
<tr>
<td>Pattern Construction</td>
<td>20</td>
</tr>
<tr>
<td>Early Number Concepts</td>
<td>20</td>
</tr>
<tr>
<td>Copying</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Composite score (mean = 100, SD = 15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal Cluster</td>
</tr>
<tr>
<td>Nonverbal Cluster</td>
</tr>
<tr>
<td>General Conceptual Ability (GCA)</td>
</tr>
<tr>
<td>GCA (Extended norms)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Domain</th>
<th>Standard score (mean = 100, SD = 15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication</td>
<td>53</td>
</tr>
<tr>
<td>Daily Living Skills</td>
<td>46</td>
</tr>
<tr>
<td>Socialization</td>
<td>57</td>
</tr>
<tr>
<td>Adaptive Behavior Composite</td>
<td>48</td>
</tr>
</tbody>
</table>

Low performance on the Early Number Concepts subtest seemed primarily due to his difficulty understanding all of what he was being asked to do from the verbal instructions.

The clinic psychologist wondered if Jean-Guy's Verbal cluster score (70) might be underestimating his degree of difficulty with expressive and receptive language skills. The Verbal cluster score is pulled up by Jean-Guy's strong performance on Naming Vocabulary, which involves only providing a one-word answer to name a picture. Some children with Williams Syndrome can initially seem quite verbally adept, with good vocabularies. However, on careful listening to their speech, it is sometimes noted that their conversations tend to be quite superficial and feature stereotyped phrases. This is sometimes referred to as "cocktail party speech" (Udwin & Yule, 1990). Indeed, Jean-Guy's score on the communication section of the Vineland suggested that he was showing a very significant deficit in his practical communication skills. Information from the Vineland indicated that Jean-Guy had difficulty
with describing events that had happened to him as well as with asking questions to seek out more information from another person. The Verbal cluster score did not seem sensitive to the full degree of difficulty Jean-Guy was experiencing. However, the relatively lower scores on Verbal Comprehension and Early Number Concepts had served to “red-flag” the problem. It should be noted that because Jean-Guy’s scores on the two verbal subtests were significantly different, the Verbal cluster score should be interpreted with caution, if at all.

There were no differences between Jean-Guy’s T scores on the Nonverbal cluster subtests. He was unable to copy even a straight line on the Copying subtest, and his only success on Pattern Construction seemed almost by accident. While his T score on Picture Similarities was also very low, Jean-Guy was able to successfully match a number of the items. He also obviously enjoyed the task and seemed happy to persist even when the items became difficult.

Since four of the six subtests had T scores of 20, the minimum possible T score, the clinic psychologist elected to use the extended norms option. Elliott (1990c) recommends this when several T scores are at floor level, in order to provide a more accurate estimate of the degree of cognitive delay. The GCA score from the extended norms option was 44. While somewhat lower than Jean-Guy’s original GCA, the extended norms option did provide evidence that Jean-Guy’s intellectual disability was in the moderate instead of the severe range. This was consistent with the adaptive behavior data from the Vineland.

**Recommendations from the Assessment**

The first goal of this assessment was to provide Jean-Guy’s parents and teachers with information about his prognosis. His parents had been encouraged by their hard-won gains with his expressive language. They had been reluctant to see his developmental problems as a permanent disability, as suggested by the school psychologist, because they did not feel his assessment was valid. From their point of view, Jean-Guy had simply not participated enough with the WISC-III subtests for them to believe that the psychologist could conclude something so major from the data. However, Jean-Guy did participate and focus more on the DAS upper preschool subtests. He experienced a number of successes along with his failures. His parents agreed, before hearing the test results, that they felt he had given a good showing of what he was capable of doing. They were still grief-stricken when the diagnosis of moderate mental retardation was discussed, but a major source of their resistance to the news was removed. Effective assessment and treatment planning for children with mental retardation involves much more than merely a test of cognitive ability. It requires a weaving together of information from multiple sources to form a comprehensive picture of the child’s
strengths and needs (see Gordon, Saklofske, & Hildebrand [1998] for more information). However, the DAS results were an important step in helping Jean-Guy's family better appreciate the degree of his disability and begin to work together with his teachers on a more effective treatment plan. Three key practical recommendations came out of the assessment.

• Jean-Guy's educational program should focus on practical or functional skills. The best environment for the development of Jean-Guy's social skills was his neighborhood school with age-level peers. However, simply having him work with the same academic content at a much slower pace than the other children seemed a plan doomed to much frustration. Instead, an ecological inventory was completed with Jean-Guy's parents and teachers to identify and prioritize the practical skills he needed to increase his independence. Wolery and Wilbers (1994) provide excellent practical resources on how to design an inclusive functional program for children with mental retardation.

• It was recommended that Jean-Guy continue to receive services from the school speech and language pathologist. While his Verbal cluster score was a relative strength, the assessment revealed a number of quite significant difficulties with Jean-Guy's receptive and expressive language skills. The focus of the next phase of Jean-Guy's treatment was to build the gains he had made with expressive vocabulary into better conversation skills as well as to increase his ability to understand verbal directions. Concurrently, Jean-Guy's teachers were encouraged to keep their verbal directions short in length and also to consider pairing them with visual prompts and cues (e.g., gesturing to what they wanted Jean-Guy to pick up, along with telling him).

• Jean-Guy showed very significant difficulties with visual-spatial activities, like many children with Williams Syndrome. An occupational therapist was invited to join the team. She recommended placing an emphasis on functional fine-motor activities such as dressing skills. She suggested adaptations such as Velcro fasteners for Jean-Guy's shoes, to compensate for his extreme difficulties in learning to tie laces. She also encouraged patience with academic content such as printing and letter recognition skills, explaining that both drew on visual-spatial skills where Jean-Guy showed his greatest weakness. She recommended using an inked stamp with Jean-Guy's name on it in order to allow him to "print" his name with the other children.

**INSIGHTS FROM EXPERIENCE**

From our clinical work, we are convinced that the DAS is a very welcome and indeed necessary addition to the roster of tests currently available to assess
cognitive abilities in children. Its strong psychometric characteristics, including its overlapping preschool and school-age batteries, the range of skills measured by the DAS clusters and subtests, and its high cluster and subtest specificity, allow the clinician to feel confident that one is obtaining a valid estimate of the client's cognitive abilities. The DAS offers unique advantages to the clinician in the assessment of a number of special populations frequently encountered in our work. We have argued in this chapter that the DAS is the test of choice in working with children with intellectual disabilities. Moreover, the DAS offers unique insights into the assessment of learning disabilities. Finally, the user-friendliness of the DAS preschool subtests offers major advantages to the clinician working with younger children.

References


3. Assessment with the Differential Ability Scales


SCOPE

The literature about brief cognitive assessment is very large, and it is not possible in a single chapter to comprehensively consider all aspects of this area. This chapter deals mainly with issues about and specific tests for the group probably seen most often by psychologists for brief intelligence testing: school-age children about 5–16 years of age. Considered first are various contexts for brief cognitive testing in school or clinic settings. Desirable characteristics and limitations of brief cognitive tests are discussed next, followed by reviews of some individually administered tests for brief intelligence testing. The occasionally overlooked, but often valuable, role of parent-informant data in screening for child cognitive dysfunction is also addressed. After review of the aforementioned issues and measures, general
recommendations for the best practice of brief cognitive assessment are offered. This chapter concludes with note of issues pertinent to this topic but that cannot here be covered in great detail. These issues include the evaluation of children who are members of minority groups or who do not speak English as a first language.

**INTRODUCTION**

The time required to administer a full-battery, individually administered cognitive ability test such as those discussed in this book—for example, the Wechsler Intelligence Scale for Children—Third Edition (WISC-III; Wechsler, 1991), the Differential Ability Scales (DAS; Elliot, 1990), the Cognitive Assessment System (CAS; Das & Naglieri, 1997), and the Woodcock-Johnson—III (WJ III; Woodcock, McGrew, & Mather, 2001)—can be significant. Although it is often possible to test younger children (e.g., <8 years old) in about an hour or so, older children may require two or more hours of testing time. While various interpretive software programs are available for some full-battery tests, the time required to initially score the results and enter raw test scores and then edit a computer-generated report is still quite substantial.¹

Although there are instances when there is really no substitute for a full-battery evaluation (more about this point, later), psychologists who work in school or clinic settings often get more requests for testing than they can manage within a reasonable time frame. For instance, school psychologists often receive numerous referrals at the very end of the school year. Pediatric psychologists who consult on medical units in hospitals are sometimes asked just before the discharge date to test a child inpatient and make a report. The use of brief cognitive assessment tests is one way to help reduce the length of a waiting list for testing or to determine priorities for a more extensive evaluation. By “brief” it is meant that the administration time is about 30 minutes or less. For very young children, the time saved by using a brief cognitive measure may be relatively little compared to use of a full-battery test, but the absolute time economy for older children will typically be more substantial.

¹Most of the newer intelligence tests are accompanied by supportive software that minimally calculates standard scores and that may provide some tables and figures. Programs such as the WISC-III and WAIS-III Writer are intended to support report writing by generating interpretive statements about test results but also enabling the psychologist to enter other important information about the client (e.g., behavior, family, language). Readers who routinely give full-battery WISC-III can consult Kline (1998b) for a review of the Kaufman WISC-III Integrated Interpretive System (K-WIIS; Kaufman, Kaufman, Doughterty, & Tuttle, 1996). This computer program accepts as input background information, examiner observations about child test behavior, and WISC-III standard scores, and then produces the text of a whole report. Although this program has limitations that are discussed in Kline (1998b), it nevertheless can significantly reduce the time needed to produce a report.
Considered next are some common contexts for which the use of brief measures of intelligence may be potentially valuable.

**CONTEXTS FOR BRIEF COGNITIVE TESTING**

In general, there are three contexts in which the use of brief cognitive ability tests may be especially pragmatic. The first is in mandatory reevaluations of children in public schools who are receiving special education services. That is, provincial, state, or federal guidelines about the definition or funding of specific special education categories may require reevaluations of students who receive such services at regular intervals, say, every 2 or 3 years (e.g., U.S. Department of Education, 1992). For children administered a full-battery intelligence test at their previous assessment, there may be little point in re-administering the whole battery, especially if it is clear, based on grades, teacher reports, or achievement test scores, that the child is not ready to be placed solely in a regular classroom. For such situations, the advice given by Ross-Reynolds (1990) seems very sensible: The time spent giving and scoring a full-battery test in reevaluations may be better spent by using a brief intelligence measure. The time saved can then be devoted to evaluation of the effectiveness of the child's special education program.

A second context for the use of brief intelligence tests is when children are referred for other types of assessments but when it is nevertheless important to screen for overall cognitive status. That is, sometimes referral questions mainly concern issues about personality characteristics or general psychological adjustment. These types of referrals are probably more common in clinic than in school settings, but there are times when examiners decide it is important to use a brief intelligence test even if the referral question did not specifically mention cognitive status. For example, in one instance the author was asked to assess a 5-year-old girl seen in a hospital-based child guidance clinic about her appropriateness for psychotherapy, specifically play therapy. The child was referred by a social worker who was seeing the child's low-income family to help them deal with recent multiple stressors. After interviewing the child, it seemed apparent to the author that her language skills and overall vocabulary breadth were very limited for her age. It is important to note that this child spoke English as a first (and only) language. Screening for cognitive status is much more difficult with multilingual children; see the discussion at the end of this chapter. The child was administered the Slossen Intelligence Test (SIT; Slossen, 1982), a brief measure of verbal skills described later in this chapter. The child's overall SIT IQ score was less than 70, which fell below the 1st percentile for her age. This result helped to quantify the author's impression of limited verbal facility, which was also of concern for the planning of psychotherapy with this child.

Another potential application of brief cognitive ability tests is in the
screening of relatively large numbers of children. For example, some school
districts offer so-called "developmental kindergartens" that offer preschool
experience for at-risk children. These children may be referred to such pro-
grams by general medical practitioners, pediatricians, or social workers. Re-
ferred children may shows signs of developmental delay. The purpose of a
special pre-kindergarten program is to prepare at-risk children for later entry
into regular classrooms or to determine the need for special services. The
availability of quickly administered tests of general cognitive status would fa-
cilitate planning within these kinds of programs.

A final context for brief intelligence tests was mentioned earlier: when a
psychologist has a waiting list of referrals for cognitive assessment. Use of
brief intelligence tests may help psychologists have initial contact with more
children and set priorities for later, more extensive testing. For example, re-
ferred children who are floundering in regular classrooms and who obtain
very low scores on a cognitive screening test may have higher priority for fol-
low-up testing than children with more normal-range scores. Also, ability/
achievement discrepancies suggestive of a learning disability or other signifi-
cant difficulties signal the need for a prompt and comprehensive assessment.
Discussed later in the chapter is the incorporation of parent-informant data
in screening for the need for special education services. Parent-informant
data may also be helpful in the context of brief intelligence testing described
above, the evaluation of large numbers of children.

However time-economical brief intelligence tests may be, there are at
least two situations when their use is inappropriate. The first is for the formal
diagnosis of mental retardation. Diagnostic criteria for mental retardation,
such as those outlined in the fourth edition of the Diagnostic and Statistical
Manual (DSM-IV; American Psychiatric Association, 1994), typically involve
IQ-score cutoffs, such as a requirement that the child's overall IQ score is
less than 70. (The other DSM-IV criterion for a diagnosis of retardation is im-
pairment of adaptive behavior skills; see Sattler, 1988, and Kamphaus, 1993,
for reviews of measures for adaptive behavior assessment.) There are pro-
found implications for a diagnosis of retardation, some of which are poten-
tially beneficial, such as eligibility for special education services for cogni-
tively impaired children. Other implications, though, such as labeling effects
or stigma associated with this disorder, may not be so benevolent. Obvi-
ously, in such instances one needs the most accurate estimate of overall
cognitive status possible. For reasons discussed later, brief intelligence tests
are simply not precise enough for this purpose; only a full-battery, individu-
ally administered measure will suffice. (See Spruill, 1998, for discussion
about the use of the WISC-III and other IQ tests in the assessment of mental
retardation.)

A second context in which brief intelligence tests are inadequate is when
test results may influence determination of the eligibility of referred children
for special education services. This is especially true when eligibility for
learning disability services is considered. Provincial, state, or federal definitions of a learning disability are often based on discrepancies between IQ and achievement scores—either absolute discrepancies or regression-based ones (e.g., Frankenberger & Fronzaglio, 1991; Gridley & Roid, 1998)—and it is thus crucial to accurately measure both domains, that of overall cognitive ability as well as that of specific scholastic achievement. Brief measures of intelligence are again inadequate for a placement decision with important implications for a child’s educational career.

**ESSENTIAL CHARACTERISTICS OF A BRIEF COGNITIVE ABILITY TEST**

Necessary attributes of a brief intelligence test, as outlined in this section, are a combination of basic requirements from measurement theory and specific requirements for the assessment of cognitive status. Properties of the first type include all the attributes necessary for any kind of psychological test, including reliability, validity, a normative sample that is large and representative, and accuracy and clarity of test manuals (e.g., see Standards for educational and psychological testing; American Psychological Association, 1985). The requirement for reliability—which generally involves the consistency, stability, and precision of test scores—is critical, due to the inherent brevity of screening tests. In general, reliability is affected by characteristics of both the sample and the test. Specifically, the reliabilities of test scores are generally lower in samples with limited amounts of individual differences (i.e., range restriction) or when they are derived from tests that have relatively few items, are subjectively scored without clear criteria, are made up of items that vary a great deal in content, or measure phenomena very susceptible to rapid temporal variation, such as current mood states (Nunnally & Bernstein, 1994).

An obvious way to make a test more reliable is to make it longer. The Spearman-Brown prophecy formula estimates the effect of changing the length of a test (e.g., see Nunnally & Bernstein, 1994, pp. 262–264). Suppose that the reliability coefficient for a 10-item measure is only .50, an inadequate value for a useful test. (Recommendations for evaluating whether a test’s reliability is satisfactory will be discussed momentarily.) If the test’s length were to be tripled to 30 items, then the predicted reliability of the longer version generated by the Spearman-Brown prophecy formula is .75. This still is not exceptionally high, but it is better than the original value, .50.² Note that this prophecy formula generates only an estimate; the actual reliability of a test

²The Spearman-Brown prophecy formula is \( r_{\text{predicted}} = \frac{k r_{\text{actual}}}{1 + (k - 1) r_{\text{actual}}} \), where \( r_{\text{actual}} \) is the reliability of the original test, \( k \) is the factor by which its length is changed, and \( r_{\text{predicted}} \) is the estimated (prophesied) reliability of the length-altered test. For the example in text, \( r_{\text{actual}} = .50, k = 3 \), and \( r_{\text{predicted}} = 3(.50)/(1 + (3 - 1).50) = .75 \).
with new items must still be evaluated. Also, the Spearman-Brown formula assumes that the psychometric properties of new items are comparable to those already in the test. Adding poor items to a test can actually reduce its reliability.

It is the general rule that longer tests are more reliable than shorter ones that, in large part, accounts for the superior reliability of full-battery intelligence tests over their brief counterparts. Thus, one crucial consideration in the selection of a brief cognitive ability test is its reliability. Specifically, if the reliability of a brief test is too low, then the administration time it saves may not be worth the cost in terms of low stability of its scores. Now, what is an acceptable level of reliability? Although there is no "gold standard" for reliability, some general guidelines can be offered. Reliability coefficients greater than .90 can be considered "good" and can be used as a minimal standard when test scores affect important decisions about individuals, such as ones discussed earlier about special education eligibility. Reliabilities of IQ or summary scores from modern, well-constructed, full-battery IQ tests are typically .90 or greater (e.g., Kamphaus, 1993; Sattler, 1988). Only some brief intelligence tests (described later) have reliabilities so high. Lower reliabilities, such as values about .70-.80, may be adequate either in earlier stages of test construction and validation or for research concerned mainly with the comparison of groups, such as when the study of group mean differences is of main interest. Reliabilities much lower than .70 are probably inadequate for just about any purpose, much less for individual assessment.

Scores from a brief intelligence test should also be valid, which, broadly speaking, concerns their interpretation. Perhaps the most critical aspect of validity in the realm of brief cognitive testing is the concurrent validity of scores from brief cognitive tests against those derived from full-battery IQ tests, such as Full Scale IQ scores from the WISC-III. Not only should IQ scores from brief tests be highly correlated with full-battery IQ scores—correlations that exceed .80 could be viewed as acceptable—but the absolute levels of scores should be comparable, too. For example, a child who obtains a WISC-III Full Scale IQ score of 90 should ideally obtain a similar score on a brief intelligence test. Note that a high correlation between two sets of scores indicates only similar rank orders, not whether their absolute levels are comparable. For example, suppose that we have a set of IQ scores for a group of children. Let us add 20 points to each child's score and then correlate the original scores with the "enhanced" ones (i.e., the ones with 20 points added). Because the rank order of the two sets of scores is identical, the correlation between them is 1.00 even though (a) the mean scores differ by 20 points and (b) no individual child has scores across the two sets (original and "enhanced") within 20 points of the each other.

Test reliability affects the interpretation of a score as an indicator of the level of some construct. That is, scores from tests with less-than-perfect reliability (which describes all applied tests) should be viewed as a "guess" around which some band of error exists. The notion of confidence intervals
constructed around observed test scores directly expresses this view (see Nunnally & Bernstein, 1994, pp. 258–260). For example, psychologists often report IQ scores with associated confidence intervals as a way to communicate a degree of uncertainty (unreliability) about the scores; the use of 95% confidence intervals, in particular, seems common (e.g., Kaufman, 1994), although interval sizes from 68 to 99% are also used. For example, a child's Full Scale IQ of 110 from the WISC-III may be presented in a test report along with the confidence interval 103–118, which says that a more reasonable estimate of the child's "true" IQ score is the range 103–118 instead of the specific score of 110.

As the reliability of a test goes down, the size of confidence intervals around its observed scores gets ever wider, which conveys increasing error. For IQ scores from full-battery intelligence tests with good reliabilities, the sizes of 95% confidence intervals are about a full standard deviation, which is about \pm 7 points for scores reported in a metric where the mean is 100 and the standard deviation is 15. For example, the confidence interval 103–118 for a WISC-III Full Scale IQ says that a child's "true" IQ score could reasonably vary within a whole standard deviation, 15 points. Of course, a confidence interval of any size does not guarantee that a child's "true" IQ score falls within that range; in fact, it may not. For this example, the child's "true" IQ score could be, say, 120, which falls outside the interval 103–118. Although such an outcome is statistically unlikely, it can still happen, which is a good reason why the upper and lower bounds of a confidence interval should not be seen as absolute limits.

It is suggested later in this chapter that scores from even the best brief intelligence tests should be seen as reasonably varying within a range of at least 10 points on either side of the obtained score. (Again, this is for standard scores where the standard deviation is 15.) This general recommendation seems to suit one of the most psychometrically sound measures described later, the Kaufman Brief Intelligence Test (K-BIT; Kaufman & Kaufman, 1990). Tests not as reliable as the K-BIT will have even greater bands of error around their scores. Greater error is the main reason why brief intelligence tests (even the K-BIT) are not suitable for classification or diagnostic testing of individual children.

Other requirements for brief intelligence tests concern norms derived from a test's standardization sample. Specifically, (1) the norms should be relatively modern and (2) based on a standardization sample that is representative in terms of relevant stratification variables such as gender, race/ethnicity, geographic region, socioeconomic status, and, for children's tests, regular versus special education status. The necessity for a modern set of norms—say, no older than 20 years—is due to a phenomenon sometimes called the "Flynn effect" after J. R. Flynn (1994), who reported that average IQ scores (in a metric where the standard deviation is 15) tend to increase by about 3–4 points (and sometimes more, depending upon the country) every 10 years. These increases may reflect genuine improvements in crystallized
knowledge and/or abstract problem-solving ability in the general population. Such gains could be due to factors like improved educational opportunities, wider or easier access to information (e.g., the Internet), or more widespread experience dealing with abstract stimuli (e.g., through the use of computers). Whatever the sources of the improvements, norms from cognitive ability tests tend to go out of date. For example, suppose that in 1960 the average IQ score in the general population on some test was 100. By the year 2000, the average IQ score on the same test using its 1960 norms may be about 110, due to the Flynn effect. If this same test were re-normed with a contemporary standardization sample, average IQ scores would drop back to about 100. Thus, IQ scores based on older norms may not be directly comparable with IQ scores based on more modern norms; specifically, the former scores tend to be too high, due to the Flynn effect.

There is another problem with the norms of older tests. Standardization samples of older tests tend not to be representative in terms of race/ethnicity. For instance, members of minority groups were not specifically represented in the standardization sample of the Stanford-Binet Intelligence Scale until the 1970s (Thorndike, 1973). This situation violated what is today considered a basic tenet of responsible test use: The interpretation of standard scores is limited to the reference group on which they are based, and the use of a test with groups not represented in its normative sample is problematic. In contrast, the standardization samples of contemporary cognitive ability tests are typically stratified to match multiple aspects of the general population according to current census data. The availability of Canadian norms for the WISC-III (Wechsler, 1996) is an example of this principle.

A final requirement concerns breadth of test content. Ideally, a brief intelligence test should tap more than a single cognitive domain. However, some tests widely used by psychologists as brief ability measures have only one kind of item content. Some of these single-domain tests (among others described later) include the Peabody Picture Vocabulary Test—III (PPVT-III; Dunn & Dunn, 1997), which assesses receptive vocabulary through pictures of familiar objects, and the Raven's Progressive Matrices (RPM; Raven, Court, & Raven, 1986), which taps nonverbal reasoning through pictures of geometric patterns. Although some single-domain measures have psychometric properties comparable to those of brief intelligence tests with more than one kind of item content (e.g., the K-BIT), it is probably better not to place all of one's measurement eggs in a single content basket. Considerations and cautions in the use of single-domain tests as brief intelligence tests are discussed later.

LIMITATIONS OF ALL BRIEF INTELLIGENCE TESTS

The two main limitations of brief intelligence tests were noted in the previous section: Brief tests tend not to be as reliable as or provide information...
about as many ability areas as full-battery tests. Both limitations are probably features of the inevitable trade-off between saving time and getting comprehensive information. Both can also be handled by using brief measures for the screening purposes for which they are intended and never using them for classification or diagnosis, points also discussed earlier.

There is another objection to brief intelligence tests, but this one may have relatively little merit, at least in one regard. The objection is that unlike full-battery scales with their dozen or so subtests, brief intelligence tests do not yield a cognitive ability profile that can be interpreted with scatter analysis. A profile is a graphical summary constructed by drawing lines that connect subtest scores on a record form of a full-battery test, such as the one for many Wechsler scales. Scatter analysis is the interpretation of "peaks and valleys" in a cognitive ability profile. In this approach, relatively high scores are taken to indicate cognitive strengths; by the same token, relatively low scores are believed to show relative cognitive weaknesses. The description of specific cognitive strengths or weaknesses is typically based on rational analyses. For instance, if a child has a relatively low score on the Picture Arrangement subtest of a Wechsler scale, then hypotheses about poor social reasoning skills might be entertained in a scatter analysis. These hypotheses are not based on empirical results; instead, they are ones that seem plausible given the item content of the Picture Arrangement subtest, and must then be "tested" to determine their utility in describing the child's needs and effecting "positive" changes.

Although the practice of scatter analysis of subtest scores dates back to the first intelligence tests, there is actually little empirical evidence that supports the practice. For instance, scores on the Picture Arrangement subtest of the Wechsler scales do not seem to covary with social reasoning skills (see, e.g., Kamphaus, 1998; Lipsitz, Dworkin, & Erlenmeyer-Kimling, 1993). There is similar dearth of evidence for many other subtest-based interpretations (see, e.g., Kline, Snyder, & Castellanos, 1996; McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992). The evidence for the validity of interpretations based on more broad-band, summary scores is much more substantial. For example, differences between Verbal and Performance IQs of Wechsler scales are associated with specific patterns of achievement and psychological adjustment among children with learning disabilities (see, e.g., Rourke, 1998). The inability to interpret a profile of subtest scores on a brief intelligence test may not be such a hindrance after all.

**TYPES AND DESCRIPTIONS OF BRIEF INTELLIGENCE TESTS**

Three categories of brief cognitive measures and specific examples from each are described in this section. The first category includes tests constructed specifically as brief intelligence tests. Tests described from this category
include the K-BIT (Kaufman & Kaufman, 1990) and the Slossen Intelligence Test—Revised (SIT-R; Slossen, 1991). Three other, recently published brief tests are also described: the Wide Range Intelligence Test (WRIT; Glutting, Adams, & Sheslow, 1999) and the Slossen Intelligence Test—Primary (SIT-P; Erford, Vitali, & Slossen, 1999), and the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999).

The second category of brief intelligence measures represents short forms of full-battery IQ tests. Short forms of full-battery tests are usually made up of two or more subtests selected from the whole battery of about a dozen or more subtests. This method of creating a brief intelligence measure may be the most common one in applied settings. It is also a topic for which the literature is quite large. There are some distinct advantages and disadvantages associated with this method, both of which are discussed later along with recommendations for using short forms.

The third category includes single-domain tests used by some practitioners for brief cognitive assessment. Tests of this type include the PPVT-III (Dunn & Dunn, 1997), the RPM (Raven et al., 1986), and the Matrix Analogies Test (MAT; Naglieri, 1985), among others. Some potential limitations of single-domain tests as brief intelligence measures were described earlier. These limitations are reiterated later, when specific single-domain tests are considered.

Tests Specifically Constructed as Brief Intelligence Measures

Kauffman Brief Intelligence Test (K-BIT)

The K-BIT by Kaufman and Kaufman (1990) is a brief measure of cognitive status that can be administered in about 15–30 minutes. Its relatively large normative sample (N = 2022), stratified by gender, geographic region, socioeconomic status, and race/ethnicity according to 1990 census data for the United States, spans ages 4–90 years. Standardization sample age groups from 4–10 years of age each have about 100 cases, age groups from 10 to 20 years may have fewer subjects for each year, and older subjects were combined into much broader age groups (ages 20–34 years, N = 200; 34–54 years, N = 150; and ages 55–90 years, N = 100). Although the test spans a wide age range, it seems that the norms are more adequate for persons younger than 20 years old than for older individuals. For school-age children, though, the K-BIT’s norms seem acceptable.

The K-BIT has two subtests, Vocabulary and Matrices, both of which are presented to subjects via an easel placed on the table that shows associated pictorial stimuli. Two tasks, Expressive Vocabulary (45 items) and Definitions (37 items), make up the Vocabulary subtest. Expressive Vocabulary items are given to subjects of all ages and require naming of pictured objects. The
Definitions part of the Vocabulary subtest requires subjects aged 8 years or older to guess a word that fits two clues, one a partial spelling of the word and the other an oral description of the word. The Definitions task is the only one on the K-BIT to have a maximum time limit (30 minutes). The clue-based format of this task is similar to the Riddles subtest of the Kaufman Assessment Battery for Children (K-ABC; Kaufman & Kaufman, 1983) and to the Definitions subtest of the Kaufman Adolescent and Adult Intelligence Test (KAIT; Kaufman & Kaufman, 1993). All items are dichotomously scored (1/0) according to criteria listed in the K-BIT manual. The number of correct items on the Expressive Vocabulary and Definitions tasks are summed to form the overall score for the K-BIT Vocabulary subtest. Altogether, the K-BIT Vocabulary subtest may measure vocabulary breadth, verbal knowledge through pictures, and word definitions, all of which could also be described as crystallized intelligence.

The second K-BIT subtest, Matrices, is administered to all subjects. This subtest has 48 dichotomously scored items and uses a multiple-choice format. Examinees are required either to select one of five choices that best goes with a stimulus picture or one of six to eight choices that solves a pattern in a $2 \times 2$ or $3 \times 3$ matrix. Early items involve pictures of familiar objects, but stimuli for later items become more abstract. The latter format is very similar to ones used in other matrix reasoning-type tasks, such as the RPM test. Note that the Matrices subtest of the K-BIT does not require a spoken response, as examinees can simply point to their choice. All task stimuli are also pictorial, so there is no manipulation of objects by examinees. The Matrices subtest may measure visual-spatial reasoning or the ability to solve visual analogies, both of which may also reflect fluid intelligence more than crystallized intelligence.

Raw scores on K-BIT Vocabulary and Matrices subtests are converted to standard scores with a mean of 100 and a standard deviation of 15. The two subtest scores are also combined to form the K-BIT Composite, which is the test’s overall estimate of cognitive ability. Standard K-BIT Composite scores have the same mean and standard deviation as for the Vocabulary and Matrices subtests.

Reliabilities of K-BIT subtest and composite scores are generally good. For example, the average split-half reliability for ages 4–19 years for the Vocabulary subtest is .91 (range = .89–.93), for the Matrices subtest it is .85 (range = .74–.92), and for the K-BIT Composite it is .92 (range = .88–.95; these values are somewhat higher for adults aged 20–90 years. Average test-retest reliabilities for a sample of 232 subjects aged 5–89 years evaluated over intervals of 12–145 days (mean = 21 days) were also generally high (Vocabulary, .94; Matrices, .85; Composite, .94).

A test’s reliability coefficients are used to estimate values of the standard error of measurement, which in turn influence the size of confidence intervals around examinees’ scores. Although values of the standard error of
measurement for the K-BIT Composite vary by age, a conservative overall figure would be 5. Multiplying this value by 2 yields a value that approximates the width of a 95% confidence interval. That is, the interval 10 points (i.e., 20 points wide) around a K-BIT Composite standard score reflects the anticipated amount of error at a confidence level often used by psychologists. This derivation is also one of the reasons why it was suggested earlier that scores from even reliable brief intelligence tests, like the K-BIT, should be seen as having a band of uncertainty of about 10 points on either side.

The K-BIT manual (Kaufman & Kaufman, 1990) also provides tables for the comparison of Vocabulary and Matrices standard scores for the same child. These tables are (1) based on information about standard errors of measurement for each subtest, and (2) indicate absolute magnitudes of differences between Vocabulary and Matrices scores required for statistical significance at the .05 and .01 levels. However, K-BIT examiners are advised to use these tables with caution in the interpretation of “significant” Vocabulary-Matrices score differences. Here’s why: First, experience with full-battery IQ tests suggests that statistically significant differences in scale scores, even at the .01 level, are very common in the general population (i.e., the base rate is high). For instance, a WISC-III Verbal-Performance IQ discrepancy of 11 points is statistically significant at the .05 level, but between 25–50% of cases in the test’s normative sample obtain a discrepancy this large or even greater (Kaufman, 1994; Sattler, 1992); these base rates are even higher for some minority groups (e.g., Prifitera, Weiss, & Saklofske, 1998). Finding a statistically significant scale score difference does not imply that it is rare or atypical. Second, results of K-BIT research summarized below suggest that differences between Vocabulary and Matrices subtest scores do not generally relate to external criteria. Finally, the K-BIT manual itself cautions against overinterpretation of Vocabulary-Matrices differences.

Results of about 20 studies described in the test’s manual indicate that correlations between K-BIT Composite scores and IQ scores from full-battery intelligence tests, such as the Wechsler Intelligence Scale for Children—Revised (WISC-R) and the Wechsler Adult Intelligence Scale—Revised (WAIS-R), are about .60–.80 (Kaufman & Kaufman, 1990). These results are generally consistent with those reported in subsequent published studies about the K-BIT, including ones conducted with nonreferred Kindergarten or Grade 1 students (Lassiter & Bardos, 1995), referred children and adolescents (Prewett, 1992b, Prewett & McCaffery, 1993; Slate, Graham, & Bower, 1996), and adjudicated delinquents (Hayes, 1999; Prewett, 1992a).

Several other studies with the K-BIT concerned differences between scores on its Vocabulary and Matrices subtests as predictors of external criteria. For example, within a sample of referred adults, Naugle, Chelune, and Tucker (1993) found high correlations between K-BIT and WAIS-R scale scores (rs of about .70–.90). Correlations between difference scores across the two tests (K-BIT Vocabulary-Matrices, WAIS-R Verbal-Performance), how-
ever, were much lower (rs of about .20–.60). Magnitudes of correlations reported by Canivez (1996) for learning disabled children for similar sets of difference scores from the K-BIT and the WISC-III are comparable.\(^3\) Within a large sample of psychiatric inpatients, Kline and Dorsey (1993) found that K-BIT Vocabulary-Matrices differences were unrelated to level of reading skill.

Findings that differences between K-BIT Vocabulary and Matrices scores may have limited external validity are not surprising. Because difference scores reflect the measurement errors of both of their constituent scores, their reliability may not be high, which in turn limits the absolute magnitudes of the correlations of difference scores with other variables (Nunnally & Bernstein, 1994). This is a general problem of difference scores, whether they are from the K-BIT or any other test. There is also some evidence that the K-BIT Vocabulary-Matrices distinction does not correspond directly to the Verbal-Performance distinction of Wechsler scales. For example, Burton, Naugle, and Schuster (1995) conducted confirmatory analyses (Kline, 1998a) of theoretical models of the factor structure of the K-BIT and WAIS-R within a sample of referred adults. The best-fitting model was one with two separate (albeit correlated) verbal factors, one defined by Verbal scale subtests of the WAIS-R and the other consisting of the Expressive Vocabulary and Definitions tasks of the K-BIT. Differences in item format across the verbal tasks of the two batteries may in part account for this finding. That is, K-BIT verbal items can typically be answered with a single word. In contrast, some items from the Verbal scales of Wechsler tests, such as those from the Comprehension task, require longer, more elaborate responses. Another format difference concerns the K-BIT Matrices subtest, which requires no essential motor involvement, and Wechsler Performance scale subtests that require the manipulation of materials, such as the Block Design task. For all the reasons cited here, (1) the K-BIT Vocabulary-Matrices model should not be viewed as a "mini" or proxy Wechsler Verbal-Performance model, and (2) most of the interpretive efforts for the K-BIT should concern its Composite score.

With the exception of criticism about the limited sizes of normative samples for individuals older than 20 years of age, reviews of the K-BIT have generally been positive (Miller, 1995; Parker, 1993; Young, 1995). Overall, the K-BIT seems to offer sound psychometric characteristics for a screening test that can typically be administered in under 30 minutes. That the K-BIT taps both verbal and visual-spatial abilities (i.e., it is not a single-domain test) is also a plus. Used within the limits of the "±10 rule" for scores from brief intelligence tests, discussed earlier, the K-BIT can be recommended.

\(^3\)The study by Canivez (1996) is a good example of how to study differences in classifications based on a full-battery IQ test and on a brief measure. Specifically, Canivez used the sensitivity-specificity-predictive values, which provides detailed information about correct classification rates of a screening measure classification given estimates of the base rates of some diagnostic entity (e.g., Glaros & Kline, 1988).
Slosen Intelligence Test—Revised (SIT-R)

The original version of the SIT (Slossen, 1963) was a brief intelligence test for ages 6 months to 27 years. SIT items were grouped by age, and the test's total score yielded a ratio IQ. For school-age children or adolescents, the SIT was essentially a single-domain measure of crystallized verbal intelligence because all its items were verbal. The second edition of the SIT (Slossen, 1982) was little changed except for an expanded normative sample. The next major version is the revised 1991 edition (SIT-R; Slossen, 1991) for ages 4 years and older. Although the SIT-R yields a deviation IQ in the form of its Total Standard Score (TSS) and its normative sample is reasonably large (N = 1854), minority groups are underrepresented in the latter. Also, SIT-R items are still primarily verbal and tap mainly language-oriented crystallized intelligence.

Evidence for the reliability and validity of the SIT-R seems generally satisfactory. For instance, internal consistency and test-retest reliabilities reported in the SIT-R manual are all about .80–.90 (Slossen, 1991). Also reported in the manual are correlations between the TSS of the SIT-R and scale scores from full-battery IQ tests; again, these values are typically .70–.90 (Slossen, 1991).

Although reviewers of the 1991 SIT-R generally note improvements over earlier editions, they also mention some important limitations of the SIT-R. Because scores on the SIT-R depend a great deal on English-language fluency, the SIT-R may be inappropriate for use with children who do not speak English as a first language or for screening situations when it is also important to assess nonverbal abilities (Kamphaus, 1995). Watson (1995) noted that analyses summarized in the SIT-R’s manual that purport to demonstrate an absence of test bias may not be technically correct. Both Kamphaus and Watson also criticized the SIT-R manual for not doing more to prevent users from over- or misinterpreting the test’s summary score, the TSS.

To summarize, the SIT-R has a relatively long history but has not generally kept pace with modern methods and standards of test construction. The recent publication of 1998 norms (N = 1800, ages 4–65 years) based more directly on current United States census data represents another in a series of incremental improvements in the test (Slossen, 1998), but it seems difficult to recommend the SIT-R over a psychometrically superior screening measure such as the K-BIT or others described below. The main drawback of the SIT-R is that it is basically a single-domain test of crystallized verbal intelligence, which also limits the usefulness of the test with children who are not fluent in English. The only way to get data about other ability areas would be to supplement use of the SIT-R with other measures, which offsets the advantage of its short administration time.

New Tests for Brief Intelligence Testing

Several new tests for brief cognitive assessment were published very recently. Consequently, it was not possible for the author to carefully review the tech-
nical merits of these tests; readers should consider the descriptions presented here as more informational than critical. The SIT-P (Erford et al., 1999) is for children ages 2–8 years and requires about 10–30 minutes to administer. Unlike the SIT-R, the SIT-P has verbal and performance items. Included among the latter are a symbol search–type task, a paper-and-pencil drawing task, and a block design–type task. The SIT-P yields three standard scores, Verbal, Performance, and a composite Total Standard Score; all three scores have a mean of 100 and a standard deviation of 15. The specific estimation of visual-spatial skills distinguishes the SIT-P from the SIT-R and makes the SIT-P more similar to the K-BIT in this sense. Unlike the K-BIT, the SIT-P has tasks that require the manipulation of objects, such as paper and pencil or blocks, which may convey useful information about fine-motor coordination. The absence of norms for children older than 8 years may limit the usefulness of the SIT-P in school settings, though.

The WRIT (Glutting et al., 1999) is another new brief cognitive ability measure for persons 4–80 years of age that requires about 20–30 minutes to administer. The WRIT's standardization sample is relatively large (N = 2285) and the test was co-normed with the Wide Range Achievement Test—3 (WRAT-3; Wilkinson, 1993). Four subtests (Vocabulary, Verbal Analogies, Matrices, and Diamonds) yield scores that contribute to three standard scale scores, a Verbal-Crystallized IQ (the first two subtests), a Visual-Fluid IQ (the last two subtests), and a General IQ for the whole test. Part of the Diamonds subtest requires the manipulation of objects, so both the WRIT and the SIT-P (but not the K-BIT) have a motor component. Also, the wide age range of the WRIT (4–80 years) relative to the much narrower one of the SIT-P (2–8 years) may favor use of the WRIT in school settings.

Of these two new measures, the WRIT seems to be more adequate as a screening measure than the SIT-P. Both are probably better than the SIT-R, and the WRIT seems psychometrically comparable to the K-BIT, at least based on the author's initial impressions. Both the SIT-P and WRIT need more critical review before well-founded recommendations about their use can be made.

The WASI (Wechsler, 1999) was developed in the tradition of the Wechsler scales. The test was carefully standardized on a representative American sample (N = 2245) and may be used with individuals in the 6–89 age range. The test includes four familiar subtests: Vocabulary, Similarities, Block Design, and Matrix Reasoning, which may be combined to yield Verbal, Performance, and Full Scale (2- and 4-subtest) IQ scores. The psychometric properties are most impressive, with average reliabilities for the FSIQ-4 of .96 for children and .98 for adults; stability coefficients are .93 and .92, respectively. FSIQ-2 and FSIQ-4 correlations with other Wechsler scales such as the WISC-III (.82, .87, respectively) and WAIS-III (.82, .92) are reported, and they have been replicated in independent studies (Saklofske, Caravan, & Schwartz, 2000). Clinical validity data are also presented in the manual. A significant advantage of the WASI is the links to other Wechsler scales thereby providing
some greater robustness to the use of this test as either a screening instrument, preceding the administration of the longer Wechsler scales, or for purposes of quick and efficient retesting and follow-up.

**Short Forms of Full-Battery IQ Tests**

The literature about the development of short forms from full-battery, individually administered IQ tests has a long history and is quite large. In fact, there are numerous citations about this topic for literally every major full-battery IQ test, including the 1911 Binet-Simon (e.g., Doll, 1917), the 1939 Wechsler-Bellevue, Form I (e.g., Robin, 1943), the 1955 WAIS and 1981 WAIS-R, (e.g., Jones, 1962; Kaufman, Ishikuma, & Kaufman-Packer, 1991), the 1974 WISC-R (e.g., Silverstein, 1982), and all contemporary tests for children and adults (e.g., WAIS-III, Ryan & Ward, 1999; WISC-III, Campbell, 1998; Fourth Edition Stanford-Binet, Volker, Guarnaccia, & Scardapane, 1999; K-ABC, Kaufman & Applegate, 1988). Indeed, the manuals of some full-battery intelligence tests include information about short forms. For instance, the manual for the Fourth Edition Stanford-Binet (Thorndike, Hagen, & Sattler, 1986) presents a "quick" four-subtest short form (Vocabulary, Pattern Analysis, Quantitative, Bead Memory) and an "abbreviated" six-subtest short form that adds the Memory for Sentences and the Memory for Digits subtests to the four of the "quick" form.

There are clear practical advantages to short forms of full-battery tests: Not only are many examiners already familiar with the full-battery version, the psychometric characteristics of the subtests are known (i.e., they are summarized in the test's manual). There are two major ways to construct a short form of a full-battery IQ test: (1) select a relatively smaller number of subtests from the whole battery and administer them in their original form, or (2) administer a "split" version of either the whole test or a subset of its tasks. An example of the former is Silverstein's two-subtest short form of the WISC-R that includes the Vocabulary and Block Design subtests, each administered in its entirety and scored using the standard norms tables. A child's total standard score across these two subtests can be converted to an estimated IQ score. An example of a "split" short form is one by Satz and Mogel (1962) for the WAIS: Examiners administered the Digit Span and Digit Symbol subtests in their original form while giving only every third item on Information, Picture Completion, and Vocabulary and every second item on the rest of the subtests. Total raw scores on "split" subtests were then multiplied by a constant before the regular norms tables were used to generate standard scores. Hobby (1980) developed a similar split short form for the WISC-R, in which only the odd items of most subtests were administered.

Of the two methods for making short forms described above, the "split" method has generally fallen out of favor. Some reviewers of short forms, such as Kaufman (1990) and Silverstein (1990), doubted whether the norms for a
full-version subtest applied when only half (or fewer) of its items were administered. For instance, eliminating items not only violates the standard administration of the subtest but also steepens the increase in difficulty across the remaining items due to the reduction in the number of easier ones. The latter condition may also reduce subjects’ opportunities for practice with easier items before they encounter more difficult ones.

The tactic of making short forms of full-battery tests by administering a small number of subtests in their entirety (e.g., a dyad, triad, or tetrad of tasks) may be superior to the “split” method, but is also not without potential problems. For example, using the standard norms tables of a full-battery test when only a few subtests were administered assumes that are no context effects, which means the same scores would presumably be obtained if the whole battery were given. For instance, when the full battery is given, the subtests are given in a specific order, and if changing this order has an effect on subjects’ scores, then there is a context effect (here due to order). Subjects’ motivation levels may also be different if they know that the total testing time will be, say, two hours instead of only 20 minutes. There is indeed evidence for order effects for at least some Wechsler subtests (e.g., Thompson, Howard, & Anderson, 1986); there may very well be order effects for subtests of other IQ tests, too. Also, the reliability and relevance of subtest combinations can vary considerably. For example, the combination of Block Design and Vocabulary from the WISC-III is a much more powerful predictor of achievement than is the combination of Object Assembly and Coding. Another problem concerns how the concurrent validity of short form tests is evaluated. The most common practice is to give a full-battery IQ test and then correlate scores from the whole scale with those from short forms, such as the Vocabulary–Block Design dyad. Because scores from Vocabulary and Block Design contribute substantially to the Full Scale IQ, correlations between the short-form dyad and the full-battery IQ may be spuriously high (e.g., Kaufman, 1990; Silverstein, 1990).

There is no magical solution to the two problems just described—they seem to be inherent limitations of using a dyad, triad, etc., of subtests as a short form in lieu of the whole test. Perhaps the best way for examiners to deal with these issues is not to overinterpret scores from short forms of full-battery tests; that is, not to use short-form scores for diagnostic or classification purposes, and to view such scores as having wider bands of error around them than scores based on whole tests.

A final problem is more a practical one, but it is not an easy one to solve: Which short form to use? For a specific test like one of the Wechsler scales, there may be numerous possible short forms. For an IQ test with 12 subtests, for example, there are 66 different two-subtest short forms, 220 possible three-subtest short forms, and 495 different four-subtest short forms of the whole test. (These values are the numbers of combinations of 12 objects taken two, three, or four at a time, respectively.) Allowing the inclusion of
short forms with at least five subtests increases the total number of candidate short forms even more. There are, typically, numerous different short forms described in the literature for a given full-battery IQ test. In his review of selected short forms of the WAIS-R, for instance, Kaufman (1990) describes at least a dozen different dyad, triad, or tetrad short forms.

Some recommendations are offered to help readers sift through what can be a confusing array of choices about short forms of full-battery intelligence tests. First, one stream of work in the literature that seems to be abating is the search for the “best” short form for specific clinical groups, such as geriatric populations or psychiatric inpatients (Kaufman, 1990). Due to problems like the unreliability of some diagnostic labels used to form special groups and poor generalization of tests developed for specific groups to other clinical populations, it seems more fruitful to focus on making short forms that have good psychometrics and the potential for general application.

Second, a short form of a full-battery IQ test should be made up of at least two subtests, one that taps verbal abilities and another that measures visual-spatial skills. Perhaps the most widely used two-subtest short form of Wechsler scales has been the dyad of Vocabulary from the Verbal scale and Block Design from the Performance scale. Both of these subtests are “g-rich” (g refers to a general ability factor), in that they seem to measure complex reasoning and also tend to be the best predictors of overall IQ scores from their respective scales. Sattler (1992) reports that the estimated reliability of this dyad for the WISC-III is about .90 and that the estimated correlation with Full Scale IQ scores is about .80. Using tables presented in sources like Sattler (1992), it is also possible to convert the sum of two subtest scores into IQ-type summary scores with a mean of 100 and a standard deviation of 15. Although this dyad may arguably be the “best” two-subtest short form, administration times for the Vocabulary and Block Design subtests can be relatively long. The scoring of the Vocabulary subtest also requires the differentiation between 1-point and 2-point responses. Another dyad with psychometric properties almost as good as those for Vocabulary–Block Design but with simpler scoring (0/1) and quicker administration times is Information–Picture Completion. This dyad would not reflect verbal fluency as well as the Vocabulary–Block Design dyad (many Information items can be answered with one word), nor does it involve the manipulation of materials.

Of course, there are other two-subtest short forms of the WISC-III (and other Wechsler scales; e.g., see tables H-6 and L-12 of Sattler, 1992). In general, readers are advised to (1) use dyads with one Verbal scale subtest and one Performance scale subtest, (2) pick a single dyad as a “standard” battery for screening purposes, and (3) avoid using the Digit Span, Symbol Search, and Object Assembly subtests in dyads. Scores from the first two subtests are not used in the derivation of WISC-III IQ scores in the standard norms tables, and the last subtest has less adequate psychometric properties (Kaufman, 1994). Also, readers should (4) “see” a band of error of at least 10 points
around IQ-type scores derived from two-subtest short forms and, as with any brief intelligence test method, should not use such scores for classification or diagnosis.

The next level of short forms for Wechsler tests with equal numbers of Verbal scale and Performance scale tasks are four-subtest short forms (tetrads). Of course, tetrads require greater administration time, but they provide more stable estimates of general cognitive level. Nevertheless, readers should "see" a band of error of at least 8 points on either side of a tetrad-based IQ-type score. A combination of the two dyads described above into a Vocabulary-Information—Picture Completion—Block Design short form yields scores that are quite reliable (> .90) and correlate highly (about .90) with Full Scale IQ scores (Sattler, 1992). Another widely used tetrad is Vocabulary-Arithmetic—Picture Arrangement—Block Design, but the inclusion of the Arithmetic task is not ideal because (1) it may be more of an achievement task than a reasoning task, and (2) it is affected by distractibility or attentional problems. Indeed, the Arithmetic subtest does not, in factor analyses, usually load on the same factor with other Verbal scale subtests (see, e.g., Kaufman, 1994). Selection from among other possible tetrads for Wechsler tests should follow the same general principles given above for dyads.

Also described in the literature for Wechsler scales are numerous short forms made up of five or more subtests, but administration times rise accordingly. Noteworthy among these shorter forms of the WISC-III is an eight-subtest form described by Prifitera et al. (1998; see also Weiss, Saklofske, Prifitera, Chen, & Hildebrand, 1999) that yields a composite score called the General Ability Index (GAI). The GAI is the sum of the four subtest scores that contribute to the WISC-III's Verbal Comprehension Index (VCI) and the four subtests that contribute to the Perceptual Organization Index (POI). Prifitera et al. note that the GAI may better estimate the general reasoning ability of children who obtain very low scores on WISC-III subtests especially susceptible to distractibility or learning problems, specifically the Arithmetic subtest (which contributes to the Verbal IQ but not to the VCI) and the Coding subtest (which contributes to the Performance IQ but not to the POI.)

### Single-Domain Tests as Brief Intelligence Measures

There are numerous quick, single-domain tests that psychologists working in various settings have used as brief intelligence tests. Some of the most popular quick tests are discussed here, but their limitations as a group are outlined first. For the reasons discussed earlier, such as the inability to assess more than one kind of cognitive skill and outdated or nonrepresentative normative samples for some quick tests, the use a single-domain test as the sole instrument in brief cognitive testing is not the best practice. Using two single-domain tests, such as a verbal task and a visual-spatial one, is better
than using just one, but there may be no straightforward way to combine scores from two different tests into a single summary score. This is especially true if the normative samples of two tests are not similar. Also, if a psychologist is going to administer two single-domain tests, then one may as well use brief tests like the K-BIT, WRIT, and WASI, which have two subtests each (two or four for the WASI), or a two-subtest short form of a full-battery IQ test. Either of these alternatives may at least yield meaningful composite scores. With these limitations in mind, individual single-domain measures are considered next.

Perhaps the single-domain test used most often as a brief measure of intelligence is the Peabody Picture Vocabulary Test (PPVT), now in its third edition (PPVT-III; Dunn & Dunn, 1997) and normed for persons 2½ to 90 years old. The test was first published in 1959 (Dunn, 1959) and was revised in 1981 (PPVT-R; Dunn & Dunn, 1981). Although the PPVT has certainly been much improved over time regarding its standardization, normative sample, and manuals, it has retained the same basic format as a multiple-choice test of receptive vocabulary. Briefly, the PPVT-III requires the child to choose the one picture from a group of four that best corresponds to a word spoken by the examiner. Because no verbal response is required from the child and only one correct picture is associated with each stimulus word, the PPVT-III taps a relatively narrow aspect of children's general vocabulary knowledge. For instance, the one word—one correct picture format of the PPVT-III does not allow observation of the child's ability to associate multiple meanings with a word or to choose appropriately among them. The PPVT-III's format also provides no information about the child's ability to combine words for meaningful communication, such as the conveyance of intentions to other people (Sattler, 1988).

As with its predecessors the PPVT and the PPVT-R, correlations between PPVT-III summary scores with IQs from full-battery cognitive ability tests are generally about .70–.90. Not surprisingly, correlations with Verbal IQ scores from the Wechsler scales are among the highest (Dunn & Dunn, 1997). The overall magnitudes of these correlations with full-battery IQ scores are about the same as for the K-BIT, WRIT, SIT-P, and short forms of full-battery tests, so the PPVT-III predicts IQ scores just as well as these other measures. The assessment of a relatively narrow range of ability (i.e., hearing vocabulary) is a drawback relative to tests like the K-BIT, WRIT, or SIT-P, though. To offset this limitation, the PPVT-III could be used in combination with one of tests described below.

Other single-domain tests include the RPM (Raven et al., 1986) and the MAT (Naglieri, 1985). As mentioned, the Matrices subtests of the K-BIT, WRIT, and WASI have essentially the same format as the RPM and MAT: multiple choice, where the child picks the best solution to a visual-geometric analogy. Figural reasoning tests with matrices, like the RPM and MAT, have a long history in the cognitive test literature, and their applications include the study of cultural differences (e.g., Raven & Summers, 1986) and generational
changes in intelligence (e.g., Flynn, 1994), among others. Matrices tests have very good internal consistency and test-retest correlations, and do tend to correlate very highly with each other (e.g., Saklofske, Yakulic, Murray, & Naglieri, 1992). There is also abundant evidence that summary scores from tests like the RPM and MAT correlate with IQ scores from full-battery intelligence tests in the range of about .50-.80 (e.g., Kamphaus, 1993; Sattler, 1988).

Like the PPVT-III, the RPM and MAT predict overall IQ scores just about as well as brief tests like the K-BIT, WRIT, and SIT-P. Also, figural reasoning tests like the RPM and MAT provide useful information about abstract visual-spatial reasoning skills, especially for children with poor motor coordination or with a language or hearing deficit. The RPM in particular has a companion vocabulary test (Raven et al., 1986), which means that verbal ability can be assessed along with visual-spatial skills. This feature of the RPM offsets the potential limitation of assessing a single cognitive domain (i.e., visual-spatial). The MAT (or the RPM without its vocabulary test) could be used together with the PPVT-III, but there may be no straightforward way to combine scores based on different normative samples into a single composite score.

Paper-and-pencil drawing tests are a third major category of single-domain tasks considered here. These tests include the Bender Visual Motor Gestalt Test (or more simply, the Bender-Gestalt [B-G]; Bender, 1938) and the Developmental Test of Visual-Motor Integration (VMI; Beery, 1982), both of which require the child to draw copies of geometric figures of generally increasing complexity. Other tasks involve the drawing of a human figure. There are several variations here, including the Goodenough-Harris Drawing Test (Goodenough, 1926; Harris, 1963) and Draw-a-Person: A Quantitative System (Naglieri, 1988), among others. Drawing tasks are also included in some full-battery IQ tests, such as the Copying subtest of the Fourth Edition Stanford-Binet.

Scoring systems for all the drawing tasks mentioned above are based either on the number of errors in reproducing geometric figures, such as in the Koppitz (1975) system for the B-G, or on the number of details (eyes, ears, nose, etc.) included in human figure drawings. In general, there is a wide-enough range of individual differences among children about 5-10 years of age to reasonably discriminate among them with drawing tasks. Scores on drawing tasks outside this age range are too uniform, either because the tasks are too hard for younger children or too easy for older ones. For example, the average number of errors on the B-G for children older than 12 years of age is about one or two. Drawing tasks like the B-G or VMI may be useful in the screening of adolescents or adults for brain damage because the tasks should be easy for these groups, but, in general, such tasks' usefulness with older children is limited.

4There are also systems to identify so-called “emotional indicators” in children’s human figure or B-G drawings (e.g., Koppitz, 1975; Machover, 1949). However, there is essentially no evidence for the external validity of such “indicators,” and readers are urged not to make personality interpretations of children’s drawings.
The absolute values of correlations of drawing tasks scores with full-battery IQ scores are typically lower than for the PPVT-III or figural reasoning-type tasks, about .20–.60 (Kamphaus, 1993; Sattler, 1988, but see also Saklofske & Braun, 1992). The use of drawing tasks as brief intelligence measures is also limited by potential confounding explanations of low scores. For example, although low-functioning children tend to perform relatively poorly on drawing tasks, the same results can be caused by deficient fine-motor coordination in children of normal overall cognitive ability. For both of these reasons, use of a drawing task as the only brief intelligence measure is not satisfactory. It is much better to use drawing tasks together with other kinds of quick measures. This would be especially advantageous for screening measures that do not require the child to manipulate materials, such as the K-BIT.

There are other, less widely used single-domain tests used for cognitive screening reviewed in Kamphaus (1993) and Sattler (1988) as well as in the Buros Mental Measurement Yearbooks, but it is beyond the scope of this chapter to review them in detail. The next section concerns a type of information that is sometimes overlooked but that nevertheless can be very useful for brief intellectual testing of children.

**PARENT-INFORMANT DATA FOR COGNITIVE SCREENING**

Federal law in the United States requires parental input when children are considered for special education placement (e.g., Federal Register, 1977). Although this demand can be met through relatively unstructured interviews, a more systematic way to gather parental observations (usually from mothers) is with an objective questionnaire. There are now several parent-informant questionnaires that are normed by child age and gender and provide information about child adjustment in several different domains. Examples of parent-informant measures include the Parent Rating Scale (PRS) of the Behavior Assessment System for Children (BASC; Reynolds & Kamphaus, 1992), the Child Behavior Checklist (CBCL; Achenbach, 1991), and the Conners' Parent Rating Scale—Revised (CPRS-R; Conners, 1997), among others.

Although each of the parent-informant measures listed above is psychometrically sound and supported by a wealth of empirical studies about its external validity, all of them are oriented much more toward the identification of emotional or behavioral problems than toward screening for cognitive or academic dysfunction. For instance, the CBCL and the PRS of the BASC have scales that assess adjustment problems like internalizing, externalizing, poor social skills, atypical behavior, or attentional problems, but they lack scales about child cognitive status. The CPRS-R does have a Cognitive Problems scale, but this type of scale is relatively rare among parent-informant inventories (e.g., Kline, 1994). Parent observations about child psychopathology
can be invaluable when the goal is to screen for adjustment problems or when evaluations concern eligibility for classroom placements or services for emotionally impaired children. However, the general absence of scales about child cognitive status on parent-informant measures limits their usefulness for screening in this area. This is unfortunate, because the use of a parent-informant questionnaire can be very time economical. For example, parents can typically complete such questionnaires with little examiner supervision, and some parent questionnaires can be administered with a computer, which afterward can score the responses and generate a profile.

An exception to the general rule that parent-informant questionnaires provide little information about cognitive status is the Personality Inventory for Children (PIC; Lachar, 1992; Wirt, Lachar, Klinedinst, & Seat, 1984). In addition to scales that assess child emotional and behavioral problems, the PIC has three scales about child cognitive status, including Achievement, Intellectual Screening, and Development. There is ample evidence that parent ratings on these and other PIC scales have validity against external criteria about child intellectual and scholastic functioning. For example, correlations between scores on PIC cognitive scales and IQ scores from individually administered tests are about .50–.70 (Kline, Lachar, & Sprague, 1985). The PIC also predicts the placement of children in regular or special education classrooms (Kline, Lachar, Gruber, & Boersma, 1994; Lachar, Kline, & Boersma, 1986). For example, Kline, Lachar, and Boersma (1993) developed within geographically heterogeneous samples of regular and special education students a series of classification rules based on PIC scales that may be useful in screening for cognitive or academic problems. Specifically, the decision rule

\[
\text{IF } (3 \times \text{ACH} + 3 \times \text{IS} + 4 \times \text{DVL} + 3 \times \text{D} + 2 \times \text{FAM} + 2 \times \text{HPR} - 2 \times \text{PSY}) \geq 700 \\
\text{Classify as SPECIAL EDUCATION} \\
\text{ELSE} \\
\text{Classify as REGULAR EDUCATION,}
\]

where ACH, IS, DVL, D, FAM, HPR, and PSY refer to standard scores on (respectively) the Achievement, Intellectual Screening, Development, Depression, Family Relations, Hyperactivity, and Psychosis scales of the PIC, is about 90% accurate in the classification of regular and special education students. Although other decision rules presented by Kline et al. (1993) concern more specific discriminations among special education categories (e.g., emotionally impaired versus learning disabled or cognitively impaired), there is potential value in classification rules like the one presented above for more general screening.\(^5\)

\(^5\)The revised second edition of the PIC should be available about the time this work is published.
A parent-informant questionnaire with scales about child cognitive or scholastic status could be incorporated into a screening battery that also includes an individually administered brief intelligence test, without a great increase in examiner time. This type of approach also more closely approximates the ideal of multi-informant assessment of children, in this case a parent, who provides observations via an objective questionnaire; the examiner, who interacts directly with the child; and the test itself (considered the third "informant"), which provides a standardized and norm-based estimate.

Considered next is another important problem in brief intelligence testing.

**ASSESSMENT OF MINORITY OR IMMIGRANT CHILDREN**

Cognitive assessment of minority children requires special consideration. Even when minority children speak English as a first language, examiners who are not also members of the same minority group need to be aware of subtle differences in language use or subcultural traditions or attitudes that can affect the child's behavior in the testing situation and his or her resulting test scores. Some cultural minorities may view mental health professionals with less trust than others, for instance, or prefer to resolve problems in a family context instead of in individual counseling with an outsider. This challenge is not diminished for brief intelligence testing. Accordingly, psychologists should approach the cognitive screening of minority children with the same preparation as for more extensive cognitive evaluations. Selecting a modern test—one with a contemporary normative sample that includes minority groups and has good psychometric properties—certainly aids in this preparation. Readers can find more information about testing minority children in Jones (1988), Puente and Salazar (1998), and Sattler (1988).

The issues mentioned above are just as pressing in the evaluation of children who do not speak English as a first language, such as the children of immigrants or of groups that have long resided in North America but speak other languages (e.g., French Canadians in Quebec, Hispanic-American children in the United States). Although it is well beyond the scope of this section to deal with the many technical issues involved, there are very few cognitive ability tests in languages other than English that have appropriate normative samples. Such tests, if generally available, would also require examiners who are fluent in the same language. Also, non-English tests would be most useful in places like Quebec, where the majority language is French, but they would have less value in places where English is the main instructional language at school, such as the United States. As readers may know, direct translation of English-language tests to other languages is also very problematic. For example, vocabulary words in English may be, once trans-
lated, more or less common in other languages, which makes the item relatively more easy or difficult in the other language. The use of the original norms tables for translated tests may also be inappropriate.

When children who speak other first languages are tested in English, their scores on brief tests should not be viewed as estimates of their general intelligence. This is especially true for brief tests with predominantly verbal tasks, which may reflect more facility with English rather than verbal reasoning ability or general intelligence per se. Tests with language-based tasks, such as the K-BIT with its Vocabulary subtest or a Wechsler short form with a Verbal scale subtest, should not be avoided; indeed, scores on such tests may provide valuable information about the child's English language skills. Examiners should also not view nonverbal tests, such as the RPM, as somehow providing more "pure" estimates of general intelligence. Readers should remember that scores on nonverbal measures are often highly correlated with scores from verbal tasks. Also, children's performances on nonverbal tasks with relatively complicated instructions may also be affected by limited language skills. Thus, basically the same sets of brief tests can be used with children who do not speak English as a first language, but examiners need to exercise even more caution about test score interpretation.

**SUMMARY: RECOMMENDATIONS FOR BEST PRACTICE**

This chapter reviewed issues about and specific instruments for brief intelligence testing of school-age children. Recommendations for best practice in this area are summarized as follows: Modern tests constructed specifically as brief cognitive measures that assess more than one type of ability are probably the best overall choice. The Kaufman Brief Intelligence Test (K-BIT) and Wechsler Abbreviated Scale of Intelligence (WASI) can be recommended with confidence; new tests like the Wide Range Intelligence Test (WRIT) and the Slossen Intelligence Test—Primary (SIT-P) require more study but seem promising. Because the older Slossen Intelligence Test—Revised (SIT-R) measures essentially just one domain (crystallized verbal intelligence) and is based on an obsolete model of test organization, it cannot be recommended as a brief intelligence measure.

A second-best, but still acceptable, choice is a short form of a full-battery IQ test. Short-form IQ tests typically consist of dyads or, at most, tetrads of subtests. Subtests selected for a short form should, as a set, measure more than one kind of ability; for instance, it would be reasonable to create dyads from Wechsler tests by pairing one subtest from the Verbal scale and one from the Performance scale. A two-or-more-subtest short form should also have good estimated reliability and correlation with IQ scores from the whole version of the test. It is also recommended that examiners select and use the
same short form for all children, which at least standardizes this aspect of cognitive screening.

The use of single-domain tests, such as the Peabody Picture Vocabulary Test—III, the Matrix Analogies Test, or the Raven's Progressive Matrices, as brief intelligence measures is less optimal than the two alternative discussed above. Although these tests have adequate psychometric properties and correlate relatively highly with IQ scores from full-battery tests, the use of one type of item content limits the range of information about a child's cognitive skills, and these tests are, therefore, of less clinical value in differential diagnosis and program creation. These tests are best used in combinations that complement what each individual test measures, such as when a figural reasoning task is used with a verbal task.

Composite scores from any brief measure should never be used for purposes of classification, such as for the determination of eligibility for special education services or for the assignment of a specific diagnosis like mental retardation. It was suggested that an error band of at least ±10 points be "seen" around a composite score from a brief intelligence test. Only a full-battery cognitive ability test administered and interpreted by a skilled psychologist can provide more precise estimates or be used for classification or diagnostic purposes.

Examiners need to be especially cautious when testing children who are members of minority groups or who do not speak English as a first language. Especially for the latter group, it is strongly recommended that scores from brief cognitive measures generally not be interpreted as estimates of general intelligence. Instead, scores for such children are probably best seen as rough estimates of English language proficiency (for verbal tasks) or of visual-spatial ability (for nonverbal tasks), with some language component required for understanding task instructions.

With these recommendations in mind, it is hoped that the time saved by using brief intelligence tests can benefit both examiners and the children they serve.

References


4. Brief Cognitive Assessment of Children


