CHAPTER 5

ASSESSMENT OF ADULT INTELLIGENCE WITH THE WAIS-III

David S. Tulsky
Jianjun Zhu
Aurelio Prifitera

INTRODUCTION

Since the publication of the Wechsler-Bellevue Intelligence Scale for adults in 1939, this scale and its revisions and derivatives, including the Wechsler Adult Intelligence Scale (WAIS) (Wechsler, 1955) and the Wechsler Adult Intelligence Scale-Revised (WAIS-R) (Wechsler, 1981), have had a tremendous influence on the field of psychology (see Kaufman, 1990; Lindemann & Matarazzo, 1984). In studies where the frequency of using assessment instruments has been examined, the Wechsler scales repeatedly come out as one of the most often-used scales. For example, in a study conducted by Harrison, Kaufman, Hickman, and Kaufman (1988), 97 percent of the respondents routinely gave the WAIS-R. More recently, Watkins, Campbell, Neiberding, and Hallmark (1995) reported that 93 percent of the 410 psychologists they surveyed administer the WAIS-R at least occasionally. Other surveys have also found that the Wechsler scales are used on such a frequent basis (Lubin, Larson, & Matarazzo, 1984; Lubin, Larson, Matarazzo, & Seever, 1985; Piotrowski & Keller, 1989).

These scales and especially the development of the new Wechsler Adult Intelligence Scale, Third Edition (WAIS-III) (Wechsler, 1997a), will be the focus of this chapter.

DAVID WECHSLER AND THE WECHSLER INTELLIGENCE SCALES

David Wechsler began using scales of intellectual functioning in his work with the U.S. Army during World War I. Dr. Wechsler was in charge of performing individual testing on people who had failed the group-administered tests. From this experience, he learned which tasks could be used to measure intelligence and used them in his testing sessions. He realized that intelligence could and should be measured by a diverse set of tasks, some verbal and some perceptual; and he saw the need for a new intelligence test, constructed for adults, that emphasized verbal and nonverbal intelligence. This idea of measuring both verbal and performance intelligence (rather than just global intelligence) revolutionized the field of cognitive testing. Wechsler (1944) wrote:

The most obviously useful feature of the Wechsler-Bellevue scales is their division into a Verbal and Performance part....Its a [sic] priori value is that it makes a possible comparison between a subject’s facility in using words and symbols and his ability to manipulate objects, and to perceive visual patterns. In practice this division is substantiated by differences between posited abilities and various occupational aptitudes. Clerical workers and teachers, in general, do much better on verbal tests, whereas manual workers and mechanics do better on perfor-
The correlations are sufficiently high to be of value in vocational guidance, particularly with adolescents of high school age.

Apart from their possible relation to vocational aptitudes, differences between verbal and performance test scores, particularly when large, have a special interest for the clinician because such discrepancies are frequently associated with certain types of mental pathology. (p. 146)

David Wechsler had been well trained in matters of intellectual functioning as well as in merging and integrating what would appear to be a set of diverse ideas about intelligence testing. At Columbia University, Dr. Wechsler spent years training with James McKeen Cattell, E. L. Thorndike, and R. S. Woodworth. He was also fortunate to have spent three months studying with Charles Spearman and Karl Pearson in London, and he took pride in being trained, first and foremost, as a psychometrician. Several of his mentors (Cattell, Thorndike, and Spearman) had strong beliefs about intelligence and intellectual testing, and Wechsler believed that "they were all right" and that he should merge these different viewpoints together into a theory and framework that everyone could accept (Shackelford, 1978).

This goal was more difficult than it might sound because two of his mentors, Thorndike and Spearman, were locked in one of the greatest debates about intelligence testing. Spearman (1904, 1927) believed that intelligence was mediated by a general "g" factor that was responsible for how one would perform on a variety of tasks. Thorndike interpreted the data differently, believing that intellect consisted of several distinct abilities (see Thorndike, Lay, & Dean, 1909). Wechsler had the difficult task of bridging the gap between the beliefs of these two individuals. Throughout his writing, Wechsler (1944) graciously paid tribute to the contributions of both of these great psychologists while not choosing "sides" in the debate.

**Wechsler's Concept of Intelligence**

Wechsler defined intelligence as "the capacity of the individual to act purposefully, to think rationally, and to deal more effectively with his environment" (Wechsler, 1944; p. 3). In this definition of intelligence, he tried to include elements from other leading theorists and researchers of the time (e.g., Thordike, Spearman, Thurstone; see proceedings from the 1921 symposium, Henmon et al., 1921 and Thorndike et al., 1921). Wechsler believed that a definition had to be accepted by one's peers first and foremost in order to gain acceptance (Shackelford, 1978).

Congruent with Spearman's ideas, Wechsler believed that global intelligence was important and meaningful as it measured the individual's overall behavior. However, similar to Thorndike, he also believed it was made up of specific abilities, each of which was important and different from one another. Hence, he emphasized the importance of sampling a variety of intellectual tasks. Wechsler (1974) wrote:

To the extent that tests are particular modes of communication, they may be regarded as different languages. These languages may be easier or harder for different subjects, but it cannot be assumed that one language is necessarily more valid than another. Intelligence can manifest itself in many forms, and an intelligence scale, to be effective as well as fair, must utilize as many different languages (tests) as possible (p. 5).

Bridging the ideas of Spearman and Thorndike, Wechsler (1939) developed a test that included a general intelligence measure (FSIQ) while, at the same time, emphasized that there were two broad types of abilities, Verbal and Performance, that should be analyzed separately to make inferences about an individual's intellectual functioning. The Full-Scale Intelligence Quotient (FSIQ) captures Spearman's idea about a general intelligence, which was characterized as a dominant "g" or general factor with much smaller, less influential "s" or specific factors to guide intelligence. Wechsler agreed with parts of Spearman's theory, namely that there was an overall intelligence. Wechsler even wrote that "Professor Spearman's generalized proof of the two factor theory of human abilities constitutes one of the greatest discoveries of psychology" (Wechsler, 1944; p. 6). Contrary to Spearman's view, however, Wechsler placed more emphasis on the importance of the specific factors and even printed tables so that examiners could review the differences between various types of abilities (e.g., Verbal-Performance discrepancies; Wechsler, 1944).

Thorndike's influence can be seen in Wechsler's writing as he discusses the importance of each subtest and the ability of the examiner to perform profile analyses (e.g., examining differences between subtests). The Wechsler-Bellevue (and all of the derivatives) contains subtests designed to measure
Table 5.1. WAIS: III Subtests Grouped According to Verbal and Performance IQ Scales

<table>
<thead>
<tr>
<th>VERBAL</th>
<th>PERFORMANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary</td>
<td>Picture Completion</td>
</tr>
<tr>
<td>Similarities</td>
<td>Digit Symbol–Coding</td>
</tr>
<tr>
<td>Arithmetic</td>
<td>Block Design</td>
</tr>
<tr>
<td>Digit Span</td>
<td>Matrix Reasoning</td>
</tr>
<tr>
<td>Information</td>
<td>Picture Arrangement</td>
</tr>
<tr>
<td>Comprehension</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2. WAIS-III Subtests Grouped According to Indexes

<table>
<thead>
<tr>
<th>VERBAL COMPREHENSION</th>
<th>PERCEPTUAL ORGANIZATION</th>
<th>WORKING MEMORY</th>
<th>PROCESSING SPEED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary</td>
<td>Picture Completion</td>
<td>Arithmetic</td>
<td>Digit Symbol–Coding</td>
</tr>
<tr>
<td>Similarities</td>
<td>Block Design</td>
<td>Digit Span</td>
<td>Symbol Search</td>
</tr>
<tr>
<td>Information</td>
<td>Matrix Reasoning</td>
<td></td>
<td>Letter-Number Sequencing</td>
</tr>
</tbody>
</table>

qualitatively different types of cognitive abilities like abstract and verbal reasoning (e.g., Similarities, Vocabulary), nonverbal reasoning (e.g., Block Design, Object Assembly), and practical intelligence (e.g., Picture Arrangement, Comprehension). Building a scale that was composed of multiple subtests, each of which could be grouped into different types of intelligence, would allow the scale to match Thorndike’s ideas, while at the same time these abilities could be aggregated into a single “global” score, which would allow the scale to coincide with Spearman’s concepts. Through the structure of the Wechsler-Bellevue, David Wechsler found a way to “walk the fine line” between a global and a multi-factorial model of intellectual functioning.

Despite the many abilities that the Wechsler tests measure, David Wechsler also believed that his scale was not a complete measure of intelligence and that there were some elements missing in his definition of intelligence. He reviewed factor-analytic studies on the Wechsler scales and knew that they only accounted for a percentage of the overall variance of intelligence. From these data, he thought that there must be something else: a group of attributes that contributed to this unexplained variance. Wechsler believed that these attributes, or nonintellective factors, as he called them, were not so much skills as they were traits and included such factors as planning and goal awareness, field dependence, persistence, and enthusiasm (Wechsler, 1950). He believed that these factors contribute to intelligent behavior. These were called the nonintellective aspects of intelligent behavior.

Introduction to the WAIS-III

The WAIS-III is an individually-administered test of intellectual ability for people aged 16–89 years. It is administered in 60–75 minutes and consists of 14 subtests. Like the previous versions, the WAIS-III yields three intelligence composite scores: a Verbal Intelligence Quotient (VIQ), a Performance Intelligence Quotient (PIQ), and a Full Scale Intelligence Quotient (FSIQ). The IQs have a mean of 100 and a standard deviation of 15. Table 5.1 shows the set of six Verbal and five Performance subtests that can be combined to yield VIQ, PIQ, and FSIQ on the WAIS-III. A new Matrix Reasoning subtest has replaced Object Assembly (used in previous Wechsler editions) on the Performance and Full-Scale IQ score.

Object Assembly has been included as an optional subtest for the IQ scales. It can be used to replace a spoiled Performance subtest when deriving IQ scores or it can replace another Performance subtest during retesting to help reduce the practice effects. Also, for those who want to use the same subtests as on the WAIS-R, to calculate PIQ and FSIQ, Object Assembly can be substituted for Matrix Reasoning.

A different subset of 11 subtests can also be combined to obtain a set of four index scores. Table 5.2 lists these subtests and how they relate to...
the four Index scores: Verbal Comprehension Index (VCI), Perceptual Organization Index (POI), Working Memory Index (WMI), and Processing Speed Index (PSI). These index scores consist of more refined domains of cognitive functioning than do the IQ scores. For practical reasons, the index scores were limited to 11 subtests, with three subtests each for the VCI, POI, and WMI, and two subtests for the PSI.

The subtests vary in content from tasks such as defining vocabulary words, stating abstract relations between two objects or concepts, repeating a string of digits, putting puzzles together, putting blocks together to match a pattern, and sequencing a set of pictures to tell a story. Descriptions of each subtest and what they are purported to assess are discussed in the literature (Matarazzo, 1972; Kaufman, 1991, 1994; Sattler, 1992).

The scoring of each subtest differs. Some are dichotomously scored, some have consistent partial credits (0, 1, 2) and some vary because of differential weighting and time bonuses among the items. On 10 of the subtests, the item order is based on difficulty, which we believe approximates a Guttman pattern (Guttman, 1944). There are discontinue rules (e.g., 3 consecutive scores of 0), that are built on the assumption that the examinee would receive scores of 0 on any items that would be administered beyond the discontinue rule. This serves to reduce administration time and to not tax an individual. Digit Symbol-Coding and Symbol Search differ from the other subtests in that they are timed subtests on which the examinee completes as many items as possible within a 120-second time limit.

Scaled scores are presented in a lookup table based on the sum of the item scores for each subtest by age group. The WAIS-III deviates from its predecessors by basing subtest scores on age corrected scaled scores rather than on the performance of a younger reference group made up of individuals between the ages of 20 and 34 years. The distribution of each subtest was normed to a scale with a mean of 10 and standard deviation of 3. The subtest scores are normed according to 13 age bands (ranging from 16 to 89 years). These age-corrected scaled scores would then be summed to develop composite IQ or Index Scores.

Goals of the WAIS-III Revision

The first goal of the revision, to update the norms, stems from the fact that the normative information for intelligence tests becomes outdated over time and IQ scores become inflated. Joseph Matarazzo (1972) and James Flynn (1984, 1987) have written about this phenomenon of shifts in IQ norms. Dr. Matarazzo wrote that "it is imperative that such [age] norms be periodically updated lest they be less than fully efficient for the re-examination of individuals living in a social-cultural-educational milieu potentially very different from the one which influenced the individuals constituting the norms for that same age group in an earlier era." (Matarazzo, 1972; p. 11). Flynn’s systematic review of this issue has shown that IQ scores tend to become inflated over time (Flynn, 1984) with the average IQ score drifting upward. Individuals appear to gain approximately 3-5 IQ points over a 10-year period. Generally, the phenomenon is more prevalent in the performance scales than it is in the verbal scales.

Based on these findings, the WAIS-III contains a contemporary, representative sample from which the IQ norms have been “re-anchored” at 100. Comparisons between the WAIS-III and WAIS-R scores reveal how “outdated” the norms on the WAIS-R had become. The WAIS-III—WMS-III Technical Manual (The Psychological Corporation, 1997) reports data on 192 individuals who completed both the WAIS-R and the WAIS-III. Examinees took the two scales in two sessions, 2–12 weeks apart, in a counterbalanced order. Consistent with the a priori predictions, the average FSIQ and PIQ scores were higher for the WAIS-R than the WAIS-III and the VIQ scores were relatively unchanged. The average FSIQ score on the WAIS-R was 2.9 IQ points higher than the corresponding average score on the WAIS-III, and the WAIS-R PIQ score was 4.8 IQ points higher. This finding adds further support to the hypothesis that IQ inflation is truly occurring. This inflation rate, however, is slightly lower than that which would have been expected from previously reported values (Flynn, 1984). Based upon the so-called “Flynn effect” alone, the average FSIQ would be increasing at a constant rate each year (e.g., an increase of one-third to one-half points per year), so that the average FSIQ of the WAIS-R would have been expected to be as high as 106–109 IQ points.

There are several reasons that the WAIS-R and the WAIS-III differences might be lower than predicted (see Zhu & Tulsky, 1999). Simply adding a constant oversimplifies the relation between the two tests. Besides this overall “Flynn effect,”
many other factors, such as practice effect, design differences between the two tests, floor and ceiling effect, other psychometric factors, (Bracken, 1988; Kamphaus, 1993; Zhu & Tulsky, 1997), and the interaction among these factors may affect the score discrepancies across the two testings. For instance, there are some significant differences between the WAIS-R and the WAIS-III that may be accounting for some of the differences. Most salient, the replacing of Object Assembly with Matrix Reasoning and the de-emphasis of timed bonus points in the WAIS-III may explain the difference between the two measures. Additionally, careful effort was taken to ensure that a representative proportion of individuals across the entire range of ability was sampled on the WAIS-III. To prevent truncated norms, 29 examinees with mental retardation were added to the overall standardization sample to ensure that the correct proportion of examinees (approximately 2.3 percent) that had FSIQ scores below 70 were included in the sample (Tulsky & Zhu, 1997). This effort may shrink the difference between the WAIS-R and WAIS-III.

The second goal of the revision was to extend the age range. Individuals in the United States are living longer. Current estimates place the average life expectancy at birth at more than 78 years for women and 72 years for men (Rosenberg, Ventura, Maurer, Heuser, & Freedman, 1996; La Rue, 1992). However, the WAIS-R only has normative information for people up to 74 years of age, and hence, it is becoming less sufficient for estimating the intelligence of older adults. Previously, to compensate for this deficit, two independent research teams have conducted studies to extend the WAIS-R norms upward for an older adult population. Ryan, Paolo, & Brungardt (1990) developed norms for older adults using a sample of 130 people (60 individuals who were between the ages of 75 and 79 years and 70 who were 80 years old and up). Attempts were made to match the sampling stratification criteria of the WAIS-R as much as possible. Concurrently, in an independent project, researchers at the Mayo Clinic collected normative data on 512 individuals between 56 and 97 years of age (Ivnik, et al., 1992). They deviated from the WAIS-R scoring technique by developing “age-specific” raw-score-to-scale-score conversions rather than basing the conversion on the optimal functioning “reference” group. Using the 56–74 year-old sample as a reference point, the research group also spent a considerable amount of time investigating the similarities between the Mayo Older Adult Normative Studies (MOANS) norms and the WAIS-R standardization sample norms so that they could make their norms as similar as possible to the WAIS-R. For the WAIS-III, the goal was to extend the normative information up to 89 years of age, allowing for appropriate use of scores for individuals in this older age range.

A third goal was to improve the item content of the subtests. A number of items were outdated and needed replacement. Additionally, some examiners have criticized the WAIS-R for containing some items that appear to be biased against certain groups. Extensive bias analyses and reviews were conducted so that biased items could be removed and replaced in the new revision (Chen, Tulsky, & Tang, 1997).

The fourth goal of the project was to update the artwork and make the WAIS-III more attractive for examinees. The WAIS-R was published in 1981 using the styles from the original Wechsler-Bellevue. Not only was some of the artwork outdated and unattractive, but some of the visual stimuli were small, putting individuals with visual acuity problems at a disadvantage. Several steps were taken to make the WAIS-III stimuli more appropriate for examinees. The Picture-Completion items were redrawn, enlarged, and colorized and the Picture-Arrangement cards were redrawn, enlarged, and modernized. The Digit Symbol-Coding subtest features more space between the items and keys to help assist left-handed examinees who might otherwise block the key as they were working. Finally, the WAIS-III Object-Assembly layout shield was modified radically to include the subtest instructions, and it was constructed of heavy card stock so that it could stand up on the table. The puzzle pieces themselves have numbers printed on the back to assist the examiner in laying out the pieces.

The fifth goal was to enhance the clinical utility of the scale, and this was accomplished in several ways. First, additional index scores were included in the WAIS-III. Some researchers have written about the limitations of the IQ score (Kaplan, 1988; Lezak, 1988, 1995). Others have suggested that the scale should measure a wider spectrum of domains of cognitive functioning (Malec et al., 1992). To incorporate some of the advances in the field, when the Wechsler Intelligence Scale for Children-Third Edition (WISC-III; Wechsler, 1991) was published, new factor-based Index scores (e.g., Verbal Comprehension, Perceptual Organizational, Freedom from Distractibility, and
Processing Speed) were added in addition to the traditional IQ composite scores. The WAIS-III revision includes a similar alternate index-scoring system, in addition to the traditional IQ-scoring system. New optional subtests have been developed to assess abilities on a hypothesized 3rd factor (Working Memory) and a 4th factor (Processing Speed). Specifically, Letter-Number Sequencing was included to measure Working Memory, and a second subtest, Symbol Search, was designed to measure Processing Speed.

Additionally, some optional procedures, such as testing incidental learning after the Digit Symbol-Coding administration (Hart, Kwentus, Wade, & Hamer, 1987; Kaplan, Fein, Morris, & Delis, 1991), were added to the WAIS-III-standardization edition. These procedures were based on the “process approach” to interpretation that was advocated by Kaplan and others. They were designed to help the examiner determine the nature of errors committed on the standardized tests.

The WAIS-III Administration and Scoring Manual (Wechsler, 1997a) includes optional normative tables designed to assist the clinician in the interpretation of scores. Besides the critical values for statistical significance of discrepancy, base rates of discrepancies between scores are presented in the manual. Matarazzo & Herman (1985) were the first to publish such tables based on the WAIS-R standardization sample and they demonstrated that VIQ-PIQ difference scores could be statistically significant but not clinically meaningful. Statistically significant scores would suggest that the difference score was “real” or that it was significantly different from 0. The base rates show how frequently such differences do occur in the population and even though someone might be better at one skill (Verbal or Performance) than the other skill, it might occur in a large percentage of the general population. These base-rate tables, therefore, allow the clinician to interpret the score based on the frequency at which such discrepancies occur.

Significant effort was made to enhance the measurement of the WAIS-III in individuals with very low or impaired intellectual functioning and other clinically relevant groups (e.g., people with mental retardation, people with neuropsychological impairment). With the WAIS-R, a 70–74-year-old person who cannot answer one item correctly can still receive a VIQ score of 60 and a PIQ score of 61 points! This was likely a result of the subtests having a restricted floor, the normative sample possibly not containing enough individuals whose true score extended that low, and the subtest scaled scores not extending more than 3 standard deviations below average. The floor of the WAIS-III extends lower than its predecessors, extending down to 45 for FSIQ, 47 for PIQ, and 48 for VIQ. To help validate that accurate scores were being obtained for people with low intellectual functioning, data on 62 people with moderate mental retardation and 46 people with mild mental retardation were obtained. The original diagnosis for each examinee was made using DSM-IV criteria (which included an appropriate score on an IQ test (other than the WAIS-III) and impairment in adaptive functioning. Roughly 83 percent of IQ scores in the mild group had WAIS-III IQ scores between 53 and 70 and 82 percent of the WAIS-III scores for the examinees in the moderate group had IQ scores between 45 and 52 (Tulsky & Zhu, 1997).

The sixth goal was to decrease the emphasis on timed performance. One criticism of the WAIS-R has been that some of the subtests are too dependent upon quick performance (Kaufman, 1990). For instance, on the Object Assembly subtest of the WAIS-R, in which subjects put puzzle pieces together, an examinee may earn up to 12 raw-score points (e.g., 29 percent additional raw-score points) as time-bonus points for speedy performance. This could result in a difference between 7 and 10 subtest scaled-score points. Hence, another objective was to reduce the contribution of speed and bonus points to the Performance IQ score wherever it is possible. To help achieve this goal, a new untimed performance subtest, Matrix Reasoning, was included.

The seventh goal was to enhance the measurement of fluid reasoning. Several recent theories of cognitive functioning have emphasized the importance of measuring fluid reasoning, or the ability to perform abstract mental operations (Sternberg, 1995). Matrix-reasoning tasks are considered typical of this type of ability, hence, the addition of this subtest to the WAIS-III.

Eighth, the theoretical structure of the WAIS-III was strengthened. Contemporary research has pointed out that intelligence encompasses more than what is measured by VIQ and PIQ scores (Carroll, 1993; Carroll, 1997). Reviews of factor-analytic work on the Wechsler scales have suggested that there are either three domains of cognitive functioning (Cohen, 1952a, 1952b, 1957a, 1957b, 1959; Leckliter, Matarazzo, & Silverstein, 1986) or, in the children’s version after an optional Symbol Search subtest was included, that there are four domains of
cognitive functioning (Wechsler, 1991; Roid, Priftiera, & Weiss, 1993). Current theories of Working Memory (e.g., Baddeley, 1986; Kyllonen, 1987; Kyllonen & Christal, 1990) and Information Processing (e.g., Kyllonen, 1987) were used in developing new additional subtests on the WAIS-III. These subtests help expand the domains of cognitive functioning that are measured by the WAIS-III.

Ninth, the WAIS-III is linked with other tests such as Wechsler Individual Achievement Test (The Psychological Corporation, 1992) and Wechsler Memory Scale-Third Edition (WMS-III) (Wechsler, 1997b) to help the clinician interpret scores and patterns of scores. Significantly, the standardization sample was co-normed with the WMS-III. This linkage allows clinicians to examine IQ and memory relationships and discrepancy scores. Moreover, the linkage assists them in the interpretation of additional domains of cognitive functioning that include both intelligence and memory assessment.

Finally, extensive work has been performed to validate the new instrument and to demonstrate comparability between the WAIS-III and WAIS-R. Correlations between the WAIS-III and the WAIS-R, WISC-III, and the Stanford-Binet, 4th Edition, demonstrate that the WAIS-III is correlated with other instruments measuring intellectual functioning (The Psychological Corporation, 1997). The correlations between FSIQ on the WAIS-III and the general composite scores of these other instruments range from .88 to .93. The correlation of FSIQ with the Raven’s Standard Progressive Matrices (SPM) (Raven, 1976), a nonverbal task of abstract ability, is lower, \( r = .64 \); however, as expected, SPM has higher correlations with PIQ \( r = .79 \) and the Matrix Reasoning subtest on the WAIS-III \( r = .81 \).

The WAIS-III was also tested in a series of clinical validity studies with more than 600 individuals with neuropsychological impairment (e.g., Alzheimer’s dementia, traumatic brain injury), psychiatric diagnosis (e.g., schizophrenia, depression), learning disabilities, mental retardation, and hearing impairment or deafness. From these studies, different patterns of performance tended to occur (especially among the index scores) and they provided an initial demonstration of the construct validity and clinical utility of the WAIS-III. A detailed description of these studies has been reported in The WAIS-III–WMS-III Technical Manual (The Psychological Corporation, 1997).

**Development of the New WAIS-III Subtests**

To enhance the measure of fluid reasoning, working memory, and processing speed, the WAIS-III includes three new subtests: Matrix Reasoning, Symbol Search, and Letter-Number Sequencing. The development of these new subtests will be described in the following sections.

**Matrix Reasoning**

In the WAIS-III, the new Matrix Reasoning subtest replaces Object Assembly as a standard subtest and contributes to PIQ, FSIQ, and POI scores. As stated earlier, this subtest was added because it has long been recognized that matrix analogy tasks are good measures of “fluid” intelligence (Sternberg, 1995) and reliable estimates of general cognitive/intellectual ability or “g” (Brody, 1992; Raven, Raven, & Court, 1991). Studies have shown that IQ indices on matrix analogy tests are highly correlated with the IQ scores of the Wechsler scales (Desai, 1955; Hall, 1957; Levine & Iscoe, 1954; Watson & Klett, 1974). Research also demonstrates that, in general, matrix analogy tasks correlated higher with performance subtests than with verbal subtests of the Wechsler intelligence scales. In addition, matrix reasoning tasks are considered to be relatively culture-fair and language-free, requiring no hand manipulation and having no time limits. These features make it an appealing measure of PIQ, particularly with older adults and minorities. Such a measure also allows for contrasts with other nonverbal reasoning tasks, such as Block Design. When performance on Block Design is low, for example, the hypothesis that a person’s score may have been affected because he or she responds slowly on a timed test can be evaluated by comparison with an untimed reasoning test. Such contrasts allow for more meaningful interpretation of test scores and performance.

The Matrix Reasoning subtest was developed after careful theory and content review of the existing literature. It contains 26 items: 3 basal items and 23 regular items. Four types of items were designed to provide a reliable measure of visual information-processing and abstract reasoning skills. These four types of matrices are continuous and discrete pattern completion, classification, analogy reasoning, and serial reasoning. They are commonly seen in existing matrix-analogy tasks.
such as Raven's (1976) *Standard Progressive Matrices* and Cattell's (1973) *Culture Fair.* Figure 5.1 provides some examples of each type of item.

In addition to the type of matrices included in the subtest, content coverage was also influenced by two other dimensions. The first dimension includes the features and types of stimuli that can be manipulated during the problem-solving process. Attributes of the stimulus, such as color, pattern, shape, size, position, direction, and the number of attributes included in an item, were manipulated or controlled for each item. The second dimension involves the mental tasks performed during the problem-solving process, such as folding, rotating, mirroring, switching, cutting, adding, and flipping. A number of these tasks were carefully selected for each item. A progression of difficulty was developed by adding more stimuli and mental tasks from these two dimensions. The test format is multiple choice. For each item, the four foils among the five-choice answers were very carefully designed to enhance item-discrimination ability. There is no time limit for this test, but data from the WAIS-III standardization suggest that most individuals will provide answers within 10 to 30 seconds.

The reliability coefficients across the different age groups range from .84 to .94, with an average of .90, which is much higher than the Object Assembly subtest (.70) that it replaces. Moreover, the Matrix-Reasoning subtest minimizes speed and motor responses, and for the majority of examinees in the standardization sample, it takes less time to complete than the Object Assembly subtest, thus reducing overall test-administration time with most examinees. Data analysis indicated that the Matrix Reasoning subtest correlates the highest with Block Design (.60), and in factor analysis, loads on a factor made up by subtests measuring Perceptual Organization. Results of two validity
studies using samples of 26 nonclinical adults and 22 adults with schizophrenia found that the WAIS-III Matrix Reasoning subtest correlates at .81 and .79 with Raven’s Progressive Matrices, respectively.

There is a legitimate concern that, because this subtest is untimed, there is a potential for the administration time of this subtest to become quite lengthy. However, the benefits of having a performance subtest measuring abstract, fluid ability independent of time outweigh the potential problems. As mentioned previously, examiners now have a subtest that can be contrasted to the other WAIS-III subtests (e.g., Block Design) that place a high emphasis on timing and bonus points. Moreover, Tulsky and Chen (1998) using the WAIS-III standardization sample estimated that examinees tend to complete the Matrix Reasoning subtest quickly, generally in seven minutes or less. These estimates indicate that the median time for the subtest is 6.4 minutes, with 90 percent of examinees completing the subtest in 11.9 minutes. Comparatively, the estimated median time to complete Object Assembly is 10.7 minutes. Therefore, when contrasted with the Object Assembly subtest that it replaces, Matrix Reasoning is much shorter.

At the item level, the data show a similar trend. Almost 75 percent of the items were completed within 15 seconds and more than 90 percent were completed within 30 seconds. This supports the theory that, in general, examinees will respond quickly to these items. Occasionally, however, there will be individuals who take longer to answer the items. Based upon the data obtained from the 2,450 examinees who completed the standardization sample, it seems that additional time will not increase scores. Of those examinees who took longer than 60 seconds per item, the responses were wrong two-thirds of the time. This rate would be higher if the guessing factor was considered. This finding can be used to help guide examiners when administering the test. If an examinee has performed quite well on the scale and then takes additional time to solve the items as difficulty increases, the examiner should grant such leeway. Alternatively, if the examinee has low and inhibited output and tends to ruminate on items without any perceived benefit, the examiner should encourage the examinee to respond after 30 seconds or so, and definitely move him or her along after 45 to 60 seconds.

Symbol Search

The WAIS-III Symbol-Search subtest is designed to measure an individual’s speed at processing new information. In this task, the examinee is presented with a series of paired groups, each pair consisting of a target group and a search group. The examinee’s task is to decide whether either of the target symbols is in the search group, a group of five search-symbols.

A similar task was developed and included in the WISC-III as a supplemental subtest contributing to the 4th factor, Processing Speed (Kaufman, 1991; Wechsler, 1991; Roid, Prifitera, & Weiss, 1993; Carroll, 1993; Kamphaus, Benson, Hutchinson, & Platt, 1994). The purpose of including this subtest in the WAIS-III is to enhance the measure of processing speed of the instrument and to bring out the four-factor structure that was found on the WISC-III.

During the development of the WAIS-III Symbol Search, the following guidelines were used. First, to minimize the potential involvement of verbal encoding, only nonsense symbols were used. Second, because some nonsense symbols can be verbally coded more easily than others, the difficulty of each item was carefully evaluated across all age groups to make sure that there were no significant differences in difficulty across all items. Third, since the tasks of Symbol Search are mainly visual discrimination and visuo-perceptual scanning (Sattler, 1992), the difficulty of the test items affects the factor-loading of this subtest. If the items are too difficult, the test will tend to load more on the perceptual organization or working-memory factors rather than the speed-of-information processing factor. Therefore, the range of item difficulty is set at .80–1.00.

The test-retest reliability is .79 for the overall test-retest sample (n = 394), with a range from .74 to .82. Factor analysis suggests that the WAIS-III Symbol Search, along with Digit Symbol loads highest on the Processing Speed Index. Correlation analysis also suggests that this test correlates the highest with Digit Symbol-coding (.65).

In the WAIS-R, the Digit Symbol-coding subtest contributed the most unique variance to the scale. With the addition of Symbol Search in WAIS-III, a new dimension of functioning can now be measured. This new area of functioning appears to be sensitive to a variety of clinical conditions, such as Parkinson’s Disease, Huntington’s Disease, and Learning Disabilities (to name a few). Also,
Table 5.3. Reliability Coefficients of Object Assembly (OA) and Matrix Reasoning (MR)

<table>
<thead>
<tr>
<th>Subtest</th>
<th>16-17</th>
<th>18-19</th>
<th>20-24</th>
<th>25-29</th>
<th>30-34</th>
<th>35-44</th>
<th>45-54</th>
<th>55-64</th>
<th>65-69</th>
<th>70-74</th>
<th>75-79</th>
<th>80-84</th>
<th>85-89</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>OA</td>
<td>.73</td>
<td>.70</td>
<td>.73</td>
<td>.71</td>
<td>.75</td>
<td>.71</td>
<td>.78</td>
<td>.72</td>
<td>.77</td>
<td>.68</td>
<td>.59</td>
<td>.64</td>
<td>.50</td>
<td>.70</td>
</tr>
<tr>
<td>MR</td>
<td>.87</td>
<td>.89</td>
<td>.88</td>
<td>.91</td>
<td>.88</td>
<td>.91</td>
<td>.89</td>
<td>.93</td>
<td>.94</td>
<td>.91</td>
<td>.90</td>
<td>.89</td>
<td>.84</td>
<td>.90</td>
</tr>
</tbody>
</table>

Note: From WAIS-III-WMS-III Technical Manual. Copyright 1997 by The Psychological Corporation. Reproduced by permission. All rights reserved.

because Symbol Search requires less motor skill than Digit Symbol-coding, contrasting the two sub-tests can provide useful clinical information on the extent of motor involvement on low scores.

Letter-Number Sequencing

This is a new subtest designed to measure working memory. It was based on the work of James Gold and his colleagues at the University of Maryland (Gold, Carpenter, Randolph, Goldberg, & Weinberger, 1997). In this test, participants were presented with a mixed list of numbers and letters. Their task is to repeat the list by saying the numbers first in ascending order and then the letters in alphabetical order.

The reliability coefficients of the subtest are fairly good, ranging from .75 to .88, with an average of .79. Data-analysis results suggested that this test correlates the highest with other working memory measures, .55 with Arithmetic, and .57 with Digit Span. Factor analysis suggested that it loads substantially on working memory, together with Arithmetic and Digit Span.

The Content of the IQ scores

In developing the IQ scores, the decision to make Object Assembly optional may be considered problematic and controversial for several reasons. Object Assembly has been a core sub-test on the Wechsler scales since their inception. Therefore, many clinicians are familiar with the performance on this subtest in various clinical populations. Also, years of research on previous Wechsler editions provide empirical support for use and interpretation of this subtest. Matrix Reasoning does not have this historical and empirical base within the Wechsler clinical and research literature.

Matrix Reasoning was designed to help assess nonverbal, fluid reasoning in an untimed manner. When work on the WAIS-III began, Matrix Reasoning was going to be an optional subtest that the examiner could use if he or she had questions about an individual’s nonverbal ability and wanted to measure it independently from speeded or timed tasks or both.

The decision to replace Object Assembly with Matrix Reasoning in the IQ scores was instituted for a number of reasons. First, the statistical properties of Matrix Reasoning are far superior to those of Object Assembly. As shown in Table 5.3, the reliability of Matrix Reasoning is much higher than the reliability of Object Assembly. For Matrix Reasoning, the average of the split-half reliability coefficients is .90, which is significantly higher than the .70 average of the coefficients that was obtained on the Object Assembly subtest. Substituting Object Assembly with Matrix Reasoning allows for a smaller standard error of measurement and tighter confidence intervals in the determination of the Performance IQ. Furthermore, as can be seen in Table 5.3, the reliability coefficients of Object Assembly for adults 75 years or older are fairly low, making the measurement error too high to obtain a valid assessment of skills. This low reliability may be due, in part, to the incorporation of bonus points for quick performance on the Object Assembly subtest. Older adults generally perform at a slower pace, and this fact alone makes the Object Assembly subtest more problematic. Also, in the majority of cases, the administration time for Matrix Reasoning is less than the time needed for Object Assembly.

In deciding to replace Object Assembly with Matrix Reasoning on the IQ scores, however, a series of analyses were performed to determine if such a replacement would affect the nature of the Performance IQ scale. In one case, two “alternate” Performance sums of scaled (PSS) scores were developed and compared. The first score (PSS₁) was developed by summing scores on the following subtests: Picture Completion, Block Design, Picture Arrangement, Digit Symbol-Coding, and Object Assembly.

The second score (PSS₂) was the sum of scores on Picture Completion, Block Design, Picture
Arrangement, Digit Symbol-Coding, and Matrix Reasoning. Both summed scores were then converted to a Wechsler score with a mean of 100 and a standard deviation of 15.

Differences between these two perceptual sums of scaled scores indicated that 14 (out of the 2,450) individuals (or 1.7 percent) had difference scores of more than 0.67 SD (standard deviation) units. More important was the question of how many individuals would have fallen outside of the confidence interval of the PSS1 score. Only 2 people out of the 2,450 examinees would have had a difference that significant. This indicated that there was not too great a change in the composite scores. Providing additional evidence that such a change would improve the IQ scores, the reliability of PSS2 was slightly higher (r = .94 for PSS2 versus r = .93 for PSS1) and the standard error of estimate is slightly smaller (SEE [standard error of the estimate] = 3.41 for PSS2 versus SEE = 3.79 for PSS1).

The Domains of Intelligence: From Factor Analytic Studies to the Development of Index Scores

Background

Wechsler believed that his intelligence scales measure two domains, Verbal IQ and Performance IQ. However, evidence began to accrue after the release of the Wechsler-Bellevue that the scales could be broken down even further. For example, Balinsky (1941) performed the first factor analysis on the scale and suggested that there might be three distinct factors. In the 1950s, factor analytic studies reaffirmed this notion, demonstrating that there was at least one additional domain of functioning that was distinct from Verbal and Performance domains (see Cohen, 1952a, 1952b, 1957a, 1957b). This work showed that a third, small, yet discrete factor seemed related to the Digit Span, Arithmetic, and possibly, the Digit Symbol subtests. Though it had been given different names, Jacob Cohen’s label, “Freedom from Distractibility” (Cohen, 1952a, 1952b), became the dominant label. This was due, in part, to the use of this label by Alan Kaufman in his initial interpretive book, Intelligent Testing with the WISC-R (Kaufman, 1979) and later by the inclusion of this label in the WISC-III factor-index scores (Wechsler, 1991).

Most important is that examiners began using such factor scores in clinical settings.

The developers of the WISC-III sought to enhance the measurement of this additional factor and developed a new subtest, Symbol Search (Prifitera, Weiss, & Saklofske, 1998; Wechsler, 1991). Surprisingly, they found that this new subtest was more related to Coding, not Arithmetic or Digit Span. In factor-analytic studies, they found that four factors seemed to emerge from the analyses and they labeled the new domains of functioning Freedom From Distractibility and Processing Speed. The interpretation of a four-factor solution is not without controversy (see Sattler, 1992, for criticism of the four-factor model). Nevertheless, the four-factor model has been replicated in additional studies (Roid & Worrall, 1996; Blaha & Wallbrown, 1996; Donders, 1997). Furthermore, the additional factors of Processing Speed and Attention seem clinically relevant and are psychologically meaningful (Prifitera & Dersh, 1993; Kaufman, 1994).

Naming the Third Factor

Following the work of Cohen (1952a, 1952b) and Kaufman (1979) the third factor continued to be called Freedom From Distractibility in the WISC-III manual, and normative information was provided for this factor (Wechsler, 1991; Roid et al., 1993). Cohen’s original term Freedom From Distractibility had become entrenched in the psychological community.

In the revision of his classic text, Kaufman (1994, p. 212) criticized this name and wrote that it was a mistake not to “split with tradition” and change the label of this factor years ago. He also pointed out that this factor should have been called “by a proper cognitive name” when The Psychological Corporation published the WISC-III. His criticism appears valid; years of research indicate that the 3rd factor is more than distractibility alone (Wielkiewicz, 1990). Also, this label may lead to improper interpretation of this factor as diagnostic for attention problems, which is not necessarily the case.

Several other WISC-R researchers have echoed this concern. Some have directly stated that labeling this 3rd factor as Freedom From Distractibility is an oversimplification (Stewart & Moely, 1983; Owenby & Matthews, 1985). In a review paper, Wielkiewicz (1990) concluded that low scores on this factor of the WISC-R are not diagnostic of any
single childhood disorder and he argues against this traditional label in favor of either short-term or working memory. Other researchers, while not directly critical of the label, have demonstrated that the Digit Span and Arithmetic subtests of the WAIS-R are related to the Attention and Concentration subtests of the WMS (Larrabee, Kane, & Schuck, 1983) or to other independent measures of attention (Sherman, Strauss, Spellacy, & Hunter, 1996).

These studies and reviews seem to suggest that the subtests that make up this factor (e.g., Arithmetic, Digit Span, and sometimes, Digit Symbol) make up a higher-order cognitive ability, such as working memory. The name Freedom From Distractibility is really a misnomer that implies that this 3rd factor is nothing more than a WISC-III or WAIS-R validity measure, used solely to test a hypothesis that an obtained score underestimates an individual’s “true” score. It may also imply that it is a direct measure of attention disorders, which is an oversimplification.

Attention, an alternate label, is made up of several high-level functions like focusing, encoding, sustaining, and shifting (Mirsky, 1989; Mirsky, Anthony, Duncan, Ahearn, & Kellam, 1991), and selective attention is an even more complicated system that involves the selection of some stimuli for higher levels of processing as well as the inhibition of other signals for those high levels of processing (Posner, 1988).

Working Memory, still another label, involves the storage of information, the manipulation of information, and the storage of products. It requires individuals to track multiple tasks while actively processing information (Baddeley, 1986). Digit Span and Arithmetic tasks have been considered tasks involving working memory (Sternberg, 1993; Kyllonen & Christal, 1987). For the WAIS-III, the label Working Memory was adopted because it is conceptualized as a key process in the acquisition of new information.

**Working Memory**

Working memory is a term that denotes a person’s processing capacity. The concept of working memory has replaced (or updated) the concept of short-term memory. Newell and his colleagues coined the term “working memory” and conceptualized it as a “computational workspace” (Newell, 1973; Newell & Simon, 1972). They viewed this “workspace” as being a “more active part of the human processing system” as opposed to the traditional term, short-term memory, that is the passive storage buffer. Hence, the concepts of working memory and short-term memory are similar because both have been thought of as a place where incoming information is stored temporarily and both are limited in capacity. However, the two concepts differ in one key aspect: short term memory is a “passive” form of memory and working memory is an “active” form. Traditional short-term memory is thought of as a passive storage area for information while it either becomes encoded into long term-memory or is forgotten. Working memory, on the other hand, is an area where incoming information is stored temporarily. It is also the place where calculations and transformation-processing occurs. Furthermore, as Baddeley and Hitch (1974) point out, this component also stores the products or output of these calculations and transformations (as well as the original information).

For the WAIS-III, the definition advanced by Kyllonen & Christal (1987) was employed. Working memory can be defined as the portion of memory that is in a highly active and accessible state whenever information is being processed. This includes the memory that is involved when an individual is simply attending to information (Kyllonen & Christal, 1987).

Recent literature has suggested that working memory is a key component to learning (Kyllonen, 1987; Kyllonen & Christal, 1989; Kyllonen & Christal, 1990; Woltz, 1988). Individuals with greater working memory will be capable of processing and encoding more material than individuals with a smaller working-memory capacity, thus accounting for individual differences in attention and learning capacities. Some cognitive psychologists have come to believe that working memory is an important predictor of individual differences in learning, ability, and fluid reasoning (Sternberg, 1993; Kyllonen & Christal, 1989).

The measurement of working memory dates back to early experiments conducted by Baddeley and Hitch (1974). Traditionally, this construct has been measured by presenting a large amount of information (which the person has to retain in memory), requiring the person to first process (or transform) this information and then to retain the end product. Tasks tend to increase in complexity (e.g., the system is more likely to become “overloaded”) as the test progresses. Individual differ-
 Factor Analysis of the WAIS-III

To determine the factor structure of the WAIS-III, several exploratory analyses were conducted in different ways, using different data sets, using subsets of the data, using different sets of variables, and using different extraction techniques and rotational techniques. Overall, the primary factor loading for each subtest remained relatively consistent from analyses to analyses. A few of these key analyses are reported in the WAIS-III—WMS-III Technical Manual (The Psychological Corporation, 1997) and one of the analyses is reprinted in Table 5.4. Fairly consistently, the Vocabulary, Similarities, Information, and Comprehension subtests all had their highest loading on one factor (called the Verbal Comprehension Index); the Picture Completion, Block Design, and Matrix Reasoning subtests had the highest loading on a different factor (called the Perceptual Organization Index); the Arithmetic, Digit Span, and Letter-Number Sequencing subtests had the highest loading on a third factor (called the Working Memory Index); and Digit Symbol-Coding and Symbol Search had the highest loading on a fourth factor (called the Processing Speed Index).

To test the stability of these results and the appropriateness of this factor structure in different ethnic groups, the exploratory analyses were conducted separately by ethnic group (African-American, Hispanic, and white) using the standardization sample. The sample sizes for the three analyses were: African-American examinees, N = 279; Hispanic examinees, N = 181; white examinees, N = 1925. The exploratory factor-pattern loadings are listed in Table 5.5 for the African-American, Table 5.6 for Hispanic, and Table 5.7 for white examinees. For all three groups, the results are similar to those presented in the WAIS-III—WMS-III Technical Manual. Although there are a couple of variables with split loadings (e.g., Arithmetic is split between Verbal Comprehension and Working Memory for the group of African-American examinees and between Verbal Comprehension and Perceptual Organization for the group of Hispanic examinees), the patterns are extremely similar between these groups.
Table 5.5. WAIS-III Exploratory Factor-Pattern Loadings for Four-Factor Solutions, African-American Examinees

<table>
<thead>
<tr>
<th></th>
<th>VERBAL COMPREHENSION</th>
<th>PERCEPTUAL ORGANIZATION</th>
<th>WORKING MEMORY</th>
<th>PROCESSING SPEED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary</td>
<td>0.85</td>
<td>-0.01</td>
<td>0.03</td>
<td>0.09</td>
</tr>
<tr>
<td>Similarities</td>
<td>0.75</td>
<td>0.05</td>
<td>-0.03</td>
<td>0.11</td>
</tr>
<tr>
<td>Information</td>
<td>0.82</td>
<td>-0.01</td>
<td>0.09</td>
<td>-0.07</td>
</tr>
<tr>
<td>Comprehension</td>
<td>0.77</td>
<td>0.09</td>
<td>-0.06</td>
<td>-0.01</td>
</tr>
<tr>
<td>Picture Completion</td>
<td>0.02</td>
<td>0.57</td>
<td>-0.09</td>
<td>0.13</td>
</tr>
<tr>
<td>Block Design</td>
<td>0.05</td>
<td>0.52</td>
<td>0.15</td>
<td>0.12</td>
</tr>
<tr>
<td>Matrix Reasoning</td>
<td>-0.02</td>
<td>0.56</td>
<td>0.34</td>
<td>-0.03</td>
</tr>
<tr>
<td>Picture Arrangement</td>
<td>0.22</td>
<td>0.51</td>
<td>0.00</td>
<td>-0.02</td>
</tr>
<tr>
<td>Arithmetic</td>
<td>0.38</td>
<td>-0.02</td>
<td>0.51</td>
<td>0.09</td>
</tr>
<tr>
<td>Digit Span</td>
<td>0.05</td>
<td>0.06</td>
<td>0.67</td>
<td>0.00</td>
</tr>
<tr>
<td>Letter–Number Sequencing</td>
<td>-0.02</td>
<td>0.07</td>
<td>0.68</td>
<td>0.22</td>
</tr>
<tr>
<td>Digit Symbol–Coding</td>
<td>0.05</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.76</td>
</tr>
<tr>
<td>Symbol Search</td>
<td>0.01</td>
<td>0.12</td>
<td>0.13</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Note: Data and Table Copyright 1998 by The Psychological Corporation. All rights reserved.

aN = 279

Table 5.6. WAIS-III Exploratory Factor-Pattern Loadings for Four-Factor Solutions, Hispanic Examinees

<table>
<thead>
<tr>
<th></th>
<th>VERBAL COMPREHENSION</th>
<th>PERCEPTUAL ORGANIZATION</th>
<th>WORKING MEMORY</th>
<th>PROCESSING SPEED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary</td>
<td>0.80</td>
<td>-0.15</td>
<td>0.26</td>
<td>0.08</td>
</tr>
<tr>
<td>Similarities</td>
<td>0.73</td>
<td>0.04</td>
<td>-0.01</td>
<td>0.07</td>
</tr>
<tr>
<td>Information</td>
<td>0.76</td>
<td>0.11</td>
<td>-0.07</td>
<td>-0.01</td>
</tr>
<tr>
<td>Comprehension</td>
<td>0.70</td>
<td>0.14</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>Picture Completion</td>
<td>0.17</td>
<td>0.39</td>
<td>-0.01</td>
<td>0.10</td>
</tr>
<tr>
<td>Block Design</td>
<td>-0.04</td>
<td>0.72</td>
<td>0.11</td>
<td>0.01</td>
</tr>
<tr>
<td>Matrix Reasoning</td>
<td>0.14</td>
<td>0.54</td>
<td>0.05</td>
<td>0.14</td>
</tr>
<tr>
<td>Picture Arrangement</td>
<td>0.36</td>
<td>0.37</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>Arithmetic</td>
<td>0.25</td>
<td>0.31</td>
<td>0.32</td>
<td>0.10</td>
</tr>
<tr>
<td>Digit Span</td>
<td>-0.09</td>
<td>0.11</td>
<td>0.67</td>
<td>0.10</td>
</tr>
<tr>
<td>Letter–Number Sequencing</td>
<td>0.18</td>
<td>0.02</td>
<td>0.55</td>
<td>-0.04</td>
</tr>
<tr>
<td>Digit Symbol–Coding</td>
<td>0.01</td>
<td>-0.01</td>
<td>-0.07</td>
<td>0.81</td>
</tr>
<tr>
<td>Symbol Search</td>
<td>0.00</td>
<td>0.03</td>
<td>0.14</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Note: Data and Table Copyright 1998 by The Psychological Corporation. All rights reserved.

aN = 181

Determining the Number of Subtests to Include on Each Index

Only three subtests were included in the Verbal Comprehension and Perceptual Organization Indexes, leaving the Comprehension and Picture Arrangement subtests as supplemental to the Index scores. The decision to exclude the Comprehension subtest on the Verbal Comprehension Index and the Picture Arrangement subtest on the Perceptual Organization Index was based on practical and empirical considerations. Practically, by not including these two subtests, the examiner can save a significant amount of time. Both Picture Arrangement and Comprehension can be lengthy subtests to administer and if reliable data can be obtained from fewer subtests, the clinician can save time and possibly administer other tests to answer specific clinical hypotheses. Also, by including a maximum of three subtests on each index, the four indexes are more balanced and equally weighted. Furthermore, Picture Arrangement tends to have split loadings between the Verbal and Performance factors, so it is a less “pure” task of perceptual organization than the other scales.

Empirical evidence indicated that there was some redundancy between the subtests and that the overall VCI and POI index scores typically do not
Table 5.7. WAIS-III Exploratory Factor-Pattern Loadings for Four-Factor Solutions, White Examinees

<table>
<thead>
<tr>
<th>Subtest</th>
<th>Verbal Comprehension</th>
<th>Perceptual Organization</th>
<th>Working Memory</th>
<th>Processing Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary</td>
<td>0.92</td>
<td>-0.11</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>Similarities</td>
<td>0.74</td>
<td>0.09</td>
<td>-0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>Information</td>
<td>0.81</td>
<td>0.02</td>
<td>0.07</td>
<td>-0.03</td>
</tr>
<tr>
<td>Comprehension</td>
<td>0.80</td>
<td>0.06</td>
<td>-0.01</td>
<td>-0.02</td>
</tr>
<tr>
<td>Picture Completion</td>
<td>0.12</td>
<td>0.49</td>
<td>-0.10</td>
<td>0.21</td>
</tr>
<tr>
<td>Block Design</td>
<td>-0.03</td>
<td>0.68</td>
<td>0.08</td>
<td>0.07</td>
</tr>
<tr>
<td>Matrix Reasoning</td>
<td>0.07</td>
<td>0.60</td>
<td>0.22</td>
<td>-0.06</td>
</tr>
<tr>
<td>Picture Arrangement</td>
<td>0.27</td>
<td>0.41</td>
<td>-0.03</td>
<td>0.09</td>
</tr>
<tr>
<td>Arithmetic</td>
<td>0.22</td>
<td>0.20</td>
<td>0.47</td>
<td>-0.01</td>
</tr>
<tr>
<td>Digit Span</td>
<td>0.02</td>
<td>0.01</td>
<td>0.66</td>
<td>0.04</td>
</tr>
<tr>
<td>Letter-Number Sequencing</td>
<td>0.01</td>
<td>0.01</td>
<td>0.63</td>
<td>0.10</td>
</tr>
<tr>
<td>Digit Symbol-Coding</td>
<td>0.04</td>
<td>-0.05</td>
<td>0.07</td>
<td>0.70</td>
</tr>
<tr>
<td>Symbol Search</td>
<td>-0.02</td>
<td>0.15</td>
<td>0.06</td>
<td>0.69</td>
</tr>
</tbody>
</table>

Note: Data and Table Copyright 1998 by The Psychological Corporation. All rights reserved.

Table 5.8. R-Squared of Different Combinations of Verbal Comprehension Subtests

<table>
<thead>
<tr>
<th>Number of Independent Variables</th>
<th>R-Squared</th>
<th>Subtests</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.84 VOC</td>
<td>COM</td>
</tr>
<tr>
<td>1</td>
<td>0.79 INF</td>
<td>SIM</td>
</tr>
<tr>
<td>1</td>
<td>0.78 SIM</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.78 COM</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.93 VOC</td>
<td>SIM</td>
</tr>
<tr>
<td>2</td>
<td>0.93 VOC</td>
<td>INF</td>
</tr>
<tr>
<td>2</td>
<td>0.92 SIM</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.92 INF</td>
<td>COM</td>
</tr>
<tr>
<td>2</td>
<td>0.92 SIM</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.98 SIM</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.97 SIM</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.97 SIM</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.00 VOC</td>
<td></td>
</tr>
</tbody>
</table>

Note: VOC = Vocabulary; INF = Information; SIM = Similarities; COM = Comprehension.

Data and Table Copyright 1998 by The Psychological Corporation. All rights reserved.

Differ if either three or four subtests are included. To determine how much incremental validity is lost by omitting a subtest, a procedure similar to that employed by Glenn Smith and his colleagues at the Mayo Clinic as part of the Mayo Older Adult Normative Study (Smith et al., 1994) was used in developing the WAIS-III.

The first step was to obtain a sum of scaled scores for each examinee in the WAIS-III standardization sample on all of the subtests that loaded on the Verbal Comprehension Index (VCI) and another sum of scaled scores for those that loaded on the Perceptual Organization Index (POI). For the VCI, Vocabulary, Similarities, Information, and Comprehension were summed. For the POI, Picture Completion, Block Design, Matrix Reasoning, and Picture Arrangement were summed. Then, in a series of separate regression analyses, these two total scores were “predicted” by using their part scores. For example, the Verbal
112 HANDBOOK OF PSYCHOLOGICAL ASSESSMENT

Table 5.9. R-square of Different Combinations of Performance Subtests

<table>
<thead>
<tr>
<th>NUMBER OF INDEPENDENT</th>
<th>R-SQUARED</th>
<th>SUBTESTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>0.67 BD</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.66 MR</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.61 PA</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.61 PC</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.86 BD  PA</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.85 MR  PC</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.84 MR  PA</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.84 BD  PC</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.83 BD  MR</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.82 PC  PA</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.95 MR  PC PA</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.95 BD  PC PA</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.94 BD  MR PA</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.93 BD  MR PC</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1.00 BD  MR PC PA</td>
</tr>
</tbody>
</table>

Note: BD = Block Design; MR = Matrix Reasoning; PA = Picture Arrangement; PC = Picture Completion.

Data and Table Copyright 1998 by The Psychological Corporation. All rights reserved.

The next step was to analyze whether these “shortened” indexes would perform roughly the same as the indexes that consisted of all four subtests in a sample of “normally functioning” adults. Again, the 2,450 examinees from the standardization sample were used for these analyses. To test the effect of omitting Comprehension, two sums of scaled scores were obtained: one by summing four subtests (Vocabulary, Similarities, Information, and Comprehension) and the other by summing three subtests (Vocabulary, Similarities, and Information). Both of these sums of scaled scores were transformed to standardized scales with a mean of 100 and a standard deviation of 15.

Significant differences (p < .05) between these two verbal-comprehension sums of scaled scores were obtained, and only 31 of the 2,450 examinees had differences of more than 0.5 SD units. Moreover, only three of the 2,450 examinees had scores on the three-subtest sum of scaled scores that “fell” outside of the 90 percent confidence interval of the index scores that were based on four subtests. The standard error of estimate and the reliability of these two sums of scaled scores were roughly identical.

For the two perceptual-organization sums of scaled scores, similar procedures were performed with similar results. For the perceptual-organization subtests, two sums of scaled scores were created (one by summing Picture Completion, Block Design, Matrix Reasoning, and Picture Arrangement, and the other by omitting Picture Arrangement and summing the three subtests). As before,
these scores were standardized and then transformed to standardized scales with a mean of 100 and a standard deviation of 15.

Differences between these two perceptual sums of scaled scores indicated that 41 of the 2,450 individuals had difference scores of more than 0.67 SD units. In terms of examining how many of these individuals would have fallen outside of the confidence interval of the index score based on four subtests, only 13 of the 2,450 would have had a difference that was significant. As with the verbal sums of scaled scores, there was not a significant change in the standard error of estimate or the reliabilities of these two scores.

These results supported the conclusion that, for the vast majority of examinees, there would not be significant differences between their index scores based on three subtests and index scores based on their longer counterparts. This is not to say that it is not valuable to administer the two additional subtests. Certainly, it is more desirable to obtain the additional information provided by the Comprehension and Picture Arrangement subtests. This is especially true if there were significant and unusually large differences between Comprehension and the other Verbal Comprehension subtests or between Picture Arrangement and the other Perceptual Organization subtests. Data analysis suggests that, for example, in 3.4 percent of the standardization sample, the Comprehension subtest was at least three points lower than the mean of the verbal subtests and in 4 percent of cases, it was at least three points higher than the verbal mean. Similarly, in 6.4 percent of the standardization sample, the Picture Arrangement subtest was at least three points lower than the mean of the performance subtests and in 8.1% of cases, it was at least three points higher than the mean of the performance subtests. So, by keeping these subtests out, one might miss important information about the relative strengths and weaknesses of some individuals.

Nevertheless, the time required to administer these two additional subtests may not justify the additional information obtained in the majority of cases. Hence, it was decided to construct the Index scores the way they were. Strengths and weaknesses on the Comprehension and Picture Arrangement subtests could always be obtained through profile analysis.

Technical Characteristics of the WAIS-III

Changes in Normative Information

The WAIS-III standardization sample contains 2,450 adults, and covers an age range from 16 to 89 years of age. Extending the upper age to 89 years to adjust for the longer life span of the U.S. population is a significant improvement to the scale. The normative sample was stratified on several demographic variables (e.g., sex, ethnicity, educational level, and region of the country) using the newest census data. The WAIS-III sample includes

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Executive</td>
<td></td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.0</td>
<td>1.1</td>
<td>2.3</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>Manager</td>
<td></td>
<td>0.3</td>
<td>7.3</td>
<td>4.1</td>
<td>6.0</td>
<td>6.1</td>
<td>4.0</td>
<td>1.6</td>
<td>3.2</td>
</tr>
<tr>
<td>Supervisor</td>
<td></td>
<td>1.2</td>
<td>2.8</td>
<td>4.9</td>
<td>3.8</td>
<td>2.8</td>
<td>3.4</td>
<td>.3</td>
<td>2.2</td>
</tr>
<tr>
<td>Professional or Tech Specialist</td>
<td></td>
<td>2.9</td>
<td>12.8</td>
<td>21.6</td>
<td>28.8</td>
<td>19.0</td>
<td>11.3</td>
<td>4.8</td>
<td>11.8</td>
</tr>
<tr>
<td>Marketing or sales</td>
<td></td>
<td>4.6</td>
<td>10.6</td>
<td>7.2</td>
<td>8.1</td>
<td>15.6</td>
<td>9.6</td>
<td>3.2</td>
<td>6.8</td>
</tr>
<tr>
<td>Administrative support and clerical specialist</td>
<td></td>
<td>4.4</td>
<td>12.3</td>
<td>13.5</td>
<td>15.2</td>
<td>16.2</td>
<td>8.0</td>
<td>3.8</td>
<td>8.6</td>
</tr>
<tr>
<td>Farming, Forestry, Fishing, &amp; Related</td>
<td></td>
<td>0.3</td>
<td>0.6</td>
<td>0.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0</td>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td>Precision Production, Craft, &amp; Repair</td>
<td></td>
<td>0.9</td>
<td>1.1</td>
<td>6.1</td>
<td>8.2</td>
<td>2.8</td>
<td>4.0</td>
<td>1.4</td>
<td>3.0</td>
</tr>
<tr>
<td>Operator, Fabricator, &amp; Laborer</td>
<td></td>
<td>10.5</td>
<td>18.4</td>
<td>17.1</td>
<td>13.6</td>
<td>11.2</td>
<td>8.5</td>
<td>3.0</td>
<td>10.0</td>
</tr>
<tr>
<td>Homemaker, Retired, Not in Labor Force</td>
<td></td>
<td>74.1</td>
<td>29.1</td>
<td>11.1</td>
<td>4.9</td>
<td>14.6</td>
<td>6.8</td>
<td>5.6</td>
<td>20.5</td>
</tr>
</tbody>
</table>

Total Percentage                                | 100   | 100   | 100   | 100   | 100   | 100   | 100   | 100 |
N                                               | 343   | 179   | 346   | 184   | 179   | 176   | 690   | 2097|

*Note:* Data and Table Copyright 1998 by The Psychological Corporation. All rights reserved.
many more older adults and minority groups than samples that had been collected for previous versions of the Wechsler adult scales. In the WAIS-R, for instance, 216 people were minorities (or “Non-white” as they were labeled in the WAIS-R manual) which roughly reflected the percentage of minorities in the U.S. based upon the 1970 U.S. census report (Wechsler, 1981). The number has become significantly outdated as the population of the United States has changed. Hence, the sample collected for the WAIS-III reflects the changes that have occurred in the U.S. population over the last 25 years.

Another difference between the WAIS-III and the previous editions is the exclusion of occupational status as a demographic-stratification variable. Occupational status has been replaced by educational level, which is highly correlated with occupational level, and may be used as a predictor of socio-economic status. However, some may find the occupational status of the WAIS-III standardization sample of interest; it is not reported in the WAIS-III manual but it is shown in Table 5.10. Occupation levels were grouped into 10 categories.

Table 5.11. Demographic Characteristics of the U.S. Population: Percentages by Age and Occupational Level

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Executive</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.7</td>
<td>0.4</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>Manager</td>
<td>0.5</td>
<td>3.0</td>
<td>7.5</td>
<td>10.0</td>
<td>11.0</td>
<td>7.7</td>
<td>2.1</td>
<td>6.9</td>
</tr>
<tr>
<td>Supervisor</td>
<td>1.1</td>
<td>3.0</td>
<td>5.2</td>
<td>6.1</td>
<td>5.9</td>
<td>3.9</td>
<td>0.7</td>
<td>4.2</td>
</tr>
<tr>
<td>Professional or Tech Specialist</td>
<td>1.7</td>
<td>9.8</td>
<td>18.1</td>
<td>18.9</td>
<td>19.0</td>
<td>11.1</td>
<td>2.0</td>
<td>13.3</td>
</tr>
<tr>
<td>Marketing or sales</td>
<td>10.3</td>
<td>8.4</td>
<td>6.0</td>
<td>5.1</td>
<td>5.4</td>
<td>4.6</td>
<td>1.4</td>
<td>5.4</td>
</tr>
<tr>
<td>Administrative support and clerical specialist</td>
<td>19.0</td>
<td>24.5</td>
<td>20.5</td>
<td>19.0</td>
<td>18.1</td>
<td>14.2</td>
<td>3.2</td>
<td>16.7</td>
</tr>
<tr>
<td>Farming, Forestry, Fishing, &amp; Related</td>
<td>2.1</td>
<td>1.6</td>
<td>1.3</td>
<td>1.0</td>
<td>0.8</td>
<td>0.8</td>
<td>0.2</td>
<td>1.0</td>
</tr>
<tr>
<td>Precision Production, Craft, and Repair</td>
<td>2.0</td>
<td>5.8</td>
<td>7.7</td>
<td>8.1</td>
<td>6.6</td>
<td>4.6</td>
<td>0.6</td>
<td>5.6</td>
</tr>
<tr>
<td>Operator, Fabricator, &amp; Laborer</td>
<td>14.5</td>
<td>21.0</td>
<td>17.8</td>
<td>16.6</td>
<td>14.9</td>
<td>11.5</td>
<td>1.9</td>
<td>14.0</td>
</tr>
<tr>
<td>Homemaker, Retired, Not in Labor Force</td>
<td>48.9</td>
<td>22.9</td>
<td>15.6</td>
<td>14.8</td>
<td>17.5</td>
<td>41.2</td>
<td>87.7</td>
<td>32.7</td>
</tr>
<tr>
<td>Total Percentage</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>N</td>
<td>14,350</td>
<td>17,317</td>
<td>40,486</td>
<td>43,445</td>
<td>32,477</td>
<td>21,146</td>
<td>31,369</td>
<td>200,590</td>
</tr>
</tbody>
</table>

Employment Status of the Civilian Noninstitutional Population, 1996

<table>
<thead>
<tr>
<th>CIVILIAN NONINSTITUTIONAL POPULATION</th>
<th>CIVILIAN LABOR FORCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL PERCENT OF POPULATION</td>
<td>EMPLOYED</td>
</tr>
<tr>
<td>200,590</td>
<td>133,943</td>
</tr>
</tbody>
</table>

Nine of these are the categories that have been suggested by the National Industry-Occupational Matrix of the Bureau of Labor Statistics. The remaining category included people outside of the work force (people who were retired, homemakers, and or not in the labor force). Table 5.11 lists the population figures for occupational level. These percentages were based upon data from the U.S. Department of Labor (1996). In general, the data obtained for the WAIS-III sample reflects the occupational level of the U.S. population.

Additional Normative Information

The WAIS-III provides additional normative information for optional procedures and for special clinical analysis (e.g., profile analyses and subtest scatter, IQ and factor-index discrepancy scores, memory-ability discrepancy scores, and ability-achievement discrepancy scores). To facilitate interpretation of testing results, The WAIS-III not only provides critical values for determining statistical significance of a given discrepancy, but also
the “base rate” for evaluating whether the discrepancy is clinically meaningful. Previously, this type of normative data was only available in journal articles and related literature that were published well after the test was printed. These tables are provided in the WAIS-III manual, which should be convenient for the clinician using the test.

Reliability

The overall split-half internal consistency coefficients are from .94 to .98 for IQ scales, from .88 to .96 for factor indexes, from .70 to .75 for Verbal subtests, and from .70 to .90 for Performance subtests. The test-retest stability was evaluated using a large sample containing 394 cases, and the stability coefficients were provided for four age-bands as well as for the overall sample. The overall stability coefficients are from .91 to .96 for IQ scales, from .85 to .95 for factor-index scales, from .75 to .94 for Verbal subtests, and from .69 to .86 for Performance subtests. Interrater reliability coefficients are also in the .90s-range for the three Verbal subtests (Vocabulary, Similarities, and Comprehension) that require more judgment in scoring. These reliability coefficients are either improved from or equally as good as WAIS-III predecessors.

Correlation with Other Wechsler Intelligence Scales

The WAIS-III is highly correlated and highly consistent with the WISC-III. The WAIS-III and the WISC-III measure similar constructs and produce similar results. The correlation coefficients between the WAIS-III and WISC-III IQ scores are .88, .78, and .88 for VIQ, PIQ, and FSIQ, respectively. The correlations between index scores are also very high, ranging from .74 to .87. The means of the WAIS-III IQ scores were from 0.4 to 0.7 points higher than the corresponding means of the WISC-III IQ scores. The classification consistency is 95 percent or higher when a 95 percent-confidence interval was used. Similarly, the WAIS-III is also highly correlated and consistent with the WAIS-R. The correlation coefficients between the WAIS-III and WAIS-R IQ scores are .94, .86, and .93 for VIQ, PIQ, and FSIQ, respectively. The mean WAIS-III scores are about 1.2, 4.8, and 2.9 points lower than the corresponding WAIS-R VIQ, PIQ, and FSIQ scores, respectively. This validity ensures the meaningful transition and comparison between the WAIS-III and the WISC-III or the WAIS-R.

It is important to point out that high consistency between the WAIS-III and other Wechsler intelligence scales does not mean that the majority of individuals will obtain “identical” scores across two different Wechsler intelligence scales. When examining individuals, the majority of examinees will obtain different scores across two different tests. This may be because of many factors, such as Flynn effect, practice effect, design differences between the two tests, effects of having a restricted floor or ceiling, and other psychometric factors (Bracken, 1988; Kamphaus, 1993; Zhu & Tuls, 1997). Furthermore, there is likely to be an interaction among these factors. Therefore, clinicians should take the confidence intervals into account when comparing the testing results of WAIS-III and other Wechsler intelligence scales. Score discrepancy should be evaluated on the basis of both statistics and clinical meaningfulness. A true score discrepancy should be statistically significant and clinically meaningful (rare) (Matarazzo & Herman, 1985).

The age-overlapping between the WAIS-III and WISC-III makes it possible for test users who work with adolescents and young adults to test their clients with either tests or to compare their performance on the two tests (Sattler, 1992). A commonly asked question is: When assessing a 16-year-old, which test is more appropriate, WISC-III or WAIS-III? The answer to this question is: It depends on the ability of the 16-year-old. Because the WISC-III has better floor than the WAIS-III, its score should be more reliable for individuals with low abilities; on the other hand, because the WAIS-III has better ceiling than the WISC-III, its score is more reliable for individuals with high ability. For individuals with average ability, the testing results should be very stable across the two tests.

Clinical Group Studies

Whenever a test is revised, there is always a question of whether the patterns of scores are consistent with the scores that would have been obtained if the previous version had been used. Often, these studies are conducted by clinicians and researchers, and the results of the studies are available only in professional journals. For the
WAIS-III, more than 600 individuals who have been diagnosed with a variety of clinical conditions participated at the time of the standardization; the WAIS-III–WMS-III Technical Manual reports the results of these small validation studies. These conditions include: Alzheimer’s disease, Huntington’s disease, Parkinson’s disease, traumatic brain injury, temporal lobe epilepsy, chronic alcohol abuse, Korsakoff’s syndrome, schizophrenia, mental retardation, learning disability, attention-deficit hyperactivity disorder, and deaf and hearing impairment.

Similar to many previous clinical studies with its predecessors, these WAIS-III clinical studies demonstrated the clinical utility of the instrument. It is important to note, however, that the clinical studies reported in the technical manual are not conclusive and provide only initial construct validity. They should not be used to provide “normative” information about the typical functioning of individuals with these clinical conditions.

The samples used in these clinical studies may not be representative because there were not “tight” inclusion and exclusion criteria. Most of the data were collected by clinicians who had busy clinical practices, and often data on some of the groups were collected in different clinics, diagnostic centers, or hospitals. Often, the different sites used different diagnostic procedures, criteria, and data collection methods. Moreover, some samples, such as Parkinson’s disease and lobectomy groups, were relatively small, which increased the likelihood of sampling error.

With those cautionary points mentioned, these studies, nonetheless, provide information to the clinician and researcher. As with many of the previous clinical studies using a Wechsler scale in similar clinical groups, the WAIS-III often replicated the clinical findings that have been published in the literature. For instance, in a clinical study, 108 adolescents and adults diagnosed with mental retardation (62 mild and 46 moderate), using DSM-IV and American Association of Mental Retardation (AAMR) criteria, were tested using the WAIS-III. The results showed that the participants exhibited relatively flat-score profiles and that 99 percent of the sample obtained IQ scores 2 to 3 SDs below the mean. These results are very consistent with previous findings by Atkinson (1992), Craft and Kronenberger (1979), and Spruill (1991) for adult participants and by Wechsler (1991) for children. Further analysis showed that roughly 83 percent of the participants in the mild group had IQ scores between 53 and 70, and that 82 percent of the examinees in the moderate group had IQ scores between 45 and 52 (Tulsky & Zhu, 1997). These results suggest that the WAIS-III not only has sensitivity in identifying individuals with cognitive functioning that is 2 SDs below the mean, but the WAIS-III also has specificity in that it can separate individuals who function at mild and moderate levels.

In another study, the WAIS-III was administered to a sample of 30 adolescents and adults diagnosed with Attention-deficit hyperactivity disorder (ADHD) according to clinical interviews, DSM-IV diagnostic criteria, and the Brown Attention-Deficit Disorder Scales (Brown, 1996). The results at the IQ score level suggested that, this sample performed similarly to the standardization sample. The mean FSIQ was at the average range and there was no significant difference between the VIQ and PIQ. When the factor-index scores were evaluated, however, marked results were found. Their mean WMI score is about 8.3 points lower than their mean VCI score, and their mean PSI score is about 7.5 points lower than their mean PSI score. About 30 percent of the sample with ADHD had WMI scores at least 1 SD lower than their VCI scores, whereas 13 percent of the WAIS-III standardization sample obtained such discrepancies. About 26 percent of the ADHD sample had PSI scores at least 1 SD lower than their POI scores whereas, 14 percent of the WAIS-III standardization sample had such discrepancies. For the difference between the higher score of either the VCI or POI and the lower score of either the WMI or PSI, 61.3 percent of the sample obtained differences of 1 SD, and 16.1 percent obtained differences of 2 SDs or more; only 30.5 percent and 3.5 percent of the WAIS-III standardization sample had such differences for the VCI- or POI-score differences, and the WMIs or PSI-score differences, respectively. These results are comparable to the findings by Brown (1996) using a larger adolescent and adult ADHD sample, and the WAIS-R. They are also very consistent with the findings by Wechsler (1991), Prifitera and Dersh (1992), and the research group lead by Biederman et al. (1993).

The study using the traumatic brain injury (TBI) sample further demonstrated the clinical utilities of the new factor-index scores. The WAIS-III was administered to 22 adults who had experienced a moderate-to-severe single-closed head injury. Consistent with the previous findings, the TBI sample exhibited some overall impairment. Their
IQ scores were all at the low-average-range and no significant differences were found between the mean VIQ and PIQ scores. When the factor scores were compared, however, the relative strengths and weaknesses were obvious. The mean PSI score (73.4) of the TBI sample was significantly lower than the POI scores (92.1) and other factor-index scores. Further analysis showed that about 77 percent of the traumatic-brain-injury sample had a PSI score that was at least 1 SD lower than their POI score, while it was only 14 percent for the standardization sample.

Although it is apparent that evaluating the factor-index scores alone is usually not conclusive for clinical diagnosis, the factor scores certainly can provide extra information that will facilitate the diagnostic processes. Understanding the strengths and weaknesses can also assist in the interpretive process and intervention planning.

Interpretive Considerations

Included in the WAIS-III—WMS-III Technical Manual is a chapter devoted to basic issues in interpreting WAIS-III scores. This chapter provides descriptions of IQ and Factor Indexes, suggestions for basic interpretive consideration, and procedures for discrepancy analysis. The suggestions for interpretive considerations should not be used as a “cook book” or comprehensive guideline for interpretation. Clinical interpretation is a very complicated hypothesis-testing process that varies from situation to situation. Therefore, no single approach will work for all scenarios.

Since the WAIS-III continues the tradition of the Wechsler intelligence scales, many interpretation strategies, methods, and procedures that were developed by experienced clinicians and researchers for its predecessors should still be valid and useful for interpreting its results. Test users should refer to Kaufman (1990, 1994) and Sattler (1992) for detailed introductions and discussions of these interpretation strategies, methods, and procedures. Additionally, in response to progress in the field of cognitive assessment, the WAIS-III provides new factor-index scores that measure more refined cognitive domains, and these factor indexes have proven useful and informative in clinical diagnosis (Tulsky, Zhu, & Vasquez, 1998). Clinicians should evaluate the additional information provided by these factor indexes when interpreting the traditional IQ scores.

While detailed discussion of interpretation strategies, methods, and procedures is beyond the scope of this chapter, the authors would like to suggest a few basic interpretive considerations that may help readers understand the nature of clinical interpretation with the WAIS-III.

First, testing results should never be interpreted in isolation. Instead, interpretation must be made within the context of an individual’s current mental status, social environment, and life history. As suggested by the WAIS-III—WMS-III Technical Manual, when interpreting the WAIS-III results, clinicians should consider four broad sources of information: medical and psychological history, direct behavioral observations, quantitative test scores, and qualitative aspects of test performance.

Second, testing is different from assessment (Matarazzo, 1990; Prifitera, Weiss, & Saklofske, 1998; Robertson & Woody, 1997). Psychological testing is a data-collection process in which an individual’s behaviors are sampled and observed systematically under standardized conditions. Psychological assessment is a complicated problem-solving process that usually begins with psychological testing. Therefore, obtaining some test scores is just the beginning of assessment, not the end.

Third, interpretation is the process of “making sense” out of the test results. It includes a very complicated, multi-level process where hypotheses are systematically formed and tested using test scores and other clinical information. Interpretation integrates data collected through testing (such as quantitative test scores, qualitative aspects of test performance, and direct behavioral observations) with information about a person’s medical and psychosocial history and weaves them together into meaningful information. Each test score may be used as a piece of evidence supporting certain conclusions. Each piece of information is like a puzzle piece. Clinicians must first gather all puzzle pieces and then put all of them together in a meaningful way before any conclusions can be made. With this analogy in mind, it will be clear that even though identifying one puzzle piece is usually not sufficient to solve the whole puzzle, it is a necessary and important step. It is similar to the physician who measures a patient’s body temperature and blood pressure as just two steps along the way in reaching a diagnosis. Temperature and blood pressure are universally-performed procedures, however, neither of them, in isolation, are conclusive for a final diagnosis. Similarly, scores
on an intelligence test must be combined with scores on other tests, the examinee’s demographic information, such as socioeconomic status, life history, educational background, and other extratest information before any clinical decision can be made.

**Basic Interpretation of the WAIS-III**

**Wechsler Scores**

The WAIS-III uses a scoring metric that will be familiar to users of other Wechsler tests. Subtest raw-scores are transformed to subtest scaled-scores with a mean of 10 and a standard deviation of 3. A subtest scaled-score of 10 indicates that the individual is performing at the average level of a given group. Scores of 7 and 13 would reflect performance that is 1 SD below and above the mean, respectively, while scaled scores of 4 and 16 would reflect performance that is 2 SDs from the mean.

The WAIS-III differs from its predecessors in that the scaled scores are now age-corrected. On the WAIS-R, a reference group of subjects (ages 20–34 years) is used to convert raw scores to scaled scores. By doing this, the subtest scores are compared to the level of performance of a relatively young reference group. Since some of the skills measured by the Wechsler adult scales decline with age, these subtest scores will reflect this decline when examinees are compared with a normative group much younger than themselves. In previous editions, the composite scores were the unit that was adjusted for age (e.g., Verbal, Performance, and Full Scale IQ scores are computed separately by age group).

In the WAIS-III, the correction for age was made at the initial transformation to scaled subtest scores. This change was made in order to prevent older subjects from receiving very low scaled-scores on some (most) of the subtests because they are being compared with examinees their own age rather than examinees who are much younger. As an example of how profound this decline can be, an example of converting raw scores to scaled scores for an 85-year-old is presented in Table 5.12. In the first line of the table, the subtest raw-scores are presented. The raw scores for this example were selected so that they corresponded with an average performance (e.g., scaled score of 10) of an 85- to 89-year-old examinee, and would emphasize the point of how different scores could look. The second row in the Table 5.12 shows the age-corrected scaled scores (e.g., 10) that correspond to the raw-score points. The “reference” group’s scaled scores are presented in the third row. As shown, the perceptual subtests (e.g., Matrix Reasoning or Picture Completion) and processing-speed subtests (e.g., Symbol Search or Digit Symbol) show significantly lower scaled-scores when the individual is compared to a younger-aged reference group rather than compared 85-year-olds. In the WAIS-III, a reference group comparison (at the subtest level) can still be made, but this is now an optional procedure and these reference-based scaled scores do not feed into the formula to calculate the IQ score.

Instead, it is the age-corrected scaled scores that are summed and transformed to yield composite scores. The WAIS-III IQ and Index scores have retained the common metric of a mean of 100 and a standard deviation of 15 for evaluating level of performance. A score of 100 on any of these measures defines the average performance of individuals within the same age group. Scores of 85 and 115 correspond to 1 SD below and above the mean, respectively, whereas scores of 70 and 130 are 2 SDs below and above the mean. About 68 percent of all examinees obtain scores between 85 and 115, about 95 percent score in the 70–130 range, and
nearly all examinees obtain scores between 55 and 145 (3 SDs on either side of the mean).

Scores should be reported in terms of confidence intervals so that the actual score is evaluated in light of the reliability of that test score. Confidence intervals assist the examiner in test interpretation by delineating a range of scores in which the examinee’s “true” score most likely falls, and remind the examiner that the observed score contains measurement error.

**Level of Performance**

The level of performance refers to the rank that is obtained by an individual in comparison to the performance by an appropriate normative group. Clinical decisions can then be made if the level of performance of the individual is significantly lower than the normative group. Alternatively to this normative approach, clinical decisions can also be made if a specific score is lower than the individual’s other scores (relative weaknesses). In nonclinical settings (e.g., industrial and occupational settings), the emphasis on level of performance shifts slightly, as more weight is placed on competency and the patterns of a person’s strengths and weaknesses without necessarily implying any type of impairment. As described in the *WAIS-III–WMS-III Technical Manual*, test results can be described in a manner similar to the following example:

Relative to individuals of comparable age [or, alternatively, of a reference group of younger adults], this individual is currently functioning in the [_____] range of functioning on a standardized measure of [IQ or Index name]. (p. 185)

IQ and index scores are estimates of overall functioning in an area that should always be evaluated. As composite scores, they should be interpreted within the context of the subtests that contribute to the overall IQ scale or index score. The IQ and index scores are much more reliable measures than the subtest scores, and, in general, these are the first scores to examine when one begins to review WAIS-III data. Sometimes, the VIQ or PIQ scores and the various index scores are discrepant from one another, indicating that the examinee has some areas of functioning that are stronger or weaker than other areas of functioning.

Alternatively, sometimes the subtests that make up the IQ and index scores are substantially different from one another. It is important to realize that when two component subtest-scores are substantially different from one another, with one unusually high and the other unusually low, it will push the index score toward the arithmetic mean and thus toward the average range. Such an average score reflects a dramatically different pattern of abilities than does an average index score obtained from two subtest scores that are both in the average range. It is common practice for examiners to closely examine profiles in an ipsative fashion (e.g., examine the subtests against the examinee’s own anchor point rather than against the subtests of a norm-referenced group) to see which scores show relative strengths and which show relative weaknesses. This technique is called profile analysis.

**Profile Analysis and Cluster Interpretation**

In clinical practice, clinicians compare the examinee’s performance on the 11 (WAIS-R) or 13 (WISC-III) subtests to see if any “patterns” emerge from which they can make inferences about an examinee. Glutting, McDermott, & Konold (1997), reported that there are more than 75 different patterns of subtest variation. Some have suggested various ipsative analyses (see Kaufman, 1994; Sattler, 1992) while others have stressed using a normative approach (McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992; McDermott, Fantuzzo, & Glutting; 1990) to analyze patterns of scores. McDermott et al. (1992) have even used cluster-analytic techniques to develop different subtest taxonomies as an alternative to profile analyses (McDermott, Glutting, Jones, Watkins, & Kush, 1989; McDermott, Fantuzzo, & Glutting; 1990) to analyze patterns of scores. McDermott et al. (1992) have even used cluster-analytic techniques to develop different subtest taxonomies as an alternative to profile analyses (McDermott, Glutting, Jones, Watkins, & Kush, 1989; McDermott, Glutting, Jones, & Noonan 1989). Unfortunately, by taking a strictly normative approach, the examiner may miss some information about an individual’s strengths and weaknesses. In fact, it is very common for an individual to function at different ability levels in different cognitive areas. By examining deviations from the individual’s average level of functioning (e.g., significant and unusual differences between subtests and the average of all subtests), the examiner may see a pattern and generate additional hypotheses. Furthermore, in the WAIS-III, the examiner can use the frequency data of deviations from a mean that were obtained in the WAIS-III standardization study to decide how rare the obtained difference is in a nor-
mative sample and interpret such normative information within the context of other facts and life history data that the examiner has accumulated about the individual.

**IQ-Score and Index-Score Discrepancies**

In Wechsler’s (1939) initial work on his first intelligence scale, he placed most of the emphasis on the FSIQ score and believed that an examinee’s FSIQ score is always an average of the person’s performance on all of the subtests (Wechsler, 1944). Nevertheless, Wechsler did realize that it was still important, at times, to view the VIQ and PIQ scores separately. He thought that this procedure would usually be reserved for the occasion of testing a person with special disabilities (Wechsler, 1944).

Since the publication of the Wechsler-Bellevue scale, the practice of interpreting VIQ-PIQ differences has become a common method of determining when to modify the interpretation of an FSIQ score and to examine the VIQ and PIQ scores separately. In the WAIS-R, Wechsler (1981) included a table to show the minimum differences between the VIQ and PIQ scores required for significance at the .15 and .05 levels of confidence for each age group. “Rules of thumb” abounded, and generally, a difference score of 12–15 points became the marker at which examiners started inferring that the examinee had a clinically relevant deficit. Matarazzo and Herman (1985) documented that the frequency of 12- to 15-point differences were much more common than examiners had previously believed and they demonstrated the need for examining statistical significance as well as clinical meaningfulness (base rates). In other words, they differentiated between a statistically significant difference (which suggests that the examinee is better at one skill than another) and a clinically meaningful difference (which indicates that the obtained-difference score is of such a high magnitude that it does not occur very frequently). This latter finding may suggest that the examinee has a true clinical deficit in an area, however, this can only be concluded after the examiner has reviewed the other variables (e.g., other test scores, psychosocial history, educational level) and found results to support such an interpretation. In addition to the VIQ- and PIQ-score differences, the WAIS-III includes normative discrepancy information on all possible pairs of index scores. A variety of detailed interpretation schemes has been suggested to explain meaningful differences (e.g., Kaufman, 1990, 1994; Sattler, 1992).

The WAIS-III–WMS-III Technical Manual also presents frequency-of-score differences by ability level (The Psychological Corporation, 1997). Unfortunately, there is no presentation of these frequency data by other demographic information (e.g., educational level). These alternative tables may prove to be more useful because variables such as previous level of education would not be affected by a neuropsychological disorder or condition, whereas current overall ability may be lowered by the neuropsychological deficit.

**Interpretation in a Neuropsychological Setting**

Neuropsychology is a highly specialized approach to the understanding of individual differences (Hynd & Semrud-Clikeman, 1990). It is the measurement and analysis of the cognitive, behavioral, and emotional consequences of brain damage or dysfunction (see the neuropsychology section, this volume). Often, the WAIS-III is used to gauge the individual’s current overall ability and will play a part in helping the neuropsychologist detect gross intellectual deterioration. The IQ scores generated by the Wechsler scales of intelligence are typically very sensitive to generalized impairment. However, these same IQ scores are also relatively insensitive to very focalized lesions of the brain (Matarazzo, 1972; Hynd & Semrud-Clikeman, 1990; Chelune, Ferguson, & Moehle, 1986). Instead, other tests that measure more distinct cognitive functions are used to supplement Wechsler IQ scores to detect specific deficits.

Since the emphasis of the evaluation typically focuses on specific abilities, the examiner may place more weight on the measurement of a person’s ability in various functional areas than on an overall IQ score. Various researchers have identified between five and seven major functional areas, including intelligence, language, spatial or perceptual ability, sensorimotor functioning, attention, memory, emotional or adaptive functioning, psychomotor speed, and learning (see Lezak, 1995; Larrabee & Curtiss; 1995; Smith et al., 1992 for a comprehensive review). With the new WAIS-III, the index scores that break down Verbal and Performance IQs into somewhat more specified scores than those obtained by the Verbal and Performance IQ scores, should be an asset to
ASSESSMENT OF ADULT INTELLIGENCE WITH THE WAIS-III

Since, the neuropsychologist is attempting to detect some of the cognitive consequences of brain damage, he or she must: (a) compare an individual's current score to his or her estimated (or known) premorbid level of functioning or use demographically-corrected scores, or both, to factor in the effects attributable to various demographic variables, (b) factor in any effects due to previous testing, and (c) examine test scores to determine strengths and weaknesses between various cognitive and memory functions. The remainder of this chapter will examine these various areas of interpretation.

Predicting Premorbid Functioning. A difficult task faced by any psychologist is determining if an individual's current test scores reflect a drop in performance from the same individual's previous ability before an accident occurred or illness began (Franzen, Burgess, & Smith-Seemiller, 1997). This process can help the neuropsychologist make a determination about whether the individual has sustained loss in functioning from the accident or illness as compared with his or her previous ability. Wechsler was the first person to propose that there was a "deterioration index" that could be derived by comparing the performance on so called "hold subtests" of the Wechsler scales (e.g., those subtests' scores that were found not to decline with the age of the examinee) to the "don't hold subtests" (e.g., those subtests' scores in which performance was not expected to remain stable over time and would ultimately deteriorate with the age of the examinee) (Wechsler, 1944). However, basing the assessment of premorbid function on "hold" tests can underestimate premorbid IQ by as much as a full standard deviation (Larrabee, Largen, & Levin, 1985; Larrabee, 1998).

Alternative methods of determining premorbid functioning utilize the relationship between Wechsler IQ scores and demographic variables such as age, education, sex, race, and occupational level. For a detailed discussion about the relation between demographic variables and IQ scores, see a review by Heaton, Ryan, Grant, and Matthews (1996), and studies by Heaton, Grant, and Matthews (1986), and Kaufman, McLean, and Reynolds (1988). In general, two methodologies have prevailed. Some have used the correlations to develop prediction equations (e.g., Wilson et al., 1978; Barona, Reynolds, & Chastain, 1984), while others have developed independent norms that use more focused reference-groups against which the examiner can compare scores.

Capitalizing on the high correlations between demographic variables and IQ scores, researchers began performing regression analyses on the WAIS (Wilson et al., 1978) and the WAIS-R (Barona, Reynolds, & Chastain, 1984) standardization samples in an effort to develop formulas to calculate premorbid IQ. Wilson et al. (1978) constructed prediction formulas based on five demographic variables (age, sex, race, education, and occupation) for Full Scale, Verbal, and Performance IQs. They also found that education and race were the most powerful predictors in each equation. Once the WAIS-R was published, Barona et al. (1984) replicated this work and constructed equations consisting of the following brain damage do much poorer than the general population on Vocabulary, a finding that contradicted Yates' hypothesis.

The more recent focus has been on using reading tests as an indicator of premorbid functioning (Nelson, 1982; Nelson & McKenna, 1975; Nelson & O'Connell, 1978). Nelson and O'Connell (1978) introduced the National Adult Reading Test (subsequently named the New Adult Reading Test [NART]), which was a reading test using irregularly pronounced words. They developed a regression-based formula for estimating WAIS IQ scores from the scores on the NART reading test and concluded that the predictions based on NART scores are fairly accurate. Subsequent revisions have included an alternative NART for American participants (AMNART) (Grober & Sliwinski, 1991), an alternative revision of the NART for American examinees (NART-R) (Blair & Spreen; 1989), and a reading subtest from the Wide Range Achievement Test-Revised (Kareken, Gur, & Saykin, 1995).
predictor variables: age, sex, race, geographic region of residence, occupation, and education). Sweet, Moberg, and Tovian (1990) reviewed these two prediction formulas and concluded that there was not strong support for using the Barona index over the Wilson index and that, at best, “modest success” in terms of adequate classifications may be achieved by both formulas. The authors conclude that these formulas may be useful in research or when used in conjunction with past records, but they should not be used “in isolation with individual patients” (Sweet et al., 1990; p. 44).

A variant of this approach that should be mentioned is a “combined approach” that would use scores (e.g., Vocabulary, achievement NART scores) as a concurrent measure of ability along with demographic variables to develop a “better” formula to predict premorbid IQ (Krull, Scott, & Sherer, 1995; Vanderploeg & Shinka, 1995). To support this methodology, the advocates of this technique stress that the amount of variance that is accounted for through multiple regression-analyses increases when the regression model includes a concurrent measure of reading and demographic variables as predictors. A word of caution should be offered, however. To the extent that a potential disorder or disability may affect current functioning, the correlation between a concurrent measure and premorbid ability will decrease, and such a methodology may be less accurate than prediction using only demographic information.

Heaton and his colleagues (Heaton, Grant, & Matthews, 1991; Heaton, 1992) proposed an alternate way to interpret IQ scores in light of demographic variables. They conducted a study with 553 neuropsychologically normal adults that investigated the relationship between neuropsychological-test scores and demographic characteristics (Heaton, Grant, & Matthews, 1986). Some scores were highly correlated with age; others were related to other demographic variables like educational level. Moreover, these demographics affected diagnostic accuracy of neuropsychological tests.

As a result, Heaton and his colleagues obtained new normative information on several measures that are commonly used in neuropsychology. They developed and published new normative information for the WAIS (Heaton et al., 1991) and for the WAIS-R (Heaton, 1992), corrected for age, education and sex. The neuropsychologist could then evaluate an individual’s performance and compare how he or she performed relative to a person of similar age, ethnicity, background, gender, and education. This score could be compared and contrasted with the traditional IQ and the examiner would have a pretty good idea of how the average individual coming from a certain culture and age would have performed. In a separate study, Malec, Ivnik, Smith and their colleagues at the Mayo Clinic (Malec et al., 1992) developed age-and education-corrected WAIS-R scores for examinees older than 74 years of age. While this methodology is different than the others proposed above (i.e., it isn’t used to predict premorbid IQ directly), the clinician uses a systematic technique to evaluate an obtained score. By comparing the overall ability score with a demographically corrected score, the examiner can judge if the score seems to reflect a deficit, given the individual’s background and socio-economic status. If so, then there is a greater probability of a neuropsychological deficit.

Throughout the development of the WAIS-III and WMS-III, there was a significant effort to develop techniques to assist the neuropsychologist. Though not included with the publication of the WAIS-III and WMS-III, research studies and development work on two of the techniques described above were included in the research design. First, an additional 437 examinees completed the WAIS-III while the WAIS-III and WMS-III were standardized. This “educational-level oversample” was collected to ensure that a minimum of 30 individuals within four educational levels were tested in each age group with the WAIS-III. This ensured that there were enough examinees at each educational level so that age-by-education levels could be created. Second, a new-word reading test was developed and co-administered with the WAIS-III and WMS-III. It was completed by 1,250 individuals. The test uses words that are phonetically difficult to decode and would probably require previous learning. Similar to the results presented by Vanderploeg and Shimka (1995), regression analyses demonstrate that this reading test adds more incremental validity in predicting IQ and Memory Scores than do equations that just include demographic variables in predicting IQ scores. Moreover, by being co-normed directly with the WAIS-III and WMS-III, the reading test should provide invaluable information to the clinician who is trying to determine premorbid IQ. It is unfortunate that these techniques were not included in the released versions of the tests.
A recent study by Smith-Seemiller, Franzen, Burgess, and Prieto (1997) suggests that such techniques have been slow to integrate into clinical practice. Smith-Seemiller and colleagues conducted a survey using a sample of some of the doctoral-level members of the National Academy of Neuropsychology. They discovered that despite all of the research that is being performed in this area, relatively few neuropsychologists are applying these techniques in their evaluations with patients. Instead, the vast majority of clinicians tend to rely solely on self-report data that is obtained in a clinical interview, and some also utilize an individual’s vocational status to make rough predictions about premorbid functioning. In the *WAIS-III-WMS-III Technical Manual*, it was emphasized that good practice means that all scores should be evaluated in light of someone’s life history, socioeconomic status, and medical and psychosocial history. It is unclear whether these examiners were not practicing in this fashion or whether they simply did not value the actuarial- or regression-based approaches that were surveyed.

**LINKS BETWEEN THE WAIS-III AND OTHER MEASURES**

**Differences Between the WAIS-III and the WMS-III**

Perhaps the most important development in the revision of the WAIS-R was to codevelop and co-norm the WAIS-III and the WMS-III. Because the scales were co-normed, examiners can directly compare IQ and memory differences, which may lead to additional power in detecting when and what type of deficits occur. Discrepancies between intelligence and memory are sometimes used to evaluate memory impairment. With this approach, learning and memory are assumed to be underlying components of general intellectual ability and, as such, to be significantly related to the examinee’s performance on tests of intellectual functioning. In fact, the examinee’s IQ scores are often used as an estimate of the individual’s actual memory ability. Several researchers have advanced the theory that when memory scores are significantly lower than IQ scores, the discrepancy is suggestive of a focal memory impairment (Milner, 1975; Prigatano, 1974; Quadfasel & Pruyser, 1955). This is especially true when the difference between the IQ and memory scores exceeds what one might expect when comparing an individual’s performance to the normative sample. For instance, if the base rate of occurrence of a large IQ-Memory discrepancy is very low, then this discrepancy score would have clinical utility. By overlapping the samples so that everyone who was part of the WMS-III sample was also part of the WAIS-III sample, more accurate base rates of IQ-memory discrepancies may be obtained. The interpretation and treatment of memory deficits is beyond the scope of this chapter and the curious reader should refer to Larrabee (this volume) for more information about memory testing in the neuropsychological evaluation or to Sohlberg, White, Evans, and Mateer (1992) and Mateer, Kerns, and Esco (1996) for a review and presentation of treatment methods.

**Differences Between the WAIS-III and Measures of Achievement**

In educational settings, the IQ scores of the Wechsler intelligence tests had been widely used in the comparison of students’ general ability level and their level of achievement. As observed by Gridley and Roid (1998), the main purpose of comparing ability and achievement is to evaluate the discrepancy between expected and observed achievement. Since the enactment of the Education for All Handicapped Children Act of 1975, the more recent Individuals with Disabilities Act (IDEA), 1990; and the reauthorization of IDEA (1997), the comparison of intellectual ability to academic achievement has become a key step in determining the presence of specific learning disabilities. Nevertheless, there are pros and cons about this methodology (Gridley & Roid, 1998).

To help clinicians analyze the discrepancy between expected and observed achievement, the WISC-III was linked to the Wechsler Individual Achievement Test (WIAT) (The Psychological Corporation, 1992). Using FSIQ as the measure of ability, one can predict what an individual’s achievement scores should be at a given level of IQ. When an individual does not achieve this predicted level (e.g., a lower-than-predicted score on the WIAT) or exceeds the predicted level (e.g., a high score on the WIAT), the examiner should examine the test scores more closely. Critical values required for a given ability-achievement discrepancy score to be significant at the .05 and .15 levels were included in the
More importantly, this technical manual also presents the frequencies of the ability-achievement discrepancy scores obtained by the standardization sample. Flanagan and Alfonso (1993a, 1993b) developed similar tables using VIQ and PIQ as measures of ability. This additional normative information has increased the clinical utility of the WISC-III in educational settings.

The tables reported in the WIAT Manual look at the relationship between the WAIS-R and the WIAT (for 17- to 19-year-olds). For the WAIS-III, a validity study was conducted in order to evaluate the relationship between the WAIS-III and the WIAT. A linking sample of 142 normal adults 16–19 years of age was used. The correlation coefficients are from .53 to .81 between the WAIS-III IQs and the WIAT composite scores, and from .37 to .82 between the WAIS-III IQs and the WIAT subtest scores. These results are comparable to those reported previously (Wechsler, 1991; The Psychological Corporation, 1992). Using similar tables as those reported in the WIAT manual (The Psychological Corporation, 1992), the technical manual reports ability-achievement discrepancy scores for both a simple difference and a regression method. Discrepancy scores are reported as both statistically significant values as well as base rate frequency data. Following the WIAT tradition, the critical values required for a given ability-achievement discrepancy score to be significant at the .05 and .15 levels and the frequencies of the ability-achievement discrepancy scores obtained by the linking sample, were provided in the technical manual for both simple difference and regression methods.

When using the simple-difference method, the examiner subtracts the IQ scores directly from the achievement scores and evaluates significance and meaningfulness in a two-step process. First, the user should determine whether a given ability-achievement discrepancy is statistically significant. If it is, then the examiner should determine how frequently such a discrepancy had occurred in the linking sample.

When using the predicted-achievement method, the steps are a little more complicated. First, the examiner should find the predicted achievement scores using the examinee’s IQ scores as a guide. The second step is to find the discrepancy score by subtracting the observed-achievement score from the predicted-achievement score. Third, the statistical significance of the difference should be decided and if it is statistically significant, then the examiner should determine if the discrepancy is rare by using the base-rate data from the linking sample.

In general, the predicted-achievement method is preferred for the ability-achievement analysis because it takes into account the measurement errors and the relationship between the measures of ability and achievement. Although it is easy to use, the simple-difference method assumes perfect correlation between the measures of ability and achievement and overlooks the measurement errors (Braden & Weiss, 1988).

**NOTES**

1. Object Assembly is still included in the WAIS-III but is considered an optional subtest; (see Wechsler, 1997a; p. 6).
2. Picture Arrangement generally had a split loading between the Perceptual Organization and Verbal Comprehension Indexes.
3. Generally, Arithmetic had a primary loading on this factor. In some of the analyses, however, it had split loadings with the Verbal Comprehension Index. Occasionally, it would have a split loading between the Working Memory and Perceptual Organization Indexes.
4. The mean of the verbal subtests was calculated using the six subtests that contribute to the Verbal IQ score (e.g., Vocabulary, Similarities, Arithmetic, Digit Span, Information, and Comprehension).
5. The mean of the performance subtests was calculated using the five subtests that contribute to the Performance IQ score (e.g., Picture Completion, Digit Symbol-Coding, Block Design, Matrix Reasoning, and Picture Arrangement).

**REFERENCES**


ASSESSMENT OF ADULT INTELLIGENCE WITH THE WAIS-III


Zhu, J., & Tulsky, D. (1997). The consistency and discrepancy between the WISC-III and WAIS-III. In D. Tulsky & J. Zhu (Co-Chairs), Methodological considerations and innovations in the development of the WAIS–III. Symposium conducted at the 105th annual American Psychological Association Convention, Chicago.

CHAPTER 6

GROUP INTELLIGENCE TESTS

Robert W. Motta
Jamie M. Joseph

The terms “intelligence test” and “IQ test” are used synonymously in this chapter. This is done more for reasons of convenience than for accuracy, for it is clear that the term “intelligence” implies a far wider range of abilities and adaptive skills than does a single IQ score. Whether referred to as IQ or intelligence, group tests of intellectual ability are used extensively in the United States and throughout the world. Over the last 30 to 40 years countries such as Belgium, France, the Netherlands, and Norway have regularly tested all young people entering military service; while countries such as Australia, Canada, former East Germany, Great Britain, and New Zealand have regularly conducted large-scale group intelligence testing of school children (Flynn, 1987).

EARLY HISTORICAL PERSPECTIVE

Written examinations of academic capabilities are a fairly recent development in the United States, but use of assessment procedures to evaluate the capabilities of groups of individuals has a long history. Records indicate that as early 2357 B.C. Chinese emperors were employing examinations of military officers. In 1115 BC civil service examinations were first used in China (Aiken, 1976). From 500 BC to 100 A.D. the Greeks employed tests for military proficiency and for college admissions. In 1200 AD the first oral examination for the Ph.D. degree as well as public and private exams for the Master of Law degree were held at the University of Bologna. It was not until the 1860s that United States’ schools and colleges began using written examinations. Modern U.S. college-admissions testing dates back to 1900, when the College Entrance Examination Board was founded as a membership association of colleges and universities (Garber & Austin, 1982).

Interest in the study of individual differences was stimulated by Charles Darwin’s work on the origin of species. Sir Francis Galton, the cousin of Charles Darwin, was interested in the hereditary nature of genius and published Hereditary Genius in 1869. Galton also devised a number of sensory-motor tests and developed procedures for studying individual differences. As an outgrowth of his efforts to measure individual variation, Galton in 1888 described a method of “co-relations,” which is the basis of modern correlational procedures (Aiken, 1976). During this time, experimental psychologists in Germany, including Fechner, Wundt, and Ebbinghaus demonstrated that, as could physical phenomena, psychological events could be quantitatively assessed.

James M. Cattell, who studied in Germany for his Ph.D. degree, became acquainted with Galton’s methods of sensory-motor assessment and attempted to relate them to what he called “mental tests” in the late 1800s. Theorists such as Spearman, Thorndike, Thurstone, Cattell, and Guilford, and more practically oriented psychologists such as Terman, Wechsler, Bayley, and Gishelli made substantial contributions to the assessment of intelligence. Yet, Alfred Binet and Theodore Simon are
DEVELOPMENT OF INDIVIDUAL AND GROUP INTELLIGENCE MEASURES

In the early 20th century Binet and Simon were commissioned by the Minister of Public Instruction in Paris to identify those children who were not able to benefit from regular public education. Binet used 30 school-related questions of increasing difficulty that were reputed to assess the ability to judge, understand, and reason. A later revision of this test in 1908 contained far more items than the original, and these were grouped into age levels from 3 to 13 years. It was in this later revision that the concept of mental age was introduced. Three years later Binet and Simon extended their test to the adult level.

Around 1915, Otis in the United States and Burt in England were experimenting with group intelligence tests for children. Not only was it seen as economical to assess children in large groups, but group tests could be administered by teachers who did not require the extensive training needed to administer Binet testing. These group tests tapped many of the processes of the individual tests, including the comprehension of relations (e.g., analogies), classification, vocabulary, problem solving, common knowledge, and so on (Vernon, 1978). The advent of World War I in 1917 served as a powerful stimulus for wide-scale group intelligence testing of young adults in the United States. The work of Arthur Otis, Lewis Terman, and Robert Yerkes, who was then president of the American Psychological Association, resulted in the development of the Army Alpha Intelligence Test. This test was used to screen large numbers of recruits for WWI so that they could be placed in service positions for which they were most suited. The Army Alpha, designed for literate recruits, and the Army Beta, for the less literate, were used in the screening of 1,726,966 men in 35 camps. Testing of military recruits by means of various group-administered assessment devices continues to this day.

Assessment of young men for military service and placement of these individuals in positions for which they were suited eventually led to civilian use of group tests of intelligence and ability. Following World War I several variations of the Army Alpha were used in hundreds of schools to assess academic capabilities. These group-administered, objectively scored tests were viewed by many as being superior to standard methods of teacher evaluation and grades. By 1923 use of group-intelligence testing devices had expanded to the point where 37 different group tests of intelligence were identified (Pintner, 1923). Thus, by the early part of the 1900s, group intelligence tests had firmly taken hold and established their utility in the identification of those who could benefit from academic instruction of various kinds.

DEFINITIONS OF INTELLIGENCE

Whether measured by group-administered or individually administered tests, the definition of intelligence has varied considerably over time. The fact that these tests correlate with academic performance does not help in clarifying what intelligence is. This ambiguity has led many to take the position that "intelligence is what intelligence tests measure." An overview of these definitions (Sattler, 1988) is presented below; however, space limitations prevent little more than a brief mentioning of them. What should be noted in the material that follows is the wide array of definitions and the equally diverse theoretical models of intelligence.

Binet and Simon (1916) focused on a set of qualities such as judgment, common sense, initiative, and adaptation; while Wechsler (1958) stressed that intelligence implies purpose, rationality, and ability to deal effectively with the environment. Factor-analytic and statistical theories of intelligences such as those of Spearman (1923) and Vernon (1950), proposed a general theory of intelligence, whereas others like Thorndike (1927) and Thurstone (1938) viewed intelligence as being composed of many independent faculties. Thurstone enumerated at least eight primary mental factors, including verbal, perceptual-speed, inductive-reasoning, number, rote-memory, deductive-reasoning, word-fluency, and space or visualization. Thorndike described three kinds of intelligence: social, concrete, and abstract. Spearman (1923) proposed a two-factor theory which emphasized a general factor (g) and one or more specific factors (s). The concept of a “g” factor has had a tremendous impact on early and current conceptualizations of intelligence.

Guilford (1967) proposed that three classes of variables must be considered when attempting to
define intelligence, and these include: the activities or operations performed (operations), the material or content on which the operations are performed (content), and the product that is the result of the operations (products). Vernon (1950) put forth a hierarchical approach to intelligence emphasizing the “g” factor. Listed under the “g” factor were verbal-educational and spatial-mechanical group factors, and these were further broken down into minor group factors. R. B. Cattell (1963) and Horn (1985) suggested two types of intelligence: fluid, which referred to capacity and which was independent of experience, and crystallized, which was learned knowledge. Campione and Brown (1978) stressed an information-processing approach to intelligence. Sternberg (1986) saw intelligence as consisting of three dimensions: the componental dimension, which related to internal mental mechanisms; the experiential dimension, which related to both the external and internal worlds; and the contextual dimension which related to the external world of the individual. Sternberg defined intelligence as “the mental activity involved in purposive adaptation to, shaping of, and selection of real-world environments relevant to one’s life” (p. 33).

Das (1973) proposed a non-hierarchical simultaneous-successive information-processing model as a way of categorizing cognitive ability. Simultaneous processing occurs in an integrated, usually semi-spatial form; successive processing is sequence dependent and temporally based. Jensen (1973) attempted to demarcate two separate but partially interdependent mental functions: associative ability, which is represented by memory and serial-learning tasks, and cognitive ability, which is represented by conceptual-reasoning tasks. Gardner, H. (1983) viewed intelligence in terms of problem solving and finding or creating problems, and he suggested the assessment of a number of competencies for solving problems.

Some have argued that traditional views of intelligence are too restrictive and that what is measured on group IQ tests does not relate to how one functions in the “real world” (e.g., McClelland, 1973; Neisser, 1976). Sternberg, Wagner, Williams, and Horvath (1995), for example, stated, “Even the most charitable view of the relation between intelligence tests scores and real-world performance leads to the conclusion that the majority of variance in real-world performance is not accounted for by intelligence test scores” (p. 913). Sternberg et al. (1995) suggested use of “practical intelligence” measures that would assess an individual’s ability to problem solve and to know how to proceed in real-life situations. Traditional group intelligence tests are said to assess functioning that is more related to school performance than to on-the-job performance or to solving problems of daily living. As might be expected, others (e.g., Barrett and Depinet, 1991) dismiss the utility of practical intelligence in favor of the more traditional (real) measures.

**CRITICISMS OF THE TESTS**

Although group measures of intelligence had been widely accepted following World War I, a number of controversies began to emerge with regard to the abilities of different groups of people. Publication of results of the Army Alpha from World War I revealed that there were considerable differences among scores when scores were classified according to the recruits’ national or racial origin. Differences were noted between the mean scores of recruits of Anglo-American or northwestern European descent and those descended from southern and eastern European backgrounds; and between American whites and those of African-American heritage. Some argued that the mental capability of the average white U.S. Army recruit was equivalent to that of a 12-year-old child. Cronbach (1975) and Haney (1981) provide detailed descriptions of the controversies that erupted in the 1920s. In addition to the scholarly debates in academic circles with regard to the use of group intelligence tests, Lippman (1922a, 1922b; 1923) led a press attack on the value of intelligence testing as a whole. Rebuttals were provided by Freeman (1922), Terman (1922a, 1922b), Brigham (1923), and Yerkes (1923).

Adding to the controversy over observed differences in intellectual capability as a function of national origin and ethnic group, was the issue of whether intelligence arose primarily through hereditary factors or environmental influences. At the time when group-intelligence measures were initially developed, it was assumed that differences in intelligence were largely because of genes. Doubts were cast upon the genetic position when it was found that those scoring poorly on the Army Alpha were usually from relatively low socio-economic backgrounds and from areas where there were scant educational resources.
In support of the environmental side of the controversy, Gordon (1923) conducted a study with gypsy and canal-boat children in England who received little, if any, formal education and found that these children, up to the age of six, scored in the average range on intelligence tests. After that their mental ages failed to progress and their IQs declined, showing the negative impact of a lack of schooling. These children did not show a decline on non-verbal performance tests, and this supported the position that environment, and specifically education, played a major role in IQ. Similar results were obtained in the United States with children living in isolated rural communities or mountainous regions of Kentucky (Hirsch, 1928).

Further research starting in the late 1920s and 1930s continued to suggest that IQ scores could be raised significantly when children were placed in enriched environments. In 1937 Newman, Freeman, and Holzinger published a study involving identical twins who were separated and reared apart shortly after birth. Despite having the same genes, IQ differences as much as 24 points were found between a few pairs whose environments were highly divergent.

The heredity-versus-environment debate continued for many years and resulted in a scholarly survey by R.S. Woodworth in 1941, concluding that heredity and environment both contribute to a given IQ score. In a somewhat similar vein, D.O. Hebb (1949) argued for two kinds of intelligence. One type of intelligence was seen as being acquired genetically; the other represented an interaction of genetically-based potential in interaction with environmental stimulation. Despite these developments, proponents of testing in the early- to mid-1900s remained unwilling to concede that environmental issues played a major role in IQ scores.

In a 1967 publication, A.R. Jensen stressed the importance of environmental influences on intellectual development. However, in 1969, when reviewing what appeared to be a failure of Head Start programs, he indicated that it was “not an unreasonable hypothesis” that genetic factors were involved in the average “Negro-white difference” (Vernon, 1978; p. 266). At this time many psychologists saw intelligence as largely malleable and responsive to a variety of environmental alterations despite a possible genetic base. Jensen (1969) put forward an alternative hypothesis which could be subject to testing. Later his views were solidified in support of genes as the major determinant of IQ (Jensen, 1973). What followed from Jensen’s support of the possibility that genes contributed more to intelligence than did environment and that IQ scores were not as malleable as once believed, was a fire storm of criticism of Jensen, his work, and of anyone else who supported his position in whole or part. Thus, Herrnstein (1973), Shockley (1971), and Eysenck (1971) were similarly pilloried for their views on genes and IQ. Vernon (1979) notes that part of the vitriol that followed Jensen’s publication may have been due to an antiestablishment Zeitgeist which was in vogue at the time of Jensen’s writing. Shuey (1958, 1966), who had also argued for the importance of genetic factors in assessing the intelligence of AfroAmericans, raised less of an uproar because of the earlier period in which that work came to print.

Since Jensen’s work and the subsequent controversies in response to his writings, there have been numerous objections to the use of intelligence tests. Many of these criticisms have centered around the potential negative impact that tests can have on certain groups. Several court battles have been fought over the use of tests for the categorization and placement of children within special-education classes. In some instances tests for the purpose of educational placement have been rejected by the courts only to be reinstated in later court decisions (Reschley, Kicklighter, & McKee, 1988). Since 1970, a number of states have banned the use of intelligence tests within schools or have considered such bans. In some instances, a ban or moratorium has been specifically directed to group intelligence testing. Those who favor testing assert that intelligence tests can be beneficial to individuals who may have good ability but are handicapped by poor past performance or a poor academic record. Gordon and Terrell (1981), who have strongly criticized the misuse of standardized tests, nevertheless, state: “To argue that standardized testing should be done away with or radically changed simply because ethnic minorities and disadvantaged groups do not earn as high scores as do middle-class whites is an untenable position” (p. 1170). These authors do suggest, however, that the wholesale use of standardized tests be greatly reduced. Gordon and Terrell propose the development of alternate devices and procedures that would be “process sensitive instruments designed to elicit data descriptive of the functional and conditional aspects of learner behavior” (p. 1170).
It would be naive to suppose, however, that if tests were developed that could do all the things Gordon and Terrell and other critics suggest, these tests would then be above criticism. Hargadon (1981), for example, states:

As a subject that invites debate and controversy, tests and their uses must rank with religion, politics and sex. Tests, at least in part, are designed to do a dirty job: they help us make discriminating judgments about ourselves, about others, about levels of accomplishment and achievement, about degrees of effectiveness. They are no less controversial when they perform their tasks well than when they perform them poorly. Indeed, it can be argued that the better the test, the more controversial its use becomes. (p. 1113)

THE BELL CURVE CONTROVERSY

Throughout most of the mid-1980s to mid-1990s, emotional debates over hereditary versus environmental factors in intelligence had largely subsided, or at least had been less emphasized in the popular press. Heated discussion in the press and on television talk-shows was rekindled when a book, by Herrnstein and Murray (1994) entitled *The Bell Curve: Intelligence and Class Structure in American Life* appeared and asserted that black-white differences in intelligence were primarily due to genetic influences. Group intelligence tests, along with individual tests, supposedly supported the gene and IQ linkage.

Debate over the relative contributions of genes and of the environment to IQ has existed for many years, as has the contention that certain racial and ethnic groups differ from others with regard to abilities. Approximately 2,000 years ago Cicero acknowledged that Britons were too stupid to make good slaves. The latest such discourse takes place in Herrnstein and Murray's 845-page book, which marshals a vast array of data, tables, and statistics to support a number of specific points. These points include the stance (a) that IQ is due primarily to genes and therefore cannot be easily altered; that blacks score 15 points lower than whites on IQ tests, but that Asians outscore whites, and these score differences are due primarily to genetic differences; (b) that social programs such as welfare and similar efforts designed to assist those at the bottom of the socioeconomic barrel are wasteful because low socio-economic functioning is probably due to low IQ which the social programs will not be able to raise; that the high-IQ readers of their book have been unfairly burdened by having to support these relatively wasteful social welfare programs; and (c) that the future belongs to those with high IQs, or what Herrnstein and Murray cavalierly refer to as the “cognitive elite.” Response to Herrnstein and Murray's work, in both the media and within the community of scholars, has been far more critical than supportive (e.g., Gardner, 1995; Gould, 1995; Kamin, 1995; Miller, 1995; Nisbett, 1995).

Those opposed to the conclusions drawn in *The Bell Curve* state their opposition on many levels. Some of the objections to the book include (a) that Herrnstein and Murray's work presents no new data or novel analyses but simply restates old eugenics arguments; (b) that the authors erroneously present the relative contributions of both genes and the environment; trained geneticists (which they are not) would be unable to delineate the relative contributions of these variables; and (c) that the expression of IQ represents the contributions of both genes and the environment, and that these two factors cannot be discussed in terms of their individual inputs to IQ. It is also argued that there is an overwhelming number of studies (including studies cited in Herrnstein and Murray's own work) showing that IQ can be raised substantially by environmental interventions; and that the observed 15-point gap between black and white IQs which they attribute to race is also seen between racially homogeneous groups, such as Catholics and Protestants in Ireland (Herbert 1995), and between Ashkenazi and Sephardic Jews. It has also been argued, as stated at the outset of this chapter, that IQ is not a comprehensive measure of intelligence and that it is inaccurate to use IQ measures constructed by a dominant social group (whites) to assess those (African Americans, Latinos, and others) who have been excluded from full participation and integration into the larger society.

The level of scientific and public ire that arose in response to *The Bell Curve* is reflected in the following vitriolic quotations: “The Bell Curve...contains no new arguments and presents no compelling data to support its anachronistic social Darwinism (Gould, 1995, p.4); “In short, the Bell Curve is not only sleazy; it is, intellectually, a mess” (Ryan, 1995, p. 28); “I gradually realized that (when reading the Bell Curve) I was encountering a style of thought previously unknown to me: scholarly brinkmanship” (Gardner, 1995, p. 63); “The Bell Curve, a scabrous
piece of racial pornography masquerading as serious scholarship" (Herbert, 1995, p. 249); “The book has nothing to do with science” (Kamin, 1995, p. 99); “The book lays out its evidence in very convincing and well thought out ways, but it is just scholarly window dressing for the same old prejudice that has plagued this country since its very conception. The most detrimental aspect of this book is that it attempts to absolve individuals and society from the necessity of positive interventions for a diverse majority of the American people. Historically, there is significant evidence that this would be a grave error” (Richardson, 1995, pp. 43-44). Many in the Asian community, whom Hernstein and Murray claim to be at the top of the IQ pyramid, have also objected to the genetic and racial positions espoused in *The Bell Curve*. “Asian Americans must not allow themselves to be misused in the service of Murray and Hernstein’s political agenda.” (Chon, 1995, p. 239). The political agenda Chon refers to is the elimination of social programs to aide those disenfranchised members of our society.

One of the most basic errors *The Bell Curve* is accused of making is its attempt to equate human intelligence with IQ scores. The single number representing IQ is only distantly related to one’s ability to perceive and understand the environment, to draw reasonable conclusions, to understand social context, to demonstrate creative cognitive processes, and to show altruism and empathy. All of these capabilities, and many more, have been tied to intelligence and cannot be meaningfully reflected in a single score.

Those supporting the position of genetically based black-white IQ differences find fault with explanations that emphasize environmental influences. Jensen (1995), for example argues that “Individual differences in adult IQ are largely genetic, with a heritability of about 70 percent” (p. 335). The three most common environmental explanations for black-white differences in tested IQs are those that point to disadvantages or oppression cultural differences, and psychological maladjustments (Frisby, 1995). Disadvantages or Oppression explanations assert that black children are unable to achieve as a group on a level commensurate with whites because they have been historically denied commensurate opportunities to develop educationally (e.g., Myrdal, Sterner, & Rose, 1962). The cultural-differences explanation proposes that blacks are hindered in academic and testing situations because they are forced to accept and learn about a culture that is alien to their natural culture (e.g., Allen & Boykin, 1992). Finally, the psychological-maladjustment explanation asserts that a combination of racism, poverty, and cultural incongruence causes psychological damage, such as impaired self-esteem, and that such maladjustment impairs academic functioning and performance on IQ tests (e.g., Boykin, 1986).

Frisby (1995) refutes these environmental explanations as being the primary reason for black-white differences and supports Hernstein and Murray’s work on the “facts” of genetic influence, by stating: “When facts and orthodoxy collide, the bearer of the facts is reflexively accused of being ‘elitist,’ ‘racist,’ ‘incompassionate [sic] toward the plight of minorities’, ‘culturally insensitive,’ and ‘ideologically reactionary’.” Thus, Frisby sees the arguments put forth in *The Bell Curve* as convincingly supporting a genetic basis of intelligence and environmental explanations as efforts to deny harsh reality.

The debate over the environment or genes and IQ scores has in no way been resolved nor is it likely to be settled anytime in the near future. It is a debate that evokes strong emotions, and seemingly compelling data can be presented to support both sides. Intelligence tests themselves, whether in group or individual format, are not constructed to shed light on the heredity-environment issue, but rather, have been used as a tool in the debate. What intelligence tests do is to reliably measure how an individual functions in response to specific items at a given point in time. The types of items seen on intelligence tests are intended to correlate with various areas of “real world” functioning, such as performance in school or in employment settings. Whether this generalization from test to real world is, in fact, achieved is another matter for debate. However, items within group intelligence tests assess a variety of areas such as reasoning, general knowledge, vocabulary, problem-solving ability, and nonverbal skills.

**TEST CONTENT**

The most common items on group intelligence tests are those that require some form of reasoning ability. For example, analogies can be presented in a verbal format or in a nonverbal manner where one is required to understand the relationship
between various forms. A typical verbal analogy that involves reasoning skills would be of the form below:

WOOD is to TREE as PAPER is to: LAKE IRON PEN MILL PULP

Other types of verbal reasoning items include similarities, such as the example below:

SMART means the same as: LIVELY HAPPY AGREEABLE CLEVER

Another type of item called oddities, is of the type that follows:

Underline the word that does not belong with the others: DESK TABLE CHAIR BOOK BOOKCASE

Other types of reasoning items involve logical reasoning, as in the example below:

Bob is shorter than John
Ralph is taller than Bob
Who is the shortest?

Items that depend upon reasonable inferences and judgments, based on the information given, are called inferential conclusions. These items are similar in form to items of reading comprehension, except that when used in intelligence tests, the level of vocabulary and reading difficulty are kept simple so that items are not dependent upon vocabulary or reading per se. An example of this kind of item, reported by Jensen (1980), is as follows:

In a particular meadow there are a great many rabbits that eat the grass. There are also many hawks that eat the rabbits. Last year a disease broke out among the rabbits and most of them died. Which one of the following things most probably occurred?

a) The grass died and the hawk population decreased.
b) The grass died and the hawk population increased.
c) The grass grew taller and the hawk population decreased.
d) The grass grew taller and the hawk population increased.
e) Neither the grass nor the hawks were affected by the death of the rabbits. (p. 151)

In a random sample of the adult population in the United States, 52 percent chose the correct answer which was "c".

Other reasoning items are of a numerical nature, as shown below:

John is twice as old as Jim, who is four years old. How old will John be when Jim is 15?

An example of a number series requiring reasoning is as follows:

Write the number that will complete the following series: 3 8 13 18 23 _____

The majority of items found on group intelligence tests are of the multiple-choice variety. Concern has been raised about multiple-choice tests in that some assert that such items measure only superficial knowledge. This argument may certainly be true in some instances. For example little reasoning is needed for the following question:

Which measure is equivalent to an average?

a) Mean
b) Median
c) Mode
d) Quartile

However, the following question requires fairly sophisticated reasoning and knowledge of statistics:

The correlation of SAT-verbal and SAT-Math among all test takers is about .5. For a group of applicants admitted to Harvard University, the correlation is probably

a) Greater than .5
b) About .5
c) Less than .5
d) No way to determine

For this question one must reason that since Harvard University is highly selective, the group of applicants who are admitted must have fairly homogeneous test scores, so the correlation will be lower than that of the national group.

In addition to items that require reasoning, there are questions that are based on knowledge of vocabulary, as shown in the example below:

Enmity means a) opponent b) hatred c) love d) vacant

Vocabulary items have been criticized because they rely heavily on educational background. Acquisition of vocabulary is not just a matter of learning and memory, but also requires discrimination, generalization, and education. Throughout one’s life everyone hears many more words than become part of their vocabularies. Some people,
however, acquire much larger vocabularies than others, and this is true even among siblings of the same family. The effective use of vocabulary requires the ability to make fine discriminations and to reason abstractly. Therefore, it is of little surprise that knowledge of vocabulary is a critical component in "g." In addition to written presentations, vocabulary items can be presented in pictorial form to be used with children and nonreaders; they sometimes appear on group intelligence tests in the lower grades.

Items that tap an individual's general fund of information may also be used on group intelligence tests and are open to the same criticism that is applied to vocabulary items. They correlate highly with other noninformation measures of intelligence because an individual's range of knowledge is a good indication of ability. These items provide the most problems with respect to cultural differences because of the difficulty in determining the range of information an individual from a different culture might be expected to know. For this reason, vocabulary items and general information items do not appear as frequently on many group intelligence tests today as they once did. There are many other types of verbal and nonverbal test items, all of which have been shown to make a contribution to "g." When tests must be administered to large groups, as most group intelligence tests are, issues such as ease of administration and ease of scoring become important factors, and these influence the selection of items.

**TEST SCORES**

Group intelligence tests used in academic settings can be subdivided into three categories, which are based on the types of score they yield (Nitk, 1983). These are the single-score omnibus tests, the three-score tests, and the multiple-aptitude tests. A single-score omnibus test reports one score that encompasses several different aspects of general scholastic ability combined into a single number. An example of an omnibus group-administered intelligence test is the Otis-Lennon School Ability Test (Otis & Lennon, 1977). Here a single score incorporates various areas of cognitive functioning. Three-score tests are divided into levels, and yield three different scores. For example, the Cognitive Abilities Test (CogAT) will yield scores measuring verbal, quantitative, and nonverbal abilities. Finally, multi-scored tests provide a wider profile about the testee, with the popular Differential Aptitude Tests (DAT) yielding nine different subtest scores.

**GROUP ABILITY TESTS**

What follows is an overview of some of the group intelligence and abilities tests that are currently in use and which fall into Nitk's (1983) categories. The tests covered are by no means an exhaustive listing, as a complete summary of group intelligence tests would become a large text. Rather, what is provided is a representative sampling of measures so that the reader will develop a perspective of the kinds of group tests that are in common use. Included in this review will be standard group intelligence measures along with tests that are less sensitive to cultural influences and specialized tests for college entry and employment.

**Otis-Lennon Mental Ability Test (OLMAT):**

One of the more popular group intelligence tests that is intended to provide a measure of "g," the general intellectual factor, is the OLMAT (Otis & Lennon, 1969). The OLMAT evolved from the Army Alpha Examination of World War I, as Arthur Otis was a contributor to both instruments. The current test reflects many characteristics of the Army Alpha, but it is more refined with regard to psychometric properties. The stated purpose of the test is to generate a comprehensive, carefully articulated assessment of general mental ability, or scholastic aptitude, of American students, through a battery of tests that provides scores from kindergarten through Grade 12 (ages 5–18 years). Items are hand-scored, and the scale has a mean of 100 (called a deviation IQ) and a standard deviation of 16. Despite the fact that an IQ is derived, the authors of the test appear to take a tentative stand on whether the test is, in fact, a measure of intelligence. At one point the test is called a measure of general mental ability, yet at a later point, the reader is informed that the tests do not measure native endowment. Regardless of whether group intelligence tests do measure intelligence, virtually all of the group measures report statistically significant reliability and validity data. The manual for the OLMAT provides extensive data on reliability and is based on samples in excess of 120,000; the
split-half and KR-20 coefficients range from 0.93 to 0.96. Reliability coefficients vary as a function of the age and grade level assessed. The test manual does not report validity data.

A further outgrowth of the Otis series of tests is the Otis-Lennon School Ability Test (OLSAT) (Otis & Lennon, 1977). Like its predecessor, the OLMAT, the OLSAT is a pencil-and-paper, multiple-choice test that is group administered and objectively scored. The test is a multilevel battery that is suitable for school settings (grades 1 through 12) and is designed to measure abstract thinking and reasoning ability. The purposes of the OLSAT are to assess the examinee’s ability to cope successfully with school-learning tasks and use the results for placing students in classes. The focus on school learning dispenses with the potential interpretational problems that arise when terms such as mental ability, intelligence, or mental maturity are used. In fact, there is a change from Deviation IQ (DIQ) as used in the OLMAT to the School Ability Index (SAI) on the OLSAT. Nevertheless, the OLSAT, like the OLMAT, is designed to assess a verbal educational factor, and the SAI has the same psychometric properties as the DIQ. The OLSAT, as its predecessor, the OLMAT, is based on a defensible standardization procedure involving 200,000 students in 200 school districts (Swerdlik, 1992). Reliability estimates range from .84 to .95, depending on the level within the test that is assessed and the method of computing reliability. Validity coefficients range from .40 to .60, and these values are typical of well-constructed psychological tests. The Primary I level of the test consists of objects familiar to the child—ice cream cones, animals, stars, etc., and is thereby helpful in holding the child’s attention.

Lorge Thorndike Intelligence Test

The Lorge Thorndike Intelligence Test (Lorge & Thorndike, 1966) is another popular group-administered scale that is applicable to grades 3 through 13. This test measures abstract intelligence, and like the OLMAT, contains both verbal and nonverbal items. The verbal battery is made up of five subtests that include vocabulary, verbal classification, sentence completion, arithmetic reasoning, and verbal analogies. The nonverbal battery contains subtests of pictorial classifications, pictorial analogies, and numerical relationships.

The current edition, called the Multi-Level Edition, has more representative norms than the earlier Separate Level Edition. Validity estimates are readily established as the Lorge-Thorndike was normed on the same samples used for the Iowa Test of Basic Skills, a group administered test for grades 3 through 8 and the Tests of Academic Progress for grades 9 through 12. Correlations with school performance are typical of the various group tests, and in one case are reported to be .87 with reading and .76 with math. Moderate but significant correlations are found between the Lorge-Thorndike and the WISC and the Stanford Binet, and range from .54 to .77. The conglomerate of different types of verbal and nonverbal items appears to represent an attempt to assess “g” by utilizing some arrangement of tests that correlate with each other and which therefore are assumed to share a common global or general intellectual process.

Multidimensional Aptitude Battery (MAB)

The MAB (Jackson, 1984) has been administered to various normal and special populations, such as business people, high school and college students, prison inmates, and psychiatric patients. The proposed purpose of the test is the assessment of intellectual abilities. The MAB attempts to transfer the structure of the WAIS-R into a format suitable for group administration. The MAB can be administered either individually or in a group setting, and it consists of five verbal scales and five nonverbal performance scales. Each subtest has a time limit of seven minutes, which means that the entire battery can be administered in 90 minutes. Verbal, Performance, and Full Scale IQs have been calibrated to match those of the WAIS-R. Test-retest reliabilities for the MAB are .95 for Verbal IQ, .96 for Performance IQ, and .97 for Full-Scale IQ. The MAB has the advantage of standardized group administration without sacrificing reliability and validity, and it is generally seen as having strong psychometric properties.

Cognitive Abilities Test (CogAT)

The CogAT, Form 5 is based on the CogAT, Forms 1–4 (Thornode & Hagen, 1986), and the Lorge-Thorndike Intelligence Test (Lorge & Thorndike, 1966). It can be used for Kindergarten through 12th grade. There are 10 different levels of
the test (Levels 1–2 and Levels A–H.) Each level of the test contains three separate batteries that yield separate scores for verbal-, quantitative- and nonverbal-reasoning abilities. A composite score is also computable. The standardization sample consisted of over 160,000 students from public, Catholic, and private non-Catholic schools. The CogAT is a popular and well-established test of educational aptitude that has undergone complete restandardization. It has strong psychometric properties with reliability and validity estimated from the .70s to .90s.

Henmon-Nelson Tests of Mental Ability

The Henmon-Nelson Tests of Mental Ability (Lamke, Nelson, & French, 1973) are group measures of mental ability that have four levels. There is a Primary Battery (Grades K–2), a battery for Grades 3–6, a battery for Grades 6–9, and a battery for grades 9–12. The Primary Battery requires approximately 30 minutes to administer, while the batteries for Grades 3–12 take exactly 30 minutes to administer.

Test items are centered around subjects related to academic functioning, such as vocabulary, sentence completion, opposites, general information, verbal analogies, verbal classification, verbal inference, number series, arithmetic reasoning, and figure analogies. No reading is required for the Primary Battery. The norms for the Henmon-Nelson Tests of Mental Ability were obtained by stratified random sampling of over 40,000 students in the years 1972–1973. Raw scores, Deviation IQ scores by age, age percentile ranks and stanines, and grade percentile ranks and stanines can be calculated.

The Culture-Fair Intelligence Test (CFIT)

The CFIT (R. B. Cattell, 1973) is a nonverbal measure of an individual’s intelligence. This assessment instrument is designed to overcome the influences of verbal fluency, cultural background, and educational level. The CFIT is said to be unique in that it was designed to measure fluid abilities, whereas traditional tests stress the measurement of crystallized abilities. Thus, in theory, the CFIT allows an evaluation of the future potential of an individual, rather than assessing past achievements or lack of achievements.

The tests are of paper-and-pencil format and have time limits for each subtest. Scale 1 is intended for children aged 4–8 years and for retarded adults. This particular scale is not considered by the test authors to be group-administered or fully culture-free. Scale 2 is for ages 8 to 13 years, and for average adults, and Scale 3 is for high school students and superior adults. The participant’s total working time is only 12 1/2 minutes, but the total administration time is closer to 30 minutes. The CFIT has been criticized for its lengthy instructions that cause children to lose attention and become bored. Another criticism is that bright adults with learning disabilities, particularly those with left-right reversal difficulties, are said to obtain low scores on this test (Vane & Motta, 1990).

Internal consistency coefficients averaged across samples are: Scale 1, .91; Scale 2, .82; and Scale 3, .85. Test-retest reliabilities are: Scale 1, .80; Scale 2, .84; and Scale 3, .82. The CFIT correlates with other intelligence measures in the mid-.70 range. Despite this, several studies of the CFIT have produced mixed results. For example, it has been shown that there are only moderate correlations from .20 to .50 with scholastic achievement, although predictive validities have been fairly impressive for certain groups and criteria. Moreover, correlations with other intelligence tests are mostly between the .50 to .70 range, suggesting that the test is measuring the “g” factor. The CFIT has been administered to many culturally diverse groups outside of the United States and produces scores that are comparable between groups. Although the tests show somewhat lower correlations with socioeconomic status than culture-loaded or other primarily verbal tests, and some bilingual immigrant groups score higher on these tests than on conventional IQ tests, the CFIT does not greatly reduce the magnitude of score differences when administered to culturally disadvantaged groups.

Raven’s Progressive Matrices

Another test that might be considered to be culture-fair or culturally reduced is the Raven’s Progressive Matrices (Raven, 1941, 1981; Raven, Court, & Raven, 1983, 1985). There are two widely used versions of the test: the Standard and the Colored versions (Naglieri & Prewett, 1990). The test was introduced in 1938 and has gone
through many revisions. Because it is nonverbal, and in most situations requires little more than having the examinee point to the correct item, it is often used in situations where examiners want a measure of ability that is not biased by educational background or by cultural or linguistic deficiencies. All of the test items are composed of geometric figures that require the test taker to select among a series of designs the one that most accurately represents or resembles the one shown in the stimulus material. The test items are presented in graded levels of difficulty and there are test booklets for different age levels. Validity measures involving the correlation of the Raven Matrices with the Stanford-Binet and the Wechsler Scales range from .54 to .86 (Raven, Court, & Raven, 1983, pp. 8–9). The authors indicate that “the scales can be described as ‘tests of observation and clear thinking….By themselves they are not tests of ‘general intelligence’….They should be used in conjunction with a vocabulary test” (p. 3). Despite this caution, the Progressive Matrices have been viewed as measures of intelligence and have been widely used in many countries to test military groups because they are considered to be independent of prior learning.

**Test of Nonverbal Intelligence-2 (TONI-2):**

The TONI-2 by Brown, Sherbenou, and Johnsen (1990) is a language-free intelligence test which does not require the examinee to read, write, speak, or listen. It can be used in small groups and is often given individually (Murphy, Conoley, & Impara, 1994). The test is intended to be useful for those who are bilingual, non-English speaking, or who have difficulty reading, writing, speaking, or hearing. The items require the subject to decide how several figures are related by choosing from four to six alternatives, and to indicate which one of these goes best with three or more of the presented stimuli. The figures are black-and-white line drawings; some are simple geometric figures; others are more abstract. The examinee is required to point to the correct response. The TONI-2 is appropriate for individuals aged 5–86 years, and requires 15 minutes to administer in either group or individual format. According to Naglieri and Prewett (1990), “The primary ability assessed by the TONI is problem solving” (p. 359). The national standardization sample consisted of 2,500 persons. There are two equivalent forms of the TONI-2; each contains a variety of problem-solving tasks presented in ascending order. A “TONI quotient” is yielded with a mean of 100 and a standard deviation of 15.

**Expressive One-Word Picture Vocabulary Test: Upper Extension (EOWPVT-U)**

This vocabulary test was designed to be an upward-age extension of the Expressive One-Word Picture Vocabulary Test-Revised (EOW-PVT-R) (Gardner, M.F., 1983), which is individually administered to individuals aged 2–12 years. The EOWPVT-U (Gardner, 1990) is used for individuals aged 12 to 16 years and can be group or individually administered. It was developed by psychologists, counselors, physicians, learning specialists, speech therapists, social workers, diagnosticians, and other professionals. It is said to provide a valid and readily obtainable assessment of a student’s verbal intelligence. The testee is shown a stimulus picture and is required to demonstrate his or her ability to understand and use words by naming simple objects or providing vocabulary words for abstract concepts that are illustrated in pictures. The EOWPVT-U can be administered individually in an oral-testing situation, or in a group format by having students write their responses. The EOW-PVT-U provides mental ages, percentiles, and stanines, and deviation-IQ scores are calculated.

**Black Intelligence Test of Cultural Homogeneity (BITCH)**

At the other end of the continuum, away from tests which claim they are culture-free or -fair, are those tests that are designed to reflect a unique knowledge of a given culture. An example of this type of scale is the BITCH Culture Specific Test by Williams (1972). The author of this test set out to demonstrate that one’s performance on a test can be affected by cultural experience. The test contains a vocabulary that reflects African-American slang, and as might be expected, African Americans score significantly better on it than do whites. The test’s value probably is that it illustrates the extent to which one’s performance can vary as a result of prior knowledge; its major drawback is that it does not correlate with known measures of intelligence (Matarazzo & Wiens, 1977). The latter finding can, of course, be countered by the notion...
that one would not expect to find a correlation between popular standardized intelligence tests that are considered by some to be culturally biased in terms of the dominant culture and one that is biased in favor of a minority group.

The Draw-A-Person (DAP) Test (A Quantitative Scoring System)

The DAP Test (A Quantitative Scoring System) (Naglieri, 1988) is a recently published system of scoring human-figure drawings “to obtain an estimate of ability” (Naglieri & Prewett, 1990, pg. 363). Although administered individually, it can also be administered in groups. There are 64 items comprising a rating scale for the drawings of a man, a woman, and one’s self. The drawings are rated according to the number of body parts that are drawn and the extent to which these parts are elaborated and detailed. Also of importance in the scoring system is the proportion of the parts of the body to one another, and the manner in which these parts are connected. The DAP Test is intended to be used as a part of the larger group of tests or for screening purposes. Since the test is nonverbal and easily administered, the influences of verbal skills, primary language, fine motor coordination, cultural diversity, and language disabilities are said to be reduced. (Naglieri & Prewett, 1990, pg. 364).

The DAP Test was normed on 2,622 individuals aged 5–17 years. According to Naglieri (1988) the normative sample is representative of the 1980 U.S. census data on the stratification variables of age, sex, race, geographic region, ethnic group, socioeconomic status, and community size. The test yields possible scores for each of the three drawings (man, woman, self). These three raw scores are then combined to form a total test score, which is then converted to a standard score with a mean of 100 and a standard deviation of 15. Percentile ranks, age-equivalent scores, and confidence intervals are available for the individual drawings and for the DAP Total Test Score.

The use of the DAP Test as a measure of nonverbal ability or intelligence has not gone without criticism. Motta, Little, and Tobin (1993), for example, point out that the DAP Test has doubtful validity because it correlates weakly with more established measures of intellectual functioning. These authors suggest that the DAP Test has little use in psychodiagnostics because of its psychometric weaknesses and that it should be used only for rapport-building purposes. Similar doubts regarding the utility of drawings in psychological assessment have been voiced by other researchers (e.g., Oakland & Dowling, 1983; Phil & Nimrod, 1976; Weerdenburg & Janzen, 1985). Despite these concerns, many continue to use figure drawings to assess intellectual functioning.

The Scholastic Aptitude Test (SAT)

In testing for college entrance, one test dominates the field: the College Entrance Examination Board’s Scholastic Aptitude Test (SAT). This test is given by the College Board to all high school students throughout the nation who wish to take it. Many selective colleges require SAT scores, but in many colleges scores are only one of the factors used in admitting students. Most colleges, however, do have minimum SAT-cutoff scores.

The SAT is a paper-and-pencil test containing 150 multiple-choice items, with five choices each. There is a verbal section involving reading comprehension, antonyms, verbal analogies, and sentence completion. The mathematics section consists of numerical and quantitative-reasoning items, but does not tap formal mathematical knowledge per se. The SAT would undoubtedly load heavily on the “g” factor in any factor-analytic study that included other mental-ability tests (Vane & Motta, 1984, 1990). The verbal section has been found to correlate higher than the quantitative score with college grade-point average.

The validity of the SAT in predicting scores of minority-group students has frequently been challenged. For example, Stanley and Porter (1967), conducted a study that involved students in three African-American, coeducational, four-year state colleges and compared them with students in 15 predominantly white state colleges in Georgia. Correlations of the combined scores with freshman grade-point averages was .72 for white females, .63 for African-American females, and .60 for both white and African-American males, suggesting that the prediction for white females is better than for the other three groups. Several studies have shown that high school grade-point averages predict college grade-point averages better than the SAT for whites, but not for African-Americans (Cleary, 1968; McKelpin, 1965; Munday, 1965; Peterson, 1968).
The Wonderlic Personnel Test (WPT)

In the area of employment, tests have been used in making employment decisions in the United States for over 70 years. Although there are many content-validated job-knowledge tests and job-sample tests such as typing tests, the most commonly used measures have been measures of cognitive skill, called either aptitude or ability tests. According to Schmidt and Hunter (1981), who performed a meta-analysis of a large number of studies in the field of employment testing, the results show that:

professionally developed cognitive ability tests are valid predictors of performance on the job and in training for all jobs in all types of settings...[and that] cognitive ability tests are equally valid for minority and majority applicants and are fair to minority applicants in that they do not underestimate the expected job performance of minority groups. (p.1128)

Schmidt and Hunter (1981) reported results of a study of 370,000 clerical workers, which showed that validity of seven cognitive abilities was essentially constant across five different clerical-job families. All seven abilities were highly valid in all five job families.

The WPT (Wonderlic, 1977) is one of the group intelligence tests designed for use in employment selection. It can be administered individually or in a group setting. The author intentionally uses the term personnel rather than intelligence to reduce the anxiety of those who must take the test. Despite this, the test manual clearly indicates that the intended use of the instrument is to assess mental ability so that a suitable match can be made between the applicant’s ability and the ability demanded for a particular job area. Dodrill and Warner (1988) found a high degree of correspondence between the WAIS and the WPT in an evaluation of psychiatric, neurological, and normal participants. They conclude that their results “point to the Wonderlic as a measure of general intelligence” (p.146)

The WPT is administered in only 10–12 minutes. There are 50 questions, which examinees usually do not finish; test items require the examinee to reason in terms of words, numbers, and symbols, and to use ideas when thinking. The test items include vocabulary, sentence rearrangement, logic, arithmetic, and interpretation of proverbs (e.g., a rolling stone gathers no moss). Reliabilities range from .82 to .94. The test has 14 different forms and has been standardized in business situations, using large numbers of people and test sites. Minimum scores are reported for professions ranging from custodian to administrator and executive. The score reported is the number correct instead of an IQ, thus reducing some of the controversy evoked by the latter term. Norms are available based on sex, age, range, and educational level. The test reportedly correlates .91–.93 with the WAIS Full-Scale IQ.

Drawbacks of the WPT are that reading skill is required to take the test and speed is a factor. As a result, it would penalize those with psychomotor deficits or reading deficiencies. Because the test provides only a single score, it may not be as diagnostically useful as longer tests. These disadvantages are offset by the obvious benefits of a reliable and valid group measure of intelligence that is easily administered and scored. If used in the right context, it is a valuable test.

SUMMARY

Group and individual assessment of human abilities have a long, colorful, and often emotional history. Many theoretical perspectives regarding the nature of intelligence have been put forward, and numerous assessment devices have been developed, and continue to be developed, for the purpose of assessing abilities. Most of the group intelligence measures being used today correlate significantly with individually administered intelligence tests, and both have proven their utility in the areas of education and employment. Questions, such as the nature of intelligence, whether greater emphasis should be placed on practical measures of ability, or whether existing measures can fairly assess minority groups or members of other cultures, remain embroiled in controversy and will continue to be debated as they have been in the past. Nevertheless, group intelligence tests provide an economical way to readily assess large numbers of individuals, and for this reason, will continue to be useful tools in helping to make personnel decisions.

REFERENCES


