CHAPTER 1

IQ Tests: Their History, Use, Validity, and Intelligent Interpretation

The field of intelligence, particularly of adolescent and adult mental development, has dominated the psychological literature for decades, and now encompasses a diversity of domains within cognitive psychology, clinical psychology, psychobiology, behavioral genetics, education, school psychology, sociology, neuropsychology, and everyday life. Excellent handbooks are available with chapters written by experts in many aspects of intellectual theory, measurement, and development (e.g., Flanagan, Genshaft, & Harrison, 1997; Groth-Marnat, 2000), and even these texts cover only a portion of the territory and quickly become outdated. Consequently, in writing this text on the assessment of adolescent and adult intelligence, we have had to make several decisions about which areas to include and how thoroughly to cover each topic.

First, this book focuses on the clinical assessment of intelligence, and every topic must bear, either directly or indirectly, on the clinical aspect of mental measurement. Because clinical assessment within the fields of neuropsychology, special education, and clinical, school, and counseling psychology involves individual evaluations, research on group-administered tests is subordinated to the more pertinent research on individual intelligence tests. The 1990 version of this text covered group-administered intelligence tests to some extent. However, the adolescent and adult assessment scene has changed during this past decade, with clinicians having options beyond Wechsler's tests. Whereas the Wechsler Adult Intelligence Scale—Third Edition (WAIS-III; Psychological Corporation, 1997; Wechsler, 1997) is still the most used test, and is clearly the featured instrument in this revised text, the availability of a variety of new in-depth and brief intelligence tests, and a proliferation of research on these instruments, has impelled us to focus on individually administered intelligence tests.

For example, the monumental efforts of Schaie (1958, 1983b, 1994) and his colleagues (Hertzog & Schaie, 1988; Schaie & Labouvie-Vief, 1974; Schaie & Strother, 1968; Schaie & Willis, 1993) to understand the development of

adult intelligence have been based on the group-administered Primary Mental Abilities Test. The key findings from these innovative cross-sequential studies are of interest to psychology in general, but have limited applicability to the work of clinical and neuropsychological practitioners. Consequently, investigations by Schaie will only be discussed in the context of aging studies on clinical instruments (e.g., Kaufman, 2000b, 2001; Kaufman & Horn, 1996), especially the WAIS-III, WAIS-R (Wechsler, 1981) and Kaufman Adolescent and Adult Intelligence Test (KAIT; Kaufman & Kaufman, 1993).

Consistent with the focus on clinical tests of intelligence, we have also eliminated sections and chapters from the first edition on clinical tools that are only tangentially related to IQ assessment, most notably neuropsychological instruments, adaptive behavior surveys, and individual achievement tests.

OUTLINE OF THE BOOK

Assessing Adolescent and Adult Intelligence (2nd ed.) has five parts:

I. Introduction to the Assessment of Adolescent and Adult Intelligence (Chapters 1–3)
II. Individual Differences on Age, Socioeconomic Status, and Other Key Variables (Chapters 4–5)
III. Integration and Application of WAIS-III Research (Chapters 6–9)
IV. Interpretation of the WAIS-III Profile: IQs, Factor Indexes, and Subtest Scaled Scores (Chapters 10–12)
V. Additional Measures of Adolescent and Adult IQ (Chapters 13–15)

Part I includes: Chapter 1, which discusses pertinent historical information, issues regarding validation of the IQ construct, and our philosophy of intelligent testing; Chapter 2, which discusses pressing issues and challenges to the IQ concept (e.g., heritability and malleability of the IQ); and Chapter 3, which provides the rationale for the WAIS-III subtests for adolescents and adults and traces the empirical and logical continuity from the Wechsler-Bellevue to the WAIS to the WAIS-R and to the WAIS-III.

Part II presents research on individual differences in intelligence associated with pertinent background variables on the WAIS-III and other instruments, notably gender, ethnicity, socioeconomic status, and urban–rural residence (all treated in Chapter 4), and aging across the adult lifespan (Chapter 5).

Parts III and IV are devoted to the WAIS-III and, occasionally, its predecessors (e.g., WAIS-R) or "alternate-form" at age 16 (WISC-III). In Part III, the focus is on research, delving into topics such as administration and scoring (Chapter 6), factor analysis (Chapter 7), and Verbal Performance (V-P) IQ differences, especially as they pertain to lateralized brain lesions (Chapter 8) and other clinical disorders (Chapter 9). The three chapters of Part IV (Chapters 10, 11, and 12) are all devoted to an empirical and clinical approach to interpretation of the WAIS-III multiscore profile.

Part V is composed of three chapters; each focuses exclusively on additional (non-Wechsler) measures for adolescent and adult assessment and integrates them with the WAIS-III: the KAIT (Chapter 13), the Woodcock-Johnson—Third Edition or WJ III (Chapter 14, authored by McGrew, Woodcock, and Ford), and a variety of brief tests of intelligence (Chapter 15). The tests discussed in the latter chapter, for example, the Peabody Picture Vocabulary Test—Third Edition (PPVT-III), the Kaufman Brief Intelligence Test (K-BIT), and the Wechsler Abbreviated Scale of Intelligence (WASI), may be used as supplements to the WAIS-III, KAIT, or WJ III, or may be used instead of comprehensive intelligence tests in certain circumstances (e.g., screening or research purposes).

The discussion of non-Wechsler tests in Part IV is essential to round out the cognitive assessment scene, but the WAIS-III, like the WAIS-R, WAIS, and Wechsler-Bellevue before it, remains the key tool for clinical and neuropsychological evaluation of adolescents and adults and, hence,

the focus of all sections of the book. The chapters on clinical applications of intelligence tests, along with the previous parts of the book, place the focus of this text squarely on the WAIS-III.

Wechsler's Scales

Even a casual observer of the clinical or neuropsychological assessment scene is aware that Wechsler's scales are uncontested as the primary cognitive measures of adolescent and adult intelligence. Individuals in their teens and adults of all ages are invariably administered the Wechsler Intelligence Scale for Children—Third Edition (WISC-III; Wechsler, 1991) or the WAIS-III when they are referred to a competent professional for a thorough assessment of their intellectual abilities, usually as part of a clinical, vocational, neuropsychological, or psychoeducational evaluation. The WISC-III is used for adolescents as old as 16 years, while the WAIS-III is used for individuals aged 16 to 89. Therefore, they overlap at age 16, giving clinicians a choice of Wechsler test for that age group.

Using the WISC-III as a clinical and psychometric tool has been discussed elsewhere in a comprehensive text (Kaufman, 1994a). For practical purposes, then, this book is primarily devoted to the WAIS-III, child of the WAIS-R (Wechsler, 1981), grandchild of the WAIS (Wechsler, 1955), and great-grandchild of the Wechsler-Bellevue Form I (Wechsler, 1939).

Clinical Relevance of Theory

To be included in this book in any depth, a topic needs to contribute to a psychologist's understanding of intelligence in the clinical arena, not in the laboratory. For example, the Cattell-Horn-Carroll (CHC; McGrew & Flanagan, 1998) theory—an amalgam of Horn's (1989) expansion of Horn-Cattell Gf-Gc theory and Carroll's (1993, 1997) model of intelligence—is treated throughout the book because it is instrumental in explaining changes in verbal and nonverbal abilities with advancing age, and it (or Horn-Cattell theory) underlies three tests of adolescent and adult intelligence: the Woodcock Johnson Psycho-Educational Battery—Third Edition (WJ III; Woodcock, McGrew, & Mather, 2000), the Stanford-Binet Intelligence Scale, Form IV (Thorndike, Hagen, & Sattler, 1986a), and the KAIT (Kaufman & Kaufman, 1993). In contrast, Sternberg's (1985) three-pronged triarchic theory of intelligence, though popular and widely discussed, is not emphasized because of its limited application to clinical assessment and the interpretation of the WAIS-III and other individual intelligence tests. Currently the Sternberg Triarchic Abilities Test (Sternberg, 1993), a group-administered measure, is available as an unpublished research instrument available from its author. However, if it is ever adapted as an individually administered, commercially published, standardized measure that translates laboratory principles to the domain of the clinical psychologist, neuroclinician, and psychoeducational diagnostician, the theory may become even more popular.

In addition, other theories of intelligence such as Gardner's (1993a, 1993b) theory of multiple intelligences—which defines intelligence as the ability to solve problems, or to create products, that are valued within one or more cultural settings—is also not emphasized in this book. The theory of multiple intelligences calls for measuring intelligences by asking individuals to solve problems in the contexts in which they naturally occur. Although the multiple intelligences theory has attracted much attention in the fields of cognition and education (Kornhaber & Krechevsky, 1995), thus far its practical application to clinical assessment and the interpretation of the WAIS-III and other major standardized individual intelligence tests is limited.

A SHORT HISTORY OF IQ TESTS

The history of intellectual assessment is largely a history of the measurement of the intelligence of children or retarded adults. Sir Francis Galton (1869, 1883) studied adults and was interested in

giftedness when he developed what is often considered the first comprehensive individual test of intelligence (Kaufman, 2000a). But despite Galton's role as the father of the testing movement (Shouksmith, 1970), he did not succeed in constructing a true intelligence test. His measures of simple reaction time, strength of squeeze, or keenness of sight proved to assess sensory and motor abilities, skills that relate poorly to mental ability, and that are far removed from the type of tasks that constitute contemporary intelligence tests.

The Binet-Simon Scales

Alfred Binet and his colleagues (Binet & Henri, 1895; Binet & Simon, 1905, 1908) developed the tasks that survive to the present day in most tests of intelligence for children and adults. Binet (1890a, 1890b) mainly studied children; beginning with systematic developmental observations of his two young daughters, Madeleine and Alice, he concluded that simple tasks like those used by Galton did not discriminate between children and adults. In 1904, the Minister of Public Instruction in Paris appointed Binet to a committee to find a way to distinguish normal from retarded children. But 15 years of qualitative and quantitative investigation of individual differences in children—along with considerable theorizing about mental organization and the development of a specific set of complex, high-level tests to investigate these differences—preceded the "sudden" emergence of the landmark 1905 Binet-Simon intelligence scale (Murphy, 1968).

The 1908 scale was the first to include age levels, spanning the range from III to XIII. This important modification stemmed from Binet and Simon's unexpected discovery that their 1905 scale was useful for much more than classifying a child at one of the three levels of retardation: moron, imbecile, idiot (Matarazzo, 1972). Assessment of older adolescents and adults, however, was not built into the Binet-Simon system until the 1911 revision. That scale was extended to age level XV and included five ungraded adult tests (Kite, 1916). This extension was not conducted with the rigor that characterized the construction of tests for children, and the primary applications of the scale were for use with school-age children (Binet, 1911).

Measuring the intelligence of adults, except those known to be mentally retarded, was almost an afterthought. But the increased applicability of the Binet-Simon tests for various child-assessment purposes dawned on Binet just prior to his untimely death in 1911: "By 1911 Binet began to foresee numerous uses for his method in child development, in education, in medicine, and in longitudinal studies predicting different occupational histories for children of different intellectual potential" (Matarazzo, 1972, p. 42).

Terman's Stanford-Binet

Lewis Terman was one of several people in the United States who translated and adapted the Binet-Simon scale for use in the United States, publishing a "tentative" revision (Terman & Childs, 1912) 4 years before releasing his painstakingly developed and carefully standardized Stanford Revision and Extension of the Binet-Simon Intelligence Scale (Terman, 1916). This landmark test, soon known simply as the Stanford-Binet, squashed competing tests developed earlier by Goddard, Kuhlmann, Wallin, and Yerkes. Terman's success was undoubtedly due in part to heeding the advice of practitioners whose demand "for more and more accurate diagnoses …raised the whole question of the accurate placing of tests in the scale and the accurate evaluation of the responses made by the child" (Pintner & Patterson, 1925, p. 11).

But, like Binet, Terman (1916) saw intelligence tests useful primarily for the detection of mental deficiency or superiority in children and for the identification of "feeblemindedness" in adults. He cited numerous studies of delinquent adolescents and adult criminals, all of which pointed to the high percentage of mentally deficient juvenile delinquents, prisoners, or prostitutes, and concluded that "there is no investigator who denies

the fearful role played by mental deficiency in the production of vice, crime, and delinquency" (p. 9). Terman also saw the potential for using intelligence tests with adults for determining "vocational fitness," but, again, he emphasized employing "a psychologist…to weed out the unfit" or to "determine the minimum 'intelligence quotient' necessary for success in each leading occupation" (p. 17).

Perhaps because of this emphasis on the assessment of children or concern with the lower end of the intelligence distribution, Terman (1916) did not use a rigorous methodology for constructing his adult-level tasks. Tests below the 14-year level were administered to a fairly representative sample of about 1,000 children and early adolescents. To extend the scale above that level, data were obtained from 30 businessmen, 50 high school students, 150 adolescent delinquents, and 150 migrating unemployed men. Based on a frequency distribution of the mental ages of a mere 62 adults (the 30 businessmen and 32 of the high school students above age 16), Terman partitioned the graph into the following MA categories: 13–15 (inferior adults), 15–17 (average adults), and above 17 (superior adults).

The World War I Tests

The infant field of adult assessment grew rapidly with the onset of World War I, particularly after U.S. entry into the war in 1917 (Anastasi & Urbina, 1997; Vane & Motta, 1984). Psychologists saw with increasing clarity the applications of intelligence tests for selecting officers and placing enlisted men in different types of service, apart from their generation-old use for identifying the mentally unfit. Under the leadership of Robert Yerkes and the American Psychological Association, the most innovative psychologists of the day helped translate Binet's tests to a group format. Arthur Otis, Terman's student, was instrumental in leading the creative team that developed the Army Alpha, essentially a group-administered Stanford-Binet, and the Army Beta, a novel group test composed of nonverbal tasks.

Yerkes (1917) opposed Binet's age-scale approach and favored a point-scale methodology, one that advocates selection of tests of specified, important functions rather than a set of tasks that fluctuates greatly with age level and developmental stage. The Army group tests reflect a blend of Yerkes's point-scale approach and Binet's notions of the kind of skills that should be measured when assessing mental ability. The Army Alpha included the Binet-like tests of Directions or Commands, Practical Judgment, Arithmetical Problems, Synonym-Antonym, Dissarranged Sentences, Analogies, and Information. Even the Army Beta had subtests resembling Stanford-Binet tasks: Maze, Cube Analysis, Pictorial Completion, and Geometrical Construction. The Beta also included novel measures like Digit Symbol, Number Checking, and X-O Series (Yoakum & Yerkes, 1920).

Never before or since have tests been normed and validated on samples so large; 1,726,966 men were tested (Vane & Motta, 1984)! Point-scores on the Army Alpha or Army Beta were converted to letter grades ranging from A to D- (the Beta was given only to illiterate and non-English-speaking candidates). Validity was demonstrated by examining the percent of A's obtained by a variety of Army ranks, for example, recruits (7.4%), corporals (16.1%), sergeants (24.0%), and majors (64.4%). In perhaps the first empirical demonstration of the Peter Principle in action, second lieutenants (59.4% A's) outperformed their direct superiors—first lieutenants (51.7%) and captains (53.4%)—while those with ranks above major performed slightly worse than majors (Yoakum & Yerkes, 1920, Table 1). Can there be any more compelling affirmation of the validity of the Army intelligence tests? Another intelligence scale was developed during the war, one that became an alternative for those who could not be tested validly by either the Alpha or Beta. This was the Army Performance Scale Examination, composed of tasks that would become the tools-of-trade for clinical psychologists, school psychologists, and

neuropsychologists into the twenty-first century: Picture Completion, Picture Arrangement, Digit Symbol, and Manikin and Feature Profile (Object Assembly). Except for Block Design (developed by Kohs in 1923), Wechsler's influential Performance Scale was added to the Army battery, "[t]o prove conclusively that a man was weakminded and not merely indifferent or malingering" (Yoakum & Yerkes, 1920, p. 10).

Wechsler's Creativity

David Wechsler assembled a test battery in the mid-1930s that comprised subtests developed primarily by Binet and World War I psychologists. His Verbal Scale was essentially a Yerkes point-scale adaptation of Stanford-Binet tasks; his Performance Scale, like other similar nonverbal batteries of the 1920s and 1930s (Cornell & Coxe, 1934; Pintner & Patterson, 1925), was a near replica of the tasks and items making up the individually administered Army Performance Scale Examination.

In essence, Wechsler took advantage of tasks developed by others for nonclinical purposes to develop a clinical test battery. He paired verbal tests that were fine-tuned to discriminate among children of different ages with nonverbal tests that were created for adult males who had flunked both the Alpha and Beta exams—nonverbal tests that were intended to distinguish between the nonmotivated and the hopelessly deficient. Like Terman, Wechsler had the same access to the available tests as did other psychologists; like Terman and Binet before him, Wechsler succeeded because he was a visionary, a man able to anticipate the needs of practitioners in the field.

While others hoped intelligence tests would be psychometric tools to subdivide retarded individuals into whatever number of categories was currently in vogue, Wechsler saw the tests as dynamic clinical instruments. While others looked concretely at intelligence tests as predictors of school success or guides to occupational choice, Wechsler looked abstractly at the tests as a mirror to the hidden personality. With the Great War over, many psychologists returned to a focus on IQ testing as a means of childhood assessment; Wechsler (1939), however, developed the first form of the Wechsler-Bellevue exclusively for adolescents and adults.

Most psychologists saw little need for nonverbal tests when assessing English-speaking individuals other than illiterates. How could it be worth 2 or 3 minutes to administer a single puzzle or block-design item when 10 or 15 verbal items can be given in the same time? Some test developers (e.g., Cornell & Coxe, 1934) felt that Performance scales might be useful for normal, English-speaking people to provide "more varied situations than are provided by verbal tests" (p. 9), and to "test the hypothesis that there is a group factor underlying general concrete ability, which is of importance in the concept of general intelligence" (p. 10).

Wechsler was less inclined to wait a generation for data to accumulate. He followed his clinical instincts and not only advocated the administration of a standard battery of nonverbal tests to everyone but placed the Performance Scale on an equal footing with the more respected Verbal Scale. Both scales would constitute a complete Wechsler-Bellevue battery, and each would contribute equally to the overall intelligence score.

Wechsler also had the courage to challenge the Stanford-Binet monopoly, a boldness not unlike Binet's when the French scientist created his own forum (the journal *L'Année Psychologique*) to challenge the preferred but simplistic Galton sensorimotor approach to intelligence (Kaufman, 2000a). Wechsler met the same type of resistance as Binet, who had had to wait until the French Ministry of Public Instruction "published" his Binet-Simon Scale. When Wechsler's initial efforts to find a publisher for his two-pronged intelligence test met failure, he had no cabinet minister to turn to, so he took matters into his own hands. With a small team of colleagues, he standardized Form I of the Wechsler-Bellevue by himself. Realizing that stratification

on socioeconomic background was more crucial than obtaining regional representation, he managed to secure a well-stratified sample from Brooklyn, New York.

The Psychological Corporation agreed to publish Wechsler's battery once it had been standardized, and the rest is history. Although an alternative form of the Wechsler-Bellevue (Wechsler, 1946) was no more successful than Terman and Merrill's (1937) ill-fated Form M, a subsequent downward extension of Form II of the Wechsler-Bellevue (to cover the age range 5 to 15 instead of 10 to 59) produced the wildly successful WISC (Wechsler, 1949). Although the Wechsler scales did not initially surpass the Stanford-Binet in popularity, serving an apprenticeship to the master in the 1940s and 1950s, the WISC and the subsequent revision of the Wechsler-Bellevue, Form I (WAIS; Wechsler, 1955) triumphed in the 1960s. "With the increasing stress on the psychoeducational assessment of learning disabilities in the 1960s, and on neuropsychological evaluation in the 1970s, the Verbal-Performance (V-P) IQ discrepancies and subtest profiles yielded by Wechsler's scales were waiting and ready to overtake the one-score Binet" (Kaufman, 1983b, p. 107).

Irony runs throughout the history of testing. Galton developed statistics to study relationships between variables—statistics that proved to be forerunners of the coefficient of correlation, later perfected by his friend Karl Pearson (DuBois, 1970). The ultimate downfall of Galton's system of testing can be traced directly to coefficients of correlation, which were too low in some crucial (but, ironically, poorly designed) studies of the relationships among intellectual variables (Sharp, 1898–99; Wissler, 1901). Similarly, Terman succeeded with the Stanford-Binet while the Goddard-Binet (Goddard, 1911), the Herring-Binet (Herring, 1922), and other Binet-Simon adaptations failed because he was sensitive to practitioners' needs. He patiently withheld a final version of his Stanford revision until he was certain that each task was appropriately placed at an age level consistent with the typical functioning of representative samples of U.S. children.

Terman continued his careful test development and standardization techniques with the first revised version of the Stanford-Binet (Terman & Merrill, 1937). But 4 years after his death in 1956, his legacy was devalued when the next revision of the Stanford-Binet comprised a merger of Forms L and M, *without a standardization* of the newly formed battery (Terman & Merrill, 1960). The following version saw a restandardization of the instrument, but without a revision of the placement of tasks at each age level (Terman & Merrill, 1973). Unfortunately for the Binet, the abilities of children and adolescents had changed fairly dramatically in the course of a generation, so the 5-year level of tasks (for example) was now passed by the average 4½-year-old!

Terman's methods had been ignored by his successors. The ironic outcome was that Wechsler's approach to assessment triumphed, at least in part because the editions of the Stanford-Binet in the 1960s and 1970s were beset by the same type of flaws as Terman's competitors in the 1910s. The newest Stanford-Binet (Thorndike, Hagen, & Sattler, 1986a, 1986b) attempted to correct these problems and even adopted Wechsler's multisubtest, multiscale format. However, these changes in the Fourth Edition of the Binet were too little and too late to be much threat to the popularity of the Wechsler scales, to offer much contribution to the field of intelligence testing, or to merit the linkage with the Binet tradition.

SURVEYS OF TEST USAGE FOR ADULTS

Surveys of test use in the United States have appeared increasingly in the literature in the past decade. These surveys are usually based on data from clinical agencies and hospitals (Lubin, Larsen, & Matarazzo, 1984; Petrowski & Keller, 1989), school systems (Goh, Teslow, & Fuller,

1981; Hutton, Dubes, & Muir, 1992; Wilson & Reschly, 1996), industry (Swenson & Lindgren, 1952), military settings (Lubin, Larsen, Matarazzo, & Seever, 1986), forensic settings (Lees-Hayley, Smith, Williams, & Dunn, 1996), or private practitioners (Archer, Maruish, Imhof, & Piotrowski, 1991; Camara, Nathan, & Puente, 2000; Harrison et al., 1988; Lubin et al., 1986; Watkins, Campbell, Nieberding, & Hallmark, 1995). Data from such studies of test use are becoming increasingly important in light of the role that managed-care companies play in reimbursement for assessment services. Data from surveys that help determine which are the typical instruments used for various types of assessment and the amount of time practitioners usually spend on an assessment may serve a function in setting standard approved rates for practitioner compensation by managed-care companies. Thus, we reviewed the recent literature to attempt to discover which instruments are most commonly used by practitioners with a variety of backgrounds and find out how much time is typically spent on assessments.

Has Test Use Changed over the Years?

Overall, little substantive change has occurred in the most popular instruments used in the last several decades (Camara et al., 2000). Test usage was first documented by Louttit and Brown (1947), with data collected spanning the mid-1930s to the mid-1940s. Since that early survey, subsequent surveys have shown that the most commonly used tests have not changed much over the years. The Wechsler family of tests has remained on the top of the assessment list for most psychologists, across a variety of settings (Ball, Archer, & Imhof, 1994; Brown & McGuire; 1976; Camara et al., 2000; Harrison et al., 1988; Lubin et al., 1971). The WAIS and WAIS-R have consistently been mentioned in surveys as the most often used adult intelligence tests by clinical psychologists, school psychologists, neuropsychologists, and forensic psychologists, and the WAIS-III will surely follow suit in future surveys.

Many studies of test usage lump together tests from all areas of assessment, including intellectual assessment, personality assessment, adaptive functioning assessment, achievement assessment, and neuropsychological assessment. Nonetheless, even when considering all these different types of assessment, the Wechsler tests remain ranked in the top 10.

Because the WAIS-III is fairly new, we were unable to find any published surveys that reported on the latest adult Wechsler test. The most recent survey at the time that this book went to press had a 2000 publication date, but the authors collected their data in late 1994, before the WAIS-R was revised (Camara et al., 2000). However, it is safe to assume that the WAIS-III will maintain the high ranking enjoyed by the WAIS-R.

Test Usage of 1,500 Psychologists and Neuropsychologists

Camara et al.'s (2000) collected survey data on test usage and assessment from 933 clinical psychologists and 567 neuropsychologists who were randomly selected from the American Psychological Association (APA) and the National Association of Neuropsychology (NAN). The authors were interested in data from practitioners who conducted assessments on a regular basis, so they ultimately conducted their analyses on data from respondents who engaged in 5 or more hours per week of assessment-related services. Thus, the final sample used for ranking test usage comprised 179 clinical psychologists (19% of the clinical psychologist respondents) and 447 neuropsychologists (79% of the neuropsychologist respondents). Table 1.1 displays the hours spent administering, scoring, and interpreting psychological tests during a typical week, for the total number of respondents to the survey (N = 1,500).

TABLE 1.1 Hours spent administering, scoring, and interpreting psychological tests during a typical week

	Clinical Psychologists		*Neuropsychologists*	
Hours	***n* (%)**	**Cumulative %**	***n* (%)**	**Cumulative %**
0–4	755 (80.9)	100.0	116 (20.5)	100.0
5–9	62 (6.6)	18.7	62 (10.9)	78.8
10–14	39 (4.2)	12.1	92 (16.2)	67.9
15–20	36 (3.9)	7.9	105 (18.5)	51.7
More than 20	37 (4.0)	4.0	188 (33.2)	33.2
No response	4 (<1)	<1	4 (<1)	<1
Total	**933 (100.0)**		**567 (100.0)**	

NOTE: Data are from "Psychological Test Usage in Professional Psychology," by W. J. Camara, J. S. Nathan, & A. E. Puente, 2000, *Professional Psychology: Research and Practice*, *31*, 141–154. Copyright © by the American Psychological Association. Reprinted with permission.

Interestingly, the sample of neuropsychologists spent many more hours per week doing assessments than did the sample of clinical psychologists. Among neuropsychologists, almost 80% spent at least 5 assessment hours per week and about half spent at least 15 hours a week conducting assessments. For clinical psychologists, the corresponding values were about 20% and 8%.

According to Camara et al. (2000), of the clinical psychologists who performed assessments 5 or more hours per week, the majority of their assessment time was spent conducting intellectual or achievement testing (34%) and personality testing (32%). For neuropsychologists, their assessment time was fairly equally divided between neuropsychological assessment (26%), intellectual or achievement assessment (20%), and personality assessment (20%). Watkins et al. (1995) reported that 8% of a clinical psychologist's total time practicing was spent on intellectual assessment, and 12% of the total time was spent on personality assessment (N = 412). In a study examining assessment practices of school psychologists (N = 389), respondents reported that they spent about one half of their time in assessment-related activities (*Mdn* = 50%) (Hutton et al., 1992).

How Frequently Are Tests Used?

As mentioned, the Wechsler tests have held on strongly to their place at the top of the heap of tests administered by practitioners over the years. In Camara et al.'s (2000) study, clinical psychologists ranked the WAIS-R the number one test administered and neuropsychologists ranked it number two. Other Wechsler tests were also at the top of the list: clinical psychologists rated the WISC-III number 3 and neuropsychologists rated the Wechsler Memory Scale—Revised number 3. Camara et al. (2000) did not separate children's tests from adults' tests, or measures of intelligence from other measures, such as personality functioning. Clinical psychologists ranked the Minnesota Multiphasic Personality Inventory—Second Edition (MMPI-II) as the number 1 most frequently used test and neuropsychologists ranked it as number 2. Other studies report similar findings: in a survey tapping tests administered by

psychologists to adolescent clients, the Wechsler scales were the number one most frequently used tests (Archer, Maruish, Imhof, & Piotrowski, 1991); in a survey of tests administered by forensic neuropsychologists, the WAIS-R, MMPI-II, and WMS-R were ranked numbers 1, 2, and 3, respectively (Lees-Hayley et al., 1996); and school psychologists also reported the Wechsler scales as the most frequently used assessment tools (Hutton et al., 1992; Wilson & Reschly, 1996).

Administration Time and Implications for Reimbursement

Camara et al. (2000) also examined the mean time to administer, score, and interpret a battery of tests. The median number of minutes reportedly spent by clinical psychologists on the WAIS-R was administration (75), scoring (20), and interpretation (30), for a total time of a little over 2 hours; similar values were reported by neuropsychologists. Considering that the WAIS-R (or WAIS-III now) is only one component of a full battery, the total time to administer, score, and interpret an entire battery is significantly greater. Clearly, the time varies depending on the type of testing necessary to answer the referral questions. That being said, Camara and colleagues found that, on average, psychologists spent about 3.5 to 4.25 hours on administering, scoring, and interpreting an assessment battery. However, the authors concede that some areas of assessment take substantially longer than these average times, especially intellectual and neuropsychological assessment.

The results from Camara et al.'s (2000) study have implications for the reimbursement of assessments by third parties, especially managed-care companies. The authors note that assessment services are often limited to 2 hours of reimbursable time, the approximate time the psychologists in Camara et al.'s (2000) study spent administering, scoring, and interpreting the WAIS-R. However, because the Camara data demonstrated that trained practitioners require at least 4 hours to complete a comprehensive assessment, it is clear that clinicians are limited in what types of assessments they can provide, if they want to be reimbursed for their time. The consequences of limited reimbursement for assessment may be that the number of psychologists conducting assessments will diminish. Already, Camara and colleagues note that almost 90% of clinical psychologists spend less than 10 hours a week on assessments (see Table 1.1).

For What Purposes Are Adults Given Intelligence Tests?

It is clear that the WAIS-R and WAIS-III are widely used in the field of assessment today, but why are these and other intelligence tests typically administered to adults? Harrison et al. (1988) asked that question specifically of a group of 277 clinical psychologists. In a survey, respondents were asked to rank seven purposes for which they would administer an intelligence test. The number 1 purpose was to measure the potential or cognitive capacity of a person. Table 1.2 lists the seven purposes and how important respondents felt each was. Although nearly 40–50% of psychologists ranked educational and vocational placement or interventions as a purpose for assessing adults, very few felt these are the main reasons for conducting an assessment (6–17%). Clearly, the data show that clinicians think that the most important reasons for assessing adults are to measure cognitive potential, obtain clinically relevant information, and assess functional integrity of the brain.

Conclusions

The Camara et al. (2000) survey results indicate that the WAIS-R, and, intuitively, the WAIS-III, is supreme among assessment tools used to assess adolescent and adult functioning by clinical psychologists and neuropsychologists. These results, in combination with results of other studies, show that the Wechsler tests are equally

TABLE 1.2 Purposes for using intelligence tests when assessing adults

Purpose	% of Psychologists Who Assess Adults for This Purpose	% of Psychologists Who Rank This Purpose as Very Important
Measure potential of capacity	85.2	58.5
Obtain clinically relevant information	85.2	53.1
Assess functional integrity of brain	77.6	43.3
Determine educational placement	48.4	17.0
Determine vocational placement	45.5	12.3
Develop educational interventions	44.0	10.8
Develop vocational interventions	39.4	5.8

NOTE: Data are from Harrison et al. (1988), based on 277 respondents asked to list all the purposes for which "you generally use a standardized intelligence test in your assessment battery" and "then rank the ones you checked in order of their importance with a 1 as the most important." The "*% of psychologists who rank this purpose as very important*" equals the percentage of the total group of 2,787 who assigned each purpose a ranking of 1 or 2.

popular in other domains such as forensic psychology, school psychology, hospital settings, and outpatient clinics. The percentage of clinical time spent conducting assessments varies across specialties within psychology (e.g., clinical, school, neuropsychology). However, the typical amount of time necessary to conduct an assessment is similar across domains, although it fluctuates depending on the type of assessment necessary. The inconsistency between the amount of time typically allowed to be reimbursed for assessment services and the actual amount of time spent in assessment-related services was pointed out by Camara et al. (2000). Such inconsistency may affect the types and numbers of assessments performed by clinicians. Notwithstanding the fees and reimbursement issues, the popularity of the Wechsler scales and the primary reasons for assessing adults remain unchanged. There appears to be a strong need for tools to assess cognitive capabilities and obtain related clinical information in adults, and the WAIS-III is there to meet those needs for those who choose to conduct assessments. However, clinicians would be wise to consider theory-based alternatives to Wechsler's scales, such as the KAIT and WJ III Tests of Cognitive Ability for adolescents and adults, and the Cognitive Assessment System (CAS; Naglieri & Das, 1997a, 1997b) for adolescents. Also, in view of time constraints imposed by managed-care criteria, reliable and valid brief intelligence tests may need to be weighed as possible assessment options (see Chapter 15).

VALIDITY OF THE IQ CONSTRUCT FOR ADOLESCENTS AND ADULTS

Matarazzo (1972, Chapters 6, 7, and 12) devoted most of three chapters to support the validity of the IQ construct, Jensen (1980) addressed the issue from both theoretical and empirical perspectives (his Chapters 6 and 8, respectively), and Brody (1985) published a thought-provoking chapter on "The Validity of Tests of Intelligence." These three esteemed psychologists concluded, in essence, that the IQ construct, as measured by

contemporary intelligence tests, is valid when defined within the societal context and when the IQ's limitations are kept fully in mind. In a survey of psychologists and educational specialists with expertise in areas related to intelligence testing, Snyderman and Rothman (1987) found that, overall, experts hold positive attitudes about the validity and usefulness of intelligence and aptitude tests. Although the validity of the IQ construct and the tests purported to assess it are important to this text, we treat it cursorily here because it has been thoroughly discussed elsewhere. Our focus is on the following aspects of IQ's validity: prediction of academic achievement, relationship to educational attainment, relationship to occupational membership, and prediction of job performance.

Prediction of Academic Achievement

The age-old IQ criterion of prediction of school achievement has been explored in thousands of studies across the age range, and Matarazzo (1972) concluded a generation ago that a correlation of about .50 exists between IQ and school performance. Coefficients are typically a bit higher in elementary school and lower in college (Brody, 1985). The overall value of .50 is high enough to support the validity of the IQ for the purpose that Binet originally intended it, but low enough to indicate that about 75% of the variance in school achievement is accounted for by factors other than IQ. Some more recent studies with newer, theory-based intelligence tests have reported higher coefficients in the .60–.70 range for the Horn-based WJ-R (McGrew, Werder, & Woodcock, 1991) and for the Luria-based K-ABC and CAS (Naglieri, 1999, Table 5.5) between intelligence and achievement. In fact, these coefficients for the theory-based tests are similar in magnitude to the values obtained with the Third Editions of the WISC and WAIS, using WIAT scores as the criteria (Psychological Corporation, 1992, Table D.6). Hence, more recent studies with new and revised instruments suggest that IQ may explain as much as 50% of school achievement; however, even that substantially higher value still leaves 50% for other variables.

For adults, the IQ-achievement correlations are illustrated by correlations between the WAIS-III and the Wechsler Individual Achievement Test (WIAT; Psychological Corporation, 1992). Overall the correlations between the WAIS-III IQs and the WIAT Composites (Reading, Math, and Language) range between .53 and .81, with most correlations in the .60s and .70s, and a median value of .70 (Psychological Corporation, 1997). The correlations between the WAIS-III Indexes and the WIAT Composites were slightly lower than those with the IQs, with *r*s ranging from .42 to .77 with a median value of .61.

Wechsler's Verbal IQ consistently correlates more strongly with achievement than does the Performance IQ. Correlations between the WAIS-III and WIAT exemplify that fact: V-IQ correlations range from .70 to .81 with the WIAT Composites, whereas P-IQ correlations range from .53 to .69 with the WIAT. Data from the WAIS-III indexes also mirror the IQ data. In WAIS-R studies (e.g., Ryan & Rosenberg, 1983; Spruill & Beck, 1986), mean correlation coefficients were .65 for Verbal and .54 for Performance. In five WAIS studies cited by Matarazzo (1972, p. 284), V-IQ correlated higher than P-IQ with high school rank (.63 versus .43) and college grade-point average (GPA) (.47 versus .24). Numerous WISC-III investigations summarized by Gridley and Roid (1998) have also shown stronger correlations between achievement ability and Verbal IQ than between Performance IQ and achievement.

In general, the use of the WAIS-III for predicting college achievement is likely to produce coefficients lower than the values in the .60s observed when standardized achievement tests are the criteria. Matarazzo (1972, p. 284), for example, cited a coefficient of .44 between WAIS FS-IQ and GPA for 335 college students with a mean IQ of 115, and Jensen (1980, p. 330) reported a median

correlation of .40 between the General Intelligence test of the General Aptitude Test Battery (GATB) and college grades in 48 different samples (comprising 5,561 students).

Even if correlation coefficients involving the WAIS-R or WAIS-III account for only 15% to 20% of the variability in college students' grades (compared to 25–50% for elementary and high school), such values nonetheless strongly support the Wechsler scales' validity for educational purposes. Correlations for college students are attenuated substantially, having nothing to do with the quality of the instrument because of (1) the restricted range of IQs found in highly selected samples, (2) the questionable reliability and validity of the GPA criterion (it, too, is usually restricted to a 5-point scale from A to F, and college grading systems fluctuate notoriously from instructor to instructor), and (3) the increasing role played by nonintellective factors such as motivation and study habits.

Relationship of IQ to Education

For children's intelligence tests, correlations between IQ and school achievement are among the best evidences of validity, but those coefficients are less valuable for adult tests. The best arguments for the validity of an adult test are the relationships between IQ and formal education and between IQ and occupational level (a variable that correlates substantially with years of schooling; Kaufman, 1990). Success in school is a key task of children and adolescents; life accomplishments are the goals of an adult.

Logically, people who score higher on a so-called intelligence test should advance higher within the formal education hierarchy and should assume positions within the more prestigious occupations. Which is cause and which is effect is not relevant to this point. Perhaps individuals score higher on IQ tests because of what they learn in school; perhaps they proceed to higher levels of education because they are smart to begin with; or perhaps these two variables combine in some way. In any case, a strong relationship between education and IQ supports the construct that underlies tests that purport to measure intelligence.

This relationship is explored in depth for the WAIS-III in Chapter 4, and again in Chapter 8 regarding V-P differences and brain damage. The present discussion gives only an overview of the relationship between years of schooling and WAIS-III scores in order to illustrate the overwhelming validity support for the WAIS-III when educational attainment is the criterion.

Educational data that are available for the WAIS-III Full Scale IQ are age-corrected *z* scores, predicted by education; these data were kindly provided by Heaton, Manly, Taylor, and Tulsky (personal communication, September, 2000), with the permission of The Psychological Corporation, and are discussed more fully in Chapter 4. Briefly, mean Full Scale IQs for 16- to 89-year-olds with different formal education levels ranged from 80.5 for individuals with 0–7 years of schooling to 116.8 for those with 17 or more years.

The two extreme educational groups differ by about 36 points, more than two standard deviations! These differences tend to be larger for Verbal than Performance subtests, but they are nonetheless substantial even for tasks like Block Design or Digit Symbol-Coding that are not specifically taught in the classroom. The mean scaled-score differences for those with 17 or more years of schooling versus those with 7 or less years of schooling (for ages 20 to 89) on two selected WAIS-III subtests, one closely related to the specific content taught in school (Vocabulary) and one unrelated to curriculum (Block Design), are 6.60 and 4.47. Specifically, on Vocabulary the mean scaled score for those with 17+ years of schooling was 13.33, whereas it was only 6.83 for those with seven or fewer years of schooling. In contrast, on Block Design mean scaled score for those with 17+ years of schooling was 11.92, yet it was only 7.45 for those with

seven or fewer years of schooling. Thus, the very highly educated adults scored 2.2 *SD* higher than relatively uneducated adults on Vocabulary and 1.5 *SD* higher on Block Design.

These data show that relatively uneducated people perform poorly on both school-related and school-unrelated tasks, and that both types of tests are substantially related to formal education. As indicated, however, highly educated adults have a greater advantage on crystallized than on fluid tasks (i.e., on Information or Vocabulary than on Block Design). Data from the Fels Longitudinal Study (McCall, 1977) reveal that childhood IQs correlate about .50 (±.10) with both adult educational and occupational attainment, stabilizing at that relatively high level at ages 7 to 8 for males and females.

The strong relationship between IQ and formal education should not obscure the considerable *variability* of IQs earned by individuals with the same educational attainment. Fluctuations in WAIS-R IQ by education level were shown by Reynolds et al. (1987), and are presented in Table 4.5. These results indicate that each level of educational attainment is accompanied by a wide range of Full Scale IQs. For example, individuals with some college education have a higher mean IQ by about 11 points than those with some high school, but their IQ ranges are fairly similar: 76–139 for those with 13–15 years of schooling compared to 59–146 for those with 9–11 years of schooling.

IQ and Occupation

For ages 20 to 54, WAIS-R data provide additional validation evidence for Wechsler's IQs by examining mean scores earned by adults actively engaged in different levels of occupation (Reynolds et al., 1987). Adolescents have been eliminated from consideration because occupational data are based on their parents' occupation, and the 55–74-year-olds have been eliminated because two thirds are categorized as "Not in Labor Force."

Occupational data are treated in depth in Chapter 4, and are summarized here to illustrate the validity of the IQ construct. Mean Full Scale IQs are shown in Table 1.3 for five categories of occupation, listed in order of the average educational level (from high to low) that typifies each category. These values range from about 87 for unskilled workers to 112 for professionals and technical workers.

The 25-point difference between professionals and unskilled workers, combined with the educational data, gives strong support to the construct underlying Wechsler's Full Scale IQs for adult samples; occupational and educational data presented in Chapter 4 give substantial validity support for the separate Verbal and Performance IQs as well.

In general, the relationship between occupation, education, and WAIS-R IQs for persons 75 years and older was similar to that found by Reynolds et al. (1987) for persons 16 to 74 (Ryan, Paolo, & Dunn, 1995). When past occupation was measured in an elderly sample (ages 75+), individuals who were retired professionals or managers earned WAIS-R Full Scale IQs that were 15.78 points higher than those who were retired

TABLE 1.3 Mean WAIS-R Full Scale IQs for 20- to 54-year-olds employed in different levels of occupation

Occupational Group	Mean WAIS-R Full Scale IQ
1. Professional and technical	112.4
2. Managers and administrators, clerical workers, and sales workers	103.6
3. Skilled workers (craftsmen and foremen)	100.7
4. Semiskilled workers (operatives, service workers—including private household—farmers, and farm managers)	92.3
5. Unskilled workers (laborers, farm laborers, farm foremen)	87.1

NOTE: Data are from Reynolds et al. (1987).

laborers or operatives. Education was also an important variable in this elderly sample, as it accounted for 30% to 43% of the variance in the WAIS-R IQs. Similar to results with younger adults, there were substantial differences (17.05 points) between those with the most education (12 or more years) and those with less formal schooling (0 to 11 years). As the relationship between education and occupation is known to be quite strong, Ryan et al. (1995) performed analyses to determine whether preretirement occupation would explain an additional amount of variance in IQ over and above age and education. Occupation did, in fact, contribute significantly to all WAIS-R IQs, explaining an additional 3% to 6% of the variance in the Verbal, Performance, and Full Scale IQs, beyond that of age and education.

When IQs are provided for specific jobs instead of general categories, even wider discrepancies emerge between diverse occupations. For example, Matarazzo (1972, pp. 178–180) cites numerous studies and his own considerable clinical experience to show that physicians, medical students, dentists, university professors, psychiatrists, executives in industry, scientists, and attorneys have consistently averaged IQs of 125 on the Wechsler-Bellevue and WAIS. In a study of 35 medical students, Mitchell, Grandy, and Lupo (1986) reported mean Full Scale IQs in the same range on both the WAIS (124.5) and WAIS-R (120.8).

The wide range of mean scores by people in different occupations is further illustrated by a comprehensive (N = 39,600) 1970 U.S. Department of Labor study cited by Jensen (1980, pp. 341–342). Mean IQs on the GATB General Intelligence scale were provided for 444 specific occupations, and ranged from 55 for Tomato Peeler to 143 for Mathematician. Although the GATB General Intelligence score correlated .89 with the WAIS (Jensen, 1980), the two scales have different standard deviations. When the GATB scores for Tomato Peelers and Mathematicians are converted to the Wechsler metric, the means become 66 and 132, respectively. This discrepancy is not as impressive as the 88-point difference on the GATB scale (mean of 100, *SD* of 20), but it nonetheless provides additional evidence of the IQ construct's validity.

Figure 1.1, adapted from Matarazzo (1972, p. 178) and Jensen (1980, p. 113) and modified based on WAIS-R data reported by Reynolds et al. (1987), presents graphically the educational or occupational referents of different IQ levels. However, these values are just the averages for different jobs or educational accomplishments. As Matarazzo (1972) and Jensen (1980) stress, adults in each occupation or educational category vary considerably in IQ range. Table 4.5 presents pertinent data that reveal the fairly wide range of IQs for individuals from the same occupational category (as mentioned previously, this same table shows the wide IQ ranges for people with different levels of education). For occupational groups, the range is relatively small for people employed in routine, menial jobs usually reserved for the mentally retarded, but substantial IQ ranges characterize members in jobs as diverse as physicians or policemen or even unskilled construction workers.

The strong relationship depicted here between IQ and occupation may be an artifact of the even stronger relationship described previously between IQ and educational attainment. Occupation and education correlate substantially, particularly because advanced formal education is frequently a prerequisite for many high-prestige occupations. Gottfredson and Brown (1981) observed an interesting age-related finding in the occupation–education relationship in their large-scale longitudinal study. Occupational status correlated a modest .17–.20 with years of schooling at ages 18–20 years, but increased at age 22 (.45) and age 24 (.60) before plateauing in the mid-.60s for 26- and 28-year-olds. Gottfredson and Brown interpreted these age-related findings as a function of the facts that (1) the later entrants into the work force are brighter and better educated, and (2) among those already employed, the smarter and more educated adults advance from low-level to high-level positions.

Crawford and Allan (1997), studying a group of 200 adults ages 16 to 83 (M = 44.3 years) from

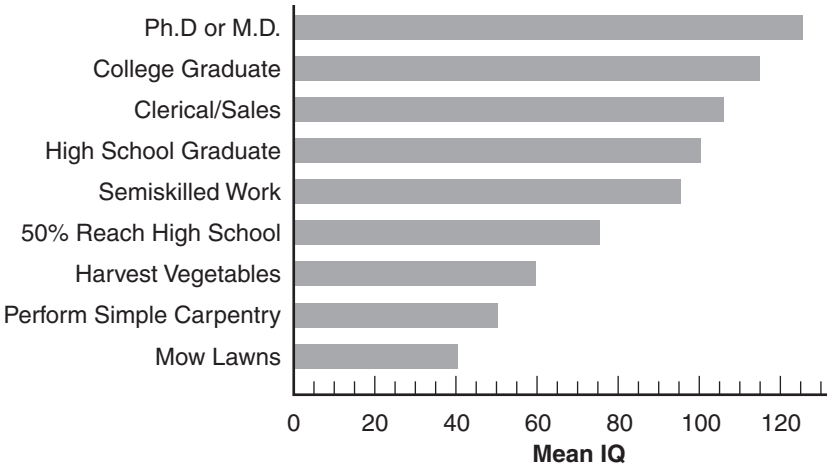

FIGURE 1.1

Mean Wechsler adult IQs that correspond to different educational and occupational accomplishments (based on data on Table 7–3 of Matarazzo, 1972, p. 178; data in Table 4.5 in Jensen, 1980, p. 113; WAIS-R standardization data reported by Reynolds et al., 1987).

the United Kingdom, found that occupation was a slightly stronger predictor of WAIS-R IQ than was education. The correlations for this sample between occupation and FS-IQ (.65), education and FS-IQ (.58), and education and occupation (.65) are within the ranges of what has been previously reported. However, Crawford and Allan found that occupation was the single best predictor of IQ for all three scales in a stepwise regression. Occupational classification accounted for 42%, 43%, and 25% of the variance in FS-IQ, V-IQ, and P-IQ, respectively. Education and age significantly increased the variance predicted, with final models predicting 32% to 53% of the variance in the IQs. Thus, it appears that occupation in and of itself is an important demographic variable contributing to IQ.

Regardless, years of schooling "is the single most important determinant of occupational status in United States society" (Brody, 1985, p. 361). Brody states further that the results of path analysis in several studies indicate that IQ has "a large influence on educational attainment and *relatively little indirect influence on occupational status*" (pp. 361–362, italics ours)—that is, separate from the IQ–education relationship.

Prediction of Job Performance

Average correlations between general intelligence and job proficiency are traditionally in the .20s (Ghiselli, 1966, 1973). However, because the predictors and criteria are typically restricted in variability due to selection factors and other practical limitations of test validation in industrial settings, some have argued that such coefficients require statistical correction to reflect more accurately the "true" relationship between IQ and job success (Hunter & Hunter, 1984). For the purpose here (i.e., to determine the validity of the *theoretical* construct underlying intelligence tests), the corrected values seem more appropriate.

In an ambitious meta-analysis of hundreds of studies relating intelligence to job performance, Hunter (1986) concluded that "general cognitive

ability has high validity predicting performance ratings and training success in all jobs" (p. 359). He organized data from three major sources, correcting coefficients for restriction of range in all cases, and for attenuation (imperfect test reliability) in the first two sets of studies: (1) Ghiselli's lifework, involving several summaries of a quarter-century's worth of validity studies in industry on the prediction of job proficiency and success in training programs; (2) 515 validation studies conducted by the U.S. Employment Service with the GATB, 425 on job proficiency (N = 32,124) and 90 on training success (N = 6,496); and (3) U.S. military studies of training success in mechanical, clerical, electronic, and general technical fields (828 studies totalling 472,539 subjects).

Coefficients of correlation between intelligence and job proficiency (performance ratings) were consistently higher for complex jobs than for those demanding less complexity. The Ghiselli studies produced substantial corrected correlations for the complex jobs of manager (.53), clerk (.54), and salesperson (.61). Coefficients in the mid-.40s were obtained for jobs of medium complexity (e.g., crafts and trades), while values in the high .20s and .30s were typical of low complexity jobs like vehicle operator. Similar averages emerged when Hunter (1986) grouped the U.S. Employment Service studies by complexity: high complexity (r = .58), medium (.51), and low (.40). Gottfredson (1997) suggested that general intelligence (*g*) has pervasive utility in work settings because it is related to one's ability to deal with cognitive complexity. She noted that the more complex a work task, the greater the advantages that higher *g* confers in performing it well.

Intelligence correlated even more impressively with success in training than it did with job performance. Further, the coefficients obtained for various training programs were about equally good, regardless of job complexity. The average corrected coefficient for the 828 studies of training success conducted by the U.S. military was .62, with values hovering around that overall value for each of the four job families (i.e., mechanical, clerical, electronic, and general technical). Coefficients from the Ghiselli summaries ranged from .37 (vehicle operator) to .87 (protective professions) with a median correlation of .65 across seven categories of jobs. The 90 training studies carried out by the U.S. Employment Service yielded average values of .50 to .65 (median = .56) for jobs grouped into four categories.

Hunter showed further that validity coefficients are even higher when objective work samples of job performance are used instead of subjective supervisor ratings. Based on a handful of particularly well-designed investigations that used objective criteria to evaluate job proficiency, corrected correlations were .75 in civilian data and .53 in military data.

In a more recent synthesis of the vocational data, Schmidt and Hunter (1998) reviewed 85 years of research in personnel selection, focusing on the results of the best meta-analyses, including much of the data reviewed in the preceding paragraphs. They concluded once again that IQ (referred to as general mental ability or GMA) had strong validity, and that the validity could be increased substantially when other predictors are considered as well: .63 (GMA + work sample or GMA + structured interview) or .65 (GMA + integrity test). Based on their review, Schmidt and Hunter concluded: (1) "of all procedures that can be used for all jobs, whether entry level or advanced, [GMA] has the highest validity and lowest application cost" (p. 264); (2) "the research evidence for the validity of GMA measures for predicting job performance is stronger than that for any other measure" (p. 264); and (3) "GMA has been shown to be the best available predictor of job-related learning" (p. 264).

Jensen's (1980) analysis of some of the same data summarized by Hunter (1986) presents a more sobering view of the ability of intelligence tests to predict job performance and training success. Coefficients reported by Hunter were corrected for restriction of range and, usually, for attenuation as well; these corrections inflate the correlations by estimating their magnitude in "what-if" situations. The correction for attenuation (test unreliability)

is particularly questionable, however, because, by definition, tests are not perfectly reliable. Jensen (1980, pp. 347–350) notes that Ghiselli's actual coefficients were in the .20 to .25 range, on the average, and that the median coefficient for the GATB General Intelligence score for 537 U.S. Employment Service studies was .27.

Similarly, Jensen demonstrates that correlations are greater for more complexions but that the values for jobs with high complexity are in the .35 to .47 range. Jensen also notes that the average correlation between IQ and success in training programs is close to .50, not the values of about .60 reported by Hunter. These criticisms apply as well to the more recent review by Schmidt and Hunter (1998).

Data from both Hunter (1986) and Jensen (1980) support the IQ construct as reasonably valid in its role as predictor of job success, although the claims made by Hunter may be exaggerated by his incautious and, perhaps, overzealous correction of obtained coefficients. From a theoretical perspective, the data set evaluated by Hunter and Schmidt and Hunter (1998) give excellent support of the construct validity of IQ in vocational settings. In a practical sense, however, the obtained correlations are often the most pertinent. In all instances, readers are wise to heed the cautions of two expert statisticians and psychometricians, Lloyd Humphreys and Robert Linn, regarding Hunter's correction procedures. Humphreys (1986), in his commentary on Hunter's article (and other papers as well) in a special issue of the *Journal of Vocational Behavior*, wrote, "Given the heterogeneity among the many studies to be aggregated, corrections for measurement error and restriction of range of talent are rough estimates at best" (p. 427). In a similar commentary, Linn (1986) asserted that "adjustments for range restriction and attenuation are nontrivial[;]… correlations that are changed dramatically by adjustments should always be viewed with caution" (pp. 440–441).

Although IQ seems to be a valid predictor of job performance, the general findings from this line of research indicate that a relatively small amount of the variance in job performance is accounted for by IQ. At worst, the average validity coefficient between measures of cognitive ability and measures of cognitive ability is .20 (Ghiselli, 1966; Wigdor & Garner, 1982), accounting for only 4% of the variance, and, at best, the average validity coefficient is about .5 (Hunter & Hunter, 1984; Schmidt & Hunter, 1998), accounting for 25% of the variance in job performance. As Sternberg, Wagner, Williams, and Horvath (1995) point out, these values leave at least three-quarters of the variance unexplained. Sternberg et al. suggest that practical intelligence (common sense) is a variable that may contribute to the prediction of job performance, above and beyond what traditional IQ contributes. Practical intelligence, or "tacit knowledge," has only a small relationship to general intelligence (Sternberg et al., 1995). When tasks of tacit knowledge are used to predict managerial performance, tacit knowledge accounts for substantial and significant increases in variance above and beyond IQ (Wagner & Sternberg, 1990). Using measures of traditional intelligence in conjunction with measures of tacit knowledge may more effectively predict job performance than reliance on one of these measures alone (Sternberg et al., 1995), although reliable and construct-valid measures of tacit knowledge are not yet available.

THE INTELLIGENT TESTING PHILOSOPHY

One's philosophy regarding the interpretation of individually administered clinical tests should be an intelligent one. The approach we will be describing has been spelled out in detail for various Wechsler tests (Kaufman, 1979a, 1994a; Kaufman & Lichtenberger, 1999, 2000), applied to the K-ABC (Kamphaus & Reynolds, 1987), and applied to a variety of other clinical and neuropsychological instruments (Reynolds & Fletcher-Janzen, 1989). Consequently, our goal here is only to summarize the assumptions underlying the approach and the basic methodology that characterizes it. The essential method is the

same, whether applied to tests for children, adolescents, or adults. Intelligent testing rests on five assumptions, discussed in the sections below:

1. IQ tasks measure what the individual has learned.
2. IQ tasks are samples of behavior and are not exhaustive.
3. IQ tests like the WAIS-III, KAIT, and WJ III assess mental functioning under fixed experimental conditions.
4. IQ tests are optimally useful when they are interpreted from an information-processing model.
5. Hypotheses generated from IQ test profiles should be supported with data from multiple sources.

IQ Tasks Measure What the Individual Has Learned

This concept comes directly from Wesman's (1968) introduction of the intelligent testing approach. The content of all tasks, whether verbal or nonverbal, is learned within a culture. The learning may take place formally in the school, casually in the home, or incidentally through everyday life. As a measure of past learning, the IQ test is best thought of as a kind of achievement test, not as a simple measure of aptitude. Like the SAT, IQ tests assess "*developed abilities*, broadly applicable intellectual skills and knowledge that develop slowly over time through the individual's experiences both in and out of school…[that are] not tied to the content of any specific course or field of study" (Anastasi, 1988, p. 330).

The interaction between learning potential and availability of learning experiences is too complex to ponder for any given person, making the whole genetics–environment issue of theoretical value, but impractical and irrelevant for the interpretation of that person's test profile. Even the sophisticated scientific challenges to the IQ construct issued by Lezak (1988a) and Siegel (1999) or the emotional, less informed indictments of IQ tests handed out by members of the public, become almost a side issue when the tests are viewed and interpreted simply as measures of accomplishment. The term *achievement* implies a societal responsibility to upgrade the level of those who have not attained it; the term *aptitude* implies something inborn and personal and can justify a withdrawal of educational resources (Flaugher, 1978).

Issues of heredity versus environment and the validity of the IQ construct are meaningful for understanding the multifaceted intelligence construct; the accumulating research helps test developers, practitioners, and theoreticians appreciate the foundation of the tests used to measure intelligence; and the IQ tests, as vehicles for the research, are essential sources of group data for use in scientific study of these topics. But all of the controversy loses meaning for each specific person referred for evaluation when the clinician administers an IQ test to study and interpret just what the person has or has not learned and to help answer the practical referral questions.

IQ Tasks Are Samples of Behavior and Are Not Exhaustive

The individual Wechsler subtests, or the subtests that compose the KAIT or WJ III, do not reflect the essential ingredients of intelligence whose mastery implies some type of ultimate life achievement. They, like tasks developed by Binet and other test constructors, are more or less arbitrary samples of behavior. Teaching people how to solve similarities, assemble blocks to match abstract designs, or repeat digits backward will not make them smarter in any broad or generalizable way. What we are able to infer from the person's success on the tasks and style of responding to them is important; the specific, unique aspect of intellect that each subtest measures is of minimal consequence.

Limitations in the selection of tasks necessarily mean that one should be cautious in generalizing the results to circumstances that are removed

from the one-on-one assessment of a finite number of skills and processing strategies. Intelligence tests should, therefore, be routinely supplemented by other formal and informal measures of cognitive, clinical, and neuropsychological functioning to facilitate the assessment of mental functioning as part of psychodiagnosis. The global IQ on any test, no matter how comprehensive, does not equal a person's total capacity for intellectual accomplishment.

IQ Tests Like the WAIS-III, KAIT, and WJ III Assess Mental Functioning under Fixed Experimental Conditions

Standardized administration and scoring means conducting an experiment with $N = 1$ every time an examiner tests someone on an intelligence test. For the results of this experiment to be meaningful, the experimenter–examiner must adhere precisely to the wording in the manual, give appropriate probes as defined in the instructions, time each relevant response diligently, and score each item exactly the way comparable responses were scored during the normative procedure. Following these rules prevents examiners from applying a flexible clinical investigatory procedure during the administration (like Piaget's semistructured *méthode clinique*), from teaching the task or giving feedback to a person who urgently desires this intervention, or from cleverly dislodging from the crevices of a person's brain his or her maximum response to each test item.

It is necessary to be an exceptional clinician to establish and maintain rapport and to weave the standardized administration into a natural, pleasant interchange between examiner and subject. Clinical skills are also essential when observing and interpreting a person's myriad behaviors during the examination and during interpretation of all available information and data when interpreting the profile of test scores. But it is vital for an examiner to follow the standardized procedures to the letter while administering the test; otherwise, the standard scores yielded for the person will be invalid and meaningless. To violate the rules is to negate the value of the meticulous set of norms obtained under experimental conditions by most major test-publishing companies for their tests.

The testing situation has a certain built-in artificiality by virtue of the stopwatch, the precise words to be spoken, and the recording of almost everything spoken by the examinee. A person with excellent visual–spatial and manipulative skills might perform slowly and ineffectively on Object Assembly because of anxiety caused by the time pressure; or a person with an impressive store of general knowledge and a good commonsense understanding of social situations may fail several Information and Comprehension items because of failure to understand some of the questions. It is tempting to give credit to a puzzle solved "just 2 or 3 seconds overtime" or to simplify the wording of a question that the person "certainly knows the answer to." But the good examiner will resist these temptations, knowing that the people in the reference group did not receive such help. Testing the limits on a subtest can often give valuable insight into the reasons for failure or confusion, so long as this flexible, supplemental testing occurs *after* the score has been recorded under appropriate conditions.

In an experiment, the empirical results are of limited value until they are interpreted and discussed in the context of pertinent research and theory by a knowledgeable researcher. By the same token, the empirical outcomes of an IQ test are often meaningless until put into context by the examiner. That is the time for a clinician's acumen and flexibility to be displayed.

IQ Tests Are Optimally Useful When They Are Interpreted from an Information-Processing Model

One of the examiner's jobs in an assessment is to identify specific areas of dysfunction. One model

that has been particularly useful to clinicians in delineating areas of dysfunction is the information-processing model (Silver, 1993). The information-processing model is applicable to the learning process in general and any given cognitive task. The four components of the model are shown in Figure 1.2.

The information-processing model can be used as a conceptual framework for interpreting IQs, Factor Indexes, and scaled scores that extends beyond the specific areas obtained (Kaufman, 1994a). With the help of this model, scores can be reorganized and translated into fundamental areas of strength and weakness within the cognitive profile.

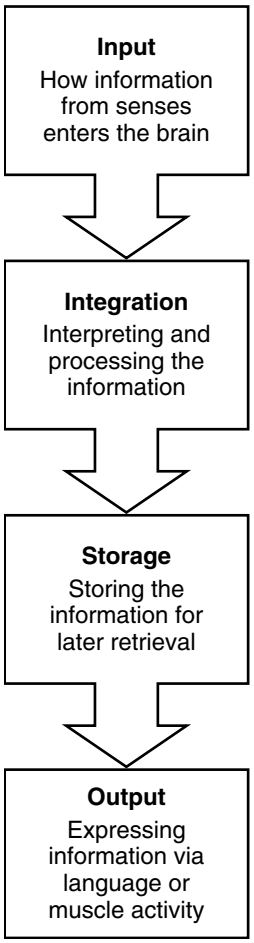

FIGURE 1.2
Information-Processing Model

Generally, the input of WAIS-III Verbal subtests tends to be auditory, while that of the Performance subtests is visual. Although it is perhaps simplistic to reduce the input of WAIS-III subtests into a verbal–visual dichotomy, in a rudimentary way, all subtests can be categorized as having one or the other types of input. For the KAIT and WJ III, there is no simple relationship between scales and modalities. For example, the KAIT Logical Steps subtest is on the Fluid Scale (akin to Performance Scale), but it requires good verbal comprehension for success.

Hypotheses Generated from IQ Test Profiles Should Be Supported with Data from Multiple Sources

Test score profiles are optimally meaningful when interpreted in the context of known background information, observed behaviors, and approach to each problem-solving task. Virtually any examiner can deduce that WAIS-III Verbal IQ, KAIT Crystallized IQ, or WJ III Comprehension-Knowledge standard score is not a very good measure of the crystallized intelligence of a person raised in a foreign culture, a person who understands Spanish or Vietnamese far better than English, or a person with a hearing impairment, and that Wechsler's Performance IQ or KAIT Memory for Block Designs does not measure nonverbal intelligence very well for a person with crippling arthritis or a visual handicap. The goal of the intelligent tester is to deduce when one or more subtests may be an invalid measure of a person's intellectual functioning for more subtle reasons: distractibility, poor arithmetic achievement in school, subcultural differences in language or custom, emotional content of the items, suspected or known lesions in specific regions of the brain, fatigue, boredom, extreme shyness, bizarre thought processes, inconsistent effort, and the like.

Being a great detective, able to follow up leads and hunches about peaks and valleys in a profile, is the hallmark of an intelligent tester. Such a tester will integrate IQ test profiles with background information, clinical observations of behaviors, and other tests administered in order to more fully understand the examinee's profile.

Tying Together the Tenets of Intelligent Testing

The principles discussed in the preceding sections direct our attention to one important point: the focus of any assessment is the person being assessed, not the test. Many psychological reports stress what the scales or subtests measure instead of what aspects of the person are particularly well developed or in need of improvement; many reports are so number-oriented that the reader loses sight of the person's uniqueness. Current IQ tests for adolescents and adults enable psychologists to better understand a person's cognitive functioning, but other facets of an individual are also revealed during an assessment and should be fully integrated to represent that person as a whole. Although the section of an assessment report that systematically reports and interprets the IQs, cluster scores, and subtest scores is valuable, the behavioral observations section of a case report is often more revealing, and ultimately of more value, if it helps to explain how or why examinees arrived at the scores that they did. The content of the responses and the person's style of responding to various types of tasks can be more important as a determiner of developmental level and intellectual maturity than the scores assigned to the items or tasks.

When several tests are administered to a person (intelligence, language, achievement, personality, visual–motor), the results must be integrated from one test battery to the other. Intelligent testing does not apply only to the interpretation of intelligence tests. The examiner's main role is to generate hypotheses that pertain mostly to assets and deficits within the information-processing model, and then confirm or deny these hypotheses by exploring multiple sources of evidence. This integrative, flexible, clinical–empirical methodology and philosophy, as outlined in the preceding tenets, represents the approach taken in this book for the interpretation of the WAIS-III, KAIT, WJ III, and other tests for adolescents and adults. The guidelines for interpreting IQ test profiles and the illustrative case reports throughout this book rest solidly on the intelligent testing framework.

SUMMARY

This chapter first delineates the goal of this book to serve as a text on individual, clinical assessment of intelligence and then outlines the five sections that make up the book: (1) introduction to the assessment of adolescent and adult intelligence; (2) individual differences on age, socioeconomic status, and other key variables; (3) integration and application of WAIS-III research; (4) interpretation of the WAIS-III profile; and (5) additional measures of adolescent and adult IQ. The remainder of the chapter sketches a brief history of the IQ, gives survey data of test usage, presents evidence for the validity of the IQ construct, and introduces the intelligent testing philosophy.

Alfred Binet was truly the pioneer of IQ testing. His concepts and approach dominated the field for years, and Terman's adaptation, the Stanford-Binet, became the criterion of intelligence in the United States. The nonverbal Performance tests developed during World War I to assess non-English-speaking recruits, low-functioning individuals, and suspected malingerers joined with the verbal-oriented Binet tradition to pave the way for David Wechsler's creative contribution of a dual Verbal and Performance approach to intellectual assessment. Wechsler went on to become a proponent of clinical, not just psychometric, assessment. The need for multiscore measurement that accompanied the learning disabilities movement in the 1960s catapulted the Wechsler series

of scales ahead of the Binet as the most popular intelligence test.

The results of recent surveys on test usage show that the Wechsler tests still are strongly popular in clinical psychology, neuropsychology, forensic psychology, school psychology, hospital settings, and outpatient clinics. The percentage of clinical time spent conducting assessments varies across specialties within psychology (e.g., clinical, school, neuropsychology), with fluctuations depending on the type of assessment necessary. The inconsistency between the amount of time typically allowed to be reimbursed for assessment services and the actual amount of time spent in assessment-related services may affect the types and numbers of assessments performed by clinicians. Notwithstanding the fees and reimbursement issues, the popularity of the Wechsler scales and the primary reasons for assessing adults remain unchanged. A strong need for tools to assess cognitive capabilities and obtain related clinical information in adults will undoubtedly keep the WAIS-III in its place at the top of the heap of assessment measures.

The validity of the IQ construct was explored for adolescents and adults. Empirical evidence supports the IQ as a good predictor of academic achievement for college students and clinical referrals, and as a strong correlate of educational attainment; IQ also relates substantially to the status of an occupation and correlates significantly with job performance, especially with success in training programs. In general, validity evidence is provided for both verbal and nonverbal measures of intelligence.

The intelligent testing philosophy, which considers the clinician's expertise and training to be more important an aspect of the assessment process than the specific instruments administered or the scores obtained, embodies the following principles: (1) IQ tasks measure what the individual has learned; (2) IQ tasks are samples of behavior and are not exhaustive; (3) IQ tests like the WAIS-III, KAIT, and WJ III assess mental functioning under fixed experimental conditions; (4) IQ tests are optimally useful when they are interpreted from an information-processing model; and (5) hypotheses generated from IQ test profiles should be supported with data from multiple sources.