CHAPTER 9

Intelligence and Its Measurement

For as long as there has been a discipline of psychology, psychologists have had differing definitions of intelligence and how best to measure it.

In this chapter we look at the varied ways intelligence has been defined and we survey various approaches to its measurement. Along the way, we will address some of the major issues surrounding how and why intelligence is measured. Before all of that, however, let's consider a question that logically precedes consideration of intelligence measurement issues.

What Is Intelligence?

We may define **intelligence** as a multifaceted capacity that manifests itself in different ways across the life span. In general, intelligence includes the abilities to:

- acquire and apply knowledge
- reason logically
- plan effectively
- infer perceptively
- make sound judgments and solve problems
- grasp and visualize concepts
- pay attention
- be intuitive
- find the right words and thoughts with facility
- cope with, adjust to, and make the most of new situations

As broad as these descriptions are, they are far from the "last word" on the matter. Rather, think of these descriptions as a point of departure for reflection on the meaning of a most intriguing term—one that, as we will see, is paradoxically both simple and complex.

JUST THINK . . .

How do *you* define intelligence?

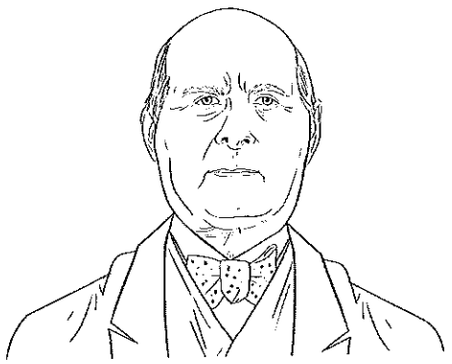
Young children may define "intelligence" in terms that emphasize positive interpersonal skills (such as acting nice, or being helpful or polite). For older children, more emphasis in the definition will typically be placed on academic skills, such as reading well (Yussen & Kane, 1980). For psychologists, universal agreement as to how intelligence should be defined has been a bit more elusive (Neisser et al., 1996). While most may view it as an abstract construct, some see it as having a more tangible existence at the level of the neuron (Grazioplene, 2015; Kreifelts et al., 2010). There are even highly respected voices in the profession who have questioned the very usefulness of the construct (Das, 2015).

Controversy surrounding the definition of intelligence is not new. In a symposium published in the *Journal of Educational Psychology* in 1921, seventeen of the country's leading psychologists addressed the following questions: (1) *What is intelligence?* (2) *How can it best be measured in group tests?* and (3) *What should be the next steps in the research?* No two psychologists agreed (Thorndike et al., 1921). Six years later, Spearman (1927, p. 14) would reflect: "In truth, intelligence has become . . . a word with so many meanings that finally it has none." And decades after the symposium was first held, Wesman (1968, p. 267) concluded that there appeared to be "no more general agreement as to the nature of intelligence or the most valid means of measuring intelligence today than was the case 50 years ago."

JUST THINK . . .

Must professionals agree on a definition of intelligence?

As Neisser (1979) observed, although the *Journal* felt that the symposium would generate vigorous discussion, it generated more heat than light and led to a general increase in exasperation with discussion on the subject. Symptomatic of that exasperation was an unfortunate statement by an experimental psychologist and historian of psychology, Edwin G. Boring. Boring (1923, p. 5), who was not a psychometrician, attempted to quell the argument by pronouncing that "intelligence is what the tests test." Although such a view is not entirely devoid of merit (see Neisser, 1979, p. 225), it is an unsatisfactory, incomplete, and circular definition. In what follows we discuss the thoughts of other behavioral scientists through history and up to contemporary times on the meaning and measurement of intelligence (see Figure 9–1).



Francis Galton (1822–1911)

Figure 9–1
"Intelligence" is. . .

Galton (1883) believed that the most intelligent persons were those equipped with the best sensory abilities. This position was intuitively appealing because, as Galton observed, "The only information that reaches us concerning outward events appears to pass through the avenues of our senses; and the more perceptive the senses are of difference, the larger is the field upon which our judgment and intelligence can act" (p. 27). Following this logic, tests of visual acuity or hearing ability are, in a sense, tests of intelligence. Galton attempted to measure this sort of intelligence in many of the sensorimotor and other perception-related tests he devised. Among his many other accomplishments, Sir Francis Galton is remembered as the first person to publish on the heritability of intelligence, thus anticipating later nature-nurture debates (McGue, 1997).

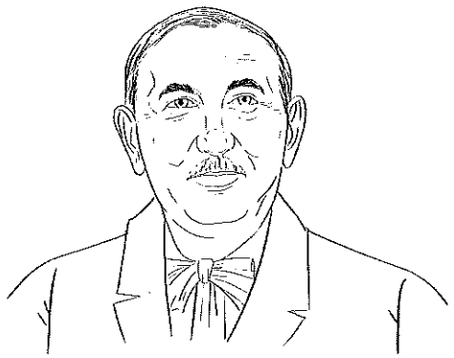
Alfred Binet (1857–1911)

In papers critical of Galton's approach to intellectual assessment, Binet and a colleague called for more complex measurements of intellectual ability (Binet & Henri, 1895a, 1895b, 1895c). Galton had viewed intelligence as a number of distinct processes or abilities that could be assessed only by separate tests. In contrast, Binet argued that when one solves a particular problem, the abilities used cannot be separated because they interact to produce the solution. For example, memory and concentration interact when a subject is asked to repeat digits presented orally. When analyzing a testtaker's response to such a task, it is difficult to determine the relative contribution of memory and concentration to the successful solution. This difficulty in determining the relative contribution of distinct abilities is the reason Binet called for more complex measurements of intelligence. Although Binet never explicitly defined intelligence, he discussed its components in terms of reasoning, judgment, memory, and abstraction (Varon, 1936).

In Wechsler's (1958, p. 7) definition of intelligence, there is an explicit reference to an "aggregate" or "global" capacity:

> *Intelligence, operationally defined, is the aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with his environment. It is aggregate or global because it is composed of elements or abilities which, though not entirely independent, are qualitatively differentiable. By measurement of these abilities, we ultimately evaluate intelligence. But intelligence is not identical with the mere sum of these abilities, however inclusive. . . . The only way we can evaluate it quantitatively is by the measurement of the various aspects of these abilities.*

Elsewhere Wechsler added that there are nonintellective factors that must be taken into account when assessing intelligence (Kaufman, 1990). Included among those factors are "capabilities more of the nature of conative, affective, or personality traits [that] include such traits as drive, persistence, and goal awareness [as well as] an individual's potential to perceive and respond to social, moral and aesthetic values" (Wechsler, 1975, p. 136). Ultimately, however, Wechsler was of the opinion that the best way to measure this global ability was by measuring aspects of several "qualitatively differentiable" abilities. Wechsler (1974) wrote of two such "differentiable" abilities, which he conceived as being primarily verbal- or performance-based in nature. Today, the conceptualization of intelligence in terms of these two factors—a verbal factor and a performance factor—is largely a matter of historical interest.

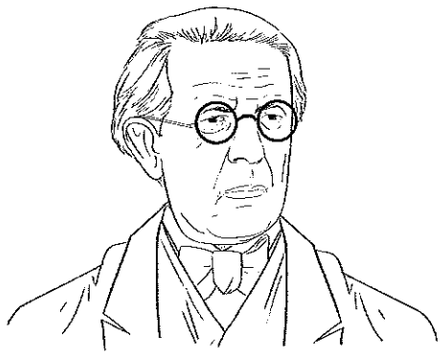


David Wechsler (1896–1981)

JUST THINK . . .

What is the role of personality in measured intelligence?

For Piaget (1954, 1971), intelligence may be conceived of as a kind of evolving biological adaptation to the outside world. As cognitive skills are gained, adaptation (at a symbolic level) increases, and mental trial and error replaces physical trial and error. Yet, according to Piaget, the process of cognitive development occurs neither solely through maturation nor solely through learning. He believed that, as a consequence of interaction with the environment, psychological structures become reorganized. Piaget described four stages of cognitive development through which, he theorized, all of us pass during our lifetimes. Although individuals can move through these stages at different rates and ages, he believed that their order was unchangeable. Piaget viewed the unfolding of these stages of cognitive development as the result of the interaction of biological factors and learning. Interested readers will find more about Piaget's theory of cognitive development in OOBAL-9-B1, which can be found in the Instructor Resources within Connect.



Jean Piaget (1896–1980)

Perspectives on Intelligence

A major thread running through the theories of Binet, Wechsler, and Piaget is a focus on interactionism. **Interactionism** refers to the complex concept by which heredity and environment are presumed to interact and influence the development of one's intelligence. As we will see, other theorists have focused on other aspects of intelligence. For example, Louis L. Thurstone conceived of intelligence as composed of what he termed *primary*

JUST THINK . . .

In everyday living, mental abilities tend to operate in unison rather than in isolation. How useful is it, therefore, to attempt to isolate and measure "primary mental abilities"?

mental abilities (PMAs). Thurstone (1938) developed and published the Primary Mental Abilities test, which consisted of separate tests, each designed to measure one PMA: verbal meaning, perceptual speed, reasoning, number facility, rote memory, word fluency, and spatial relations. Although the test was not widely used, this early model of multiple abilities inspired other theorists and test developers to explore various components of intelligence and ways to measure them.

In **factor-analytic theories,** the focus is squarely on identifying the ability or groups of abilities deemed to constitute intelligence. In **information-processing theories,** the focus is on identifying the specific mental processes that constitute intelligence. Prior to reading about factor-analytic theories of intelligence, some extended discussion of factor analysis may be helpful (see this chapter's *Close-Up*).

Factor-analytic theories of intelligence Factor analysis is a group of statistical techniques designed to determine the existence of underlying relationships between sets of variables, including test scores. In search of a definition of intelligence, theorists have used factor analysis to study correlations between tests measuring varied abilities presumed to reflect the underlying attribute of intelligence.

As early as 1904, the British psychologist Charles Spearman pioneered new techniques to measure intercorrelations between tests. He found that measures of intelligence tended to correlate to various degrees with each other. Spearman (1927) formalized these observations into an influential theory of general intelligence that postulated the existence of a general intellectual ability factor (denoted by an italic lowercase ***g***) that is partially tapped by all other mental abilities. This theory is sometimes referred to as a **two-factor theory of intelligence,** with *g* representing the portion of the variance that all intelligence tests have in common and the remaining portions of the variance being accounted for either by specific components (*s*), or by error components (*e*) of this general factor (Figure 9–2). Tests that

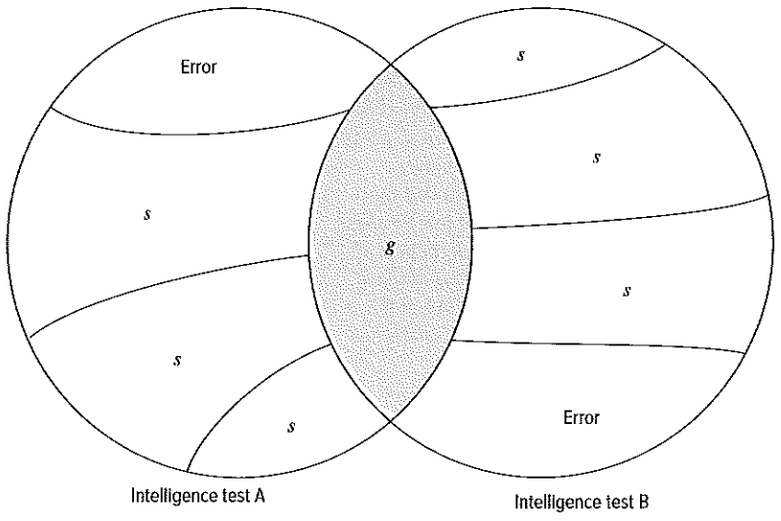


Figure 9–2
Spearman's Two-Factor Theory of Intelligence

Here, g *stands for a general intelligence factor and* s *stands for a specific factor of intelligence (specific to a single intellectual activity only).*

CLOSE-UP

Factor Analysis*

To measure characteristics of physical objects, there may be some disagreement about the best methods to use, but there is little disagreement about which dimensions are being measured. We know, for example, that we are measuring length when we use a ruler, and we know that we are measuring temperature when we use a thermometer. Such certainty is not always present in measuring psychological dimensions such as personality traits, attitudes, and cognitive abilities.

Psychologists may disagree about what to name the dimensions being measured and about the number of dimensions being measured. Consider a personality trait that one researcher refers to as *niceness*. Another researcher views *niceness* as a vague term that lumps together two related but independent traits called *friendliness* and *kindness*. Yet another researcher claims that *kindness* is too general and must be dichotomized into *kindness to friends* and *kindness to strangers*. Who is right? Is everybody right? If researchers are ever going to build on each others' findings, there needs to be some way of reaching consensus about what is being measured. Toward that end, factor analysis can be helpful.

An assumption of factor analysis is that things that co-occur tend to have a common cause. Note here that "tend to" does *not* mean "always." Fevers, sore throats, stuffy noses, coughs, and sneezes may *tend to* occur at the same time in the same person, but they do not always co-occur. When these symptoms do co-occur, they may be caused by one thing: the virus that causes the common cold. Although the virus is one thing, its manifestations are quite diverse.

In psychological assessment research, we measure a diverse set of abilities, behaviors, and symptoms and then attempt to deduce which underlying dimensions cause or account for the variations in behavior and symptoms observed in large groups of people. We measure the relations among various behaviors, symptoms, and test scores with correlation coefficients. We then use factor analysis to discover patterns of correlation coefficients that suggest the existence of underlying psychological dimensions.

All else being equal, a simple theory is better than a complicated theory. Factor analysis helps us discover the smallest number of psychological dimensions (or factors) that can account for the various behaviors, symptoms, and test scores we observe. For example, imagine that we create four different tests to measure people's knowledge of vocabulary, grammar, multiplication, and geometry. If the correlations between all of these tests were zero (or, high scorers on one test are no more likely than low scorers to score high on the other tests), then the factor analysis would suggest to us that we have measured four distinct abilities.

Of course, you probably recognize that it is most unlikely that the correlations between these tests would be zero. Therefore, imagine that the correlation between the vocabulary and grammar tests were quite high (or, high scorers on vocabulary were likely also to score high on grammar, and low scorers on vocabulary were likely to score low on grammar), and suppose also a high correlation between multiplication and geometry. Furthermore, the correlations between the verbal tests and the mathematics tests were zero. Factor analysis would suggest that we have measured not four distinct abilities but two. The researcher interpreting the results of this factor analysis would have to use his or her best judgment in deciding what to call these two abilities. In this case, it would seem reasonable to call them *language ability* and *mathematical ability*.

Now imagine that the correlations between all four tests were equally high—for example, that vocabulary was as strongly correlated with geometry as it was with grammar. In this case, factor analysis suggests that the simplest explanation for this pattern of correlations is that there is just one factor that causes all these tests to be equally correlated. We might call this factor *general academic ability*.

In reality, if you were to actually measure these four abilities, the results would not be so clear-cut. It is likely that all of the correlations would be positive and substantially above zero. It is likely that the verbal subtests would correlate more strongly with each other than with the mathematical subtests. It is likely that factor analysis would suggest that language and mathematical abilities are distinct from but not entirely independent of each other—in other words, that language abilities and mathematics abilities are substantially correlated, suggesting that a general academic (or intellectual) ability influences performance in all academic areas.

Factor analysis can help researchers decide how best to summarize large amounts of information about people by using just a few scores. For example, when we ask parents to complete questionnaires about their children's behavior problems, the questionnaires can have hundreds of items. It would take too long and would be too confusing to review every item. Factor analysis can simplify the information while minimizing the loss of detail. Here is an example of a

(continued)

*Prepared by W. Joel Schneider.

Factor Analysis *(continued)*

short questionnaire that factor analysis can be used to summarize.

> On a scale of 1 to 5, compared to other children his or her age, my child:
>
> 1. gets in fights frequently at school
> 2. is defiant to adults
> 3. is very impulsive
> 4. has stomachaches frequently
> 5. is anxious about many things
> 6. appears sad much of the time

If we give this questionnaire to a large, representative sample of parents, we can calculate the correlations between the items. Table 1 illustrates what we might find.

Note that all of the perfect 1.00 correlations in this table are used to emphasize the fact that each item correlates perfectly with itself. In the analysis of the data, the software would ignore these correlations and analyze only all of the correlations below this diagonal "line of demarcation" of 1.00 correlations.

Using the set of correlation coefficients presented in Table 1, factor analysis suggests that there are two factors measured by this behavior rating scale. The logic of factor analysis suggests that the reason items 1 through 3 have high correlations with each other is that each has a high correlation with the first factor. Similarly, Items 4 through 6 have high correlations with each other because they have high correlations with the second factor. The correlations of the items with the hypothesized factors are called *factor loadings.* The factor loadings for this hypothetical example are presented in Table 2.

Table 1
A Sample Table of Correlations

	1	2	3	4	5	6
1. gets in fights frequently at school	1.00					
2. is defiant to adults	.81	1.00				
3. is very impulsive	.79	.75	1.00			
4. has stomachaches frequently	.42	.38	.36	1.00		
5. is anxious about many things	.39	.34	.34	.77	1.00	
6. appears sad much of the time	.37	.34	.32	.77	.74	1.00

Table 2
Factor Loadings for Our Hypothetical Example

	Factor 1	Factor 2
1. gets in fights frequently at school	.91	.03
2. is defiant to adults	.88	−.01
3. is very impulsive	.86	−.01
4. has stomachaches frequently	.02	.89
5. is anxious about many things	.01	.86
6. appears sad much of the time	−.02	.87

Factor analysis tells us which items *load* on which factors, but it cannot interpret the meaning of the factors. Researchers usually look at all the items that load on a factor and use their intuition or knowledge of theory to identify what the items have in common. In this case, Factor 1 could receive any number of names, such as *Conduct Problems, Acting Out,* or *Externalizing Behaviors.* Factor 2 might also go by various names, such as *Mood Problems, Negative Affectivity,* or *Internalizing Behaviors.* Thus, the problems on this behavior rating scale can be summarized fairly efficiently with just two scores. In this example, a reduction of six scores to two scores may not seem terribly useful. In actual behavior rating scales, factor analysis can reduce the overwhelming complexity of hundreds of different behavior problems to a more manageable number of scores that help professionals more easily conceptualize individual cases.

Factor analysis also calculates the correlation among factors. If a large number of factors are identified and if there are substantial correlations among factors, then this new correlation matrix can also be factor-analyzed to obtain *second-order factors.* These factors, in turn, can be analyzed to obtain *third-order factors.* Theoretically, it is possible to have even higher-order factors, but most researchers rarely find it necessary to go beyond third-order factors. The *g* factor from intelligence test data is an example of a third-order factor that emerges because all tests of cognitive abilities are positively correlated. In our previous example, the two factors have a

correlation of .46, suggesting that children who have externalizing problems are also at risk of having internalizing problems. It is therefore reasonable to calculate a second-order factor score that measures the overall level of behavior problems.

This example illustrates the most commonly used type of factor analysis: *exploratory factor analysis*. Exploratory factor analysis is helpful when we wish to summarize data efficiently, when we are not sure how many factors are present in our data, or when we are not sure which items load on which factors. In short, when we are exploring or looking for factors, we may use exploratory factor analysis. When we think we have found factors and seek to *confirm* this, we may use another variety of factor analysis.

Researchers can use *confirmatory factor analysis* to test highly specific hypotheses. For example, a researcher might want to know if the two different types of items on the WISC-IV Digit Span subtest measure the same ability or two different abilities. On the Digits Forward type of item, the child must repeat a string of digits in the same order in which they were heard. On the Digits Backward type of item, the child must repeat the string of digits in reverse order. Some researchers believe that repeating numbers verbatim measures auditory short-term memory and that repeating numbers in reverse order measures executive control, the ability to allocate attentional resources efficiently to solve multistep problems. Typically, clinicians add the raw scores of both types of items to produce a single score. If the two item types measure different abilities, then adding the raw scores together is kind of like adding apples and oranges, peaches and pears . . . you get the idea. If, however, the two items measure the same ability, then adding the scores together may yield a more reliable score than the separate scores.

Confirmatory factor analysis may be used to determine whether the two item types measure different abilities. We would need to identify or invent several additional tests that are likely to measure the two separate abilities we believe are measured by the two types of Digit Span items. Usually, three tests per factor is sufficient. Let's call the short-term memory tests STM1, STM2, and STM3. Similarly, we can call the executive control tests EC1, EC2, and EC3.

Next, we specify the hypotheses, or models, we wish to test. There are three of them:

1. *All of the tests measure the same ability.* A graphical representation of a hypothesis in confirmatory factor analysis is called a *path diagram*. Tests are drawn with rectangles, and hypothetical factors are drawn with ovals. The correlations between tests and factors are drawn with arrows. The path diagram for this hypothesis is presented in Figure 1.
2. *Both Digits Forward and Digits Backward measure short-term memory and are distinct from executive control.* The path diagram for this hypothesis is presented in Figure 2.
3. *Digits Forward and Digits Backward measure different abilities.* The path diagram for this hypothesis is presented in Figure 3.

Confirmatory factor analysis produces a number of statistics, called *fit statistics,* that tell us which of the models or hypotheses we tested are most in agreement with the data. Studying the results, we can select the model that provides the best fit with the data or perhaps even generate a new model. Actually, factor analysis can quickly become a lot more complicated than described here, but for now, let's hope this is helpful.

Used with permission of W. Joel Schneider.

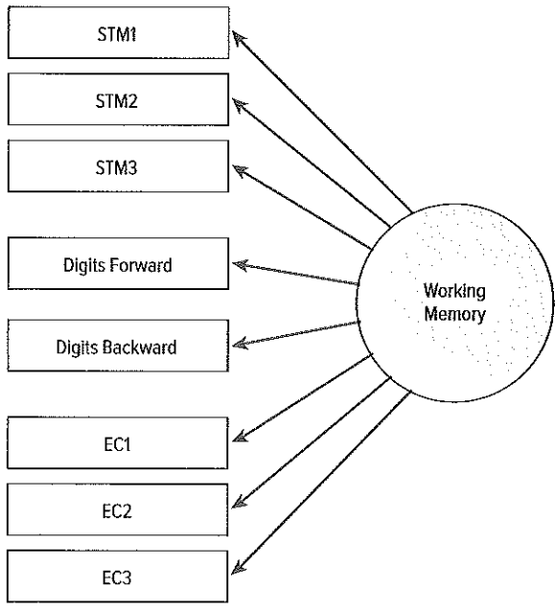


Figure 1
This path diagram is a graphical representation of the hypothesis that All of the tests measure the same ability.

(continued)

Factor Analysis *(continued)*

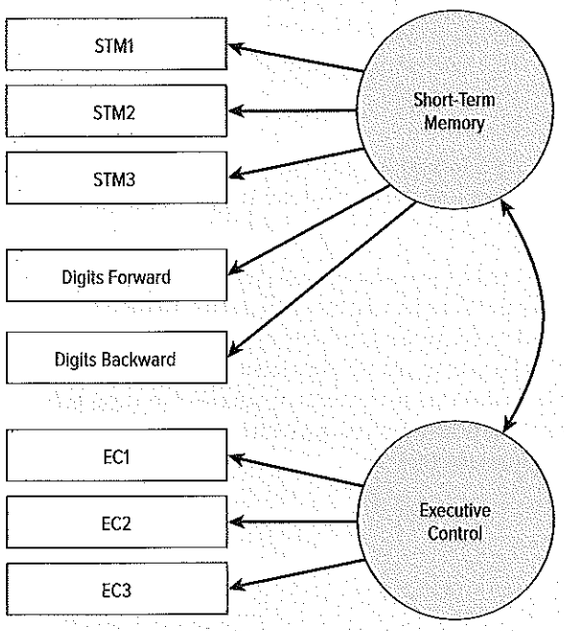


Figure 2
This path diagram is a graphical representation of the hypothesis that Both Digits Forward and Digits Backward measure short-term memory and are distinct from executive control. *Note that the curved arrow indicates the possibility that the two factors might be correlated.*

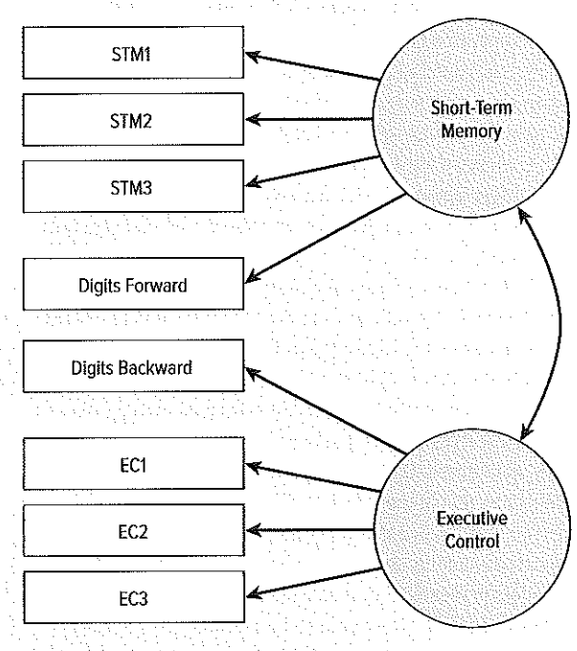


Figure 3
This path diagram is a graphical representation of the hypothesis that Digits Forward and Digits Backward measure different abilities.

exhibited high positive correlations with other intelligence tests were thought to be highly saturated with *g*, whereas tests with low or moderate correlations with other intelligence tests were viewed as possible measures of specific factors (such as visual or motor ability). The greater the magnitude of *g* in a test of intelligence, the better the test was thought to predict overall intelligence.

Spearman (1927) conceived of the basis of the *g* factor as some type of general electrochemical mental energy available to the brain for problem solving. In addition, it was associated with facility in thinking of one's own experience and in making observations and extracting principles. It was *g* rather than *s* that was assumed to afford the best prediction of overall intelligence. Abstract-reasoning problems were thought to be the best measures of *g* in formal tests. As Spearman and his students continued their research, they acknowledged the existence of an intermediate class of factors common to a group of activities but not to all. This class of factors, called **group factors,** is neither as general as

g nor as specific as *s*. Examples of these broad group factors include linguistic, mechanical, and arithmetical abilities.

Other theorists attempted to "dig deeper," to be even more specific about identifying and describing factors other than *g* in intelligence. The number of factors listed to define intelligence in a factor-analytic theory of intelligence may depend, in part, on just how specific the theory is in terms of defining discrete cognitive abilities. These abilities may be conceived of in many ways, from very broad to highly specific. As an example, consider that one researcher has identified an ability "to repeat a chain of verbally presented numbers" that he labels "Factor R." Another researcher analyzes Factor R into three "facilitating abilities" or subfactors, which she labels "ability to process sound" (R1), "ability to retain verbally presented stimuli" (R2), and "speed of processing verbally presented stimuli" (R3). Both researchers present factor-analytic evidence to support their respective positions.[1] Which of these two models will prevail? All other things being equal, it will probably be the model that is perceived as having the greater real-world application, the greater intuitive appeal in terms of how intelligence should be defined, and the greater amount of empirical support.

Many multiple-factor models of intelligence have been proposed. Some of these models, such as that developed by Guilford (1967), have sought to explain mental activities by deemphasizing, if not eliminating, any reference to *g*. Thurstone (1938) initially conceived of intelligence as being composed of seven "primary abilities." However, after designing tests to measure these abilities and noting a moderate correlation between the tests, Thurstone became convinced it was difficult, if not impossible, to develop an intelligence test that did not tap *g*. Gardner (1983, 1994) developed a theory of multiple (seven, actually) intelligences: logical-mathematical, bodily-kinesthetic, linguistic, musical, spatial, interpersonal, and intrapersonal. Gardner (1983) described the last two as follows:

> Interpersonal intelligence is the ability to understand other people: what motivates them, how they work, how to work cooperatively with them. Successful sales people, politicians, teachers, clinicians, and religious leaders are all likely to be individuals with high degrees of interpersonal intelligence. Intrapersonal intelligence, a seventh kind of intelligence, is a correlative ability, turned inward. It is a capacity to form an accurate, veridical model of oneself and to be able to use that model to operate effectively in life. (p. 9)

Aspects of Gardner's writings, particularly his descriptions of **interpersonal intelligence** and **intrapersonal intelligence,** have found expression in popular books written by others on the subject of so-called **emotional intelligence.** But whether or not constructs related to empathy and self-understanding qualify more for the study of emotion and personality than the study of intelligence has been a subject of debate (Davies et al., 1998).

In recent years, a theory of intelligence first proposed by Raymond B. Cattell (1941, 1971) and subsequently modified by Horn (Cattell & Horn, 1978; Horn & Cattell, 1966, 1967) has received increasing attention from test developers as well as test users. As originally conceived by

JUST THINK . . .

Is it possible to develop an intelligence test that does not tap *g*?

1. Recall that factor analysis can take many forms. In exploratory factor analysis, the researcher essentially explores what relationships exist. In confirmatory factor analysis, the researcher is typically testing the viability of a proposed model or theory. Some factor-analytic studies are conducted on the subtests of a single test (such as a Wechsler test), whereas other studies are conducted on subtests from two (or more) tests (such as the current versions of a Wechsler test and the Binet test). The type of factor analysis employed by a theorist may well be the tool that presents that theorist's conclusions in the best possible light.

Cattell, the theory postulated the existence of two major types of cognitive abilities: crystallized intelligence and fluid intelligence. The abilities that make up **crystallized intelligence** (symbolized ***Gc***) include acquired skills and knowledge that are dependent on exposure to a particular culture as well as on formal and informal education (vocabulary, for example). Retrieval of information and application of general knowledge are conceived of as elements of crystallized intelligence. The abilities that make up **fluid intelligence** (symbolized ***Gf***) are nonverbal, relatively culture-free, and independent of specific instruction (such as memory for digits). Through the years, Horn (1968, 1985, 1988, 1991, 1994) proposed the addition of several factors: visual processing (*Gv*), auditory processing (*Ga*), quantitative processing (*Gq*), speed of processing (*Gs*), facility with reading and writing (*Grw*), short-term memory (*Gsm*), and long-term storage and retrieval (*Glr*). According to Horn (1989; Horn & Hofer, 1992), some of the abilities (such as *Gv*) are **vulnerable abilities** in that they decline with age and tend not to return to preinjury levels following brain damage. Others of these abilities (such as *Gq*) are **maintained abilities;** they tend not to decline with age and may return to preinjury levels following brain damage.

Another influential multiple-intelligences model based on factor-analytic studies is the **three-stratum theory of cognitive abilities** (Carroll, 1997). In geology, a stratum is a layer of rock formation having the same composition throughout. Strata (the plural of *stratum*) are illustrated in Figure 9–3, along with a representation of each of the three strata in Carroll's theory. The top stratum or level in Carroll's model is *g*, or general intelligence. The second stratum is composed of eight abilities and processes: fluid intelligence (*Gf*), crystallized intelligence (*Gc*), general memory and learning (*Y*), broad visual perception (*V*),

JUST THINK . . .

Moving from an analogy based on geology to one based on chemistry, think of the periodic table, which lists all known elements. Will it ever be possible to develop a comparable, generally agreed-upon "periodic table" of human abilities?

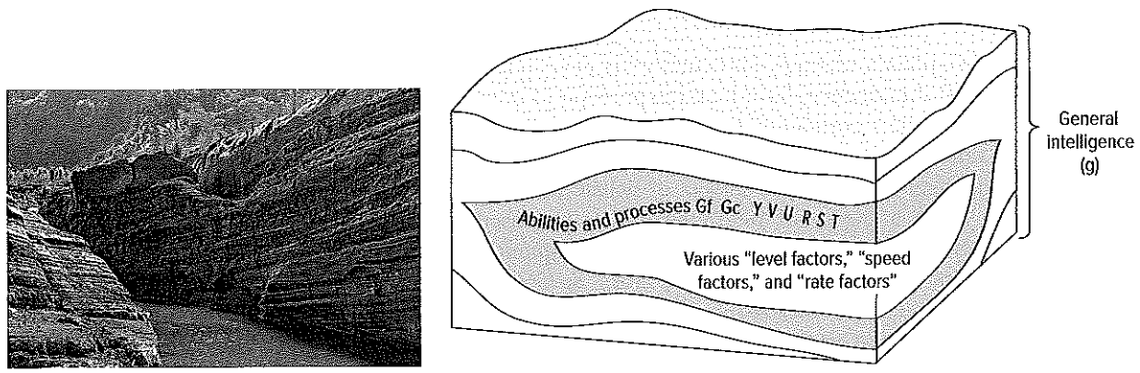


Figure 9–3
Strata in Geology and Carroll's Three-Stratum Theory

Erosion can bare multiple levels of strata on a cliff. In psychology, theory can bare the strata of hypothesized mental structure and function. In Carroll's three-stratum theory of cognitive ability, the first level is g, *followed by a stratum made up of eight abilities and processes, followed by a stratum containing what Carroll refers to as varying "level factors" and "speed factors."*

broad auditory perception (*U*), broad retrieval capacity (*R*), broad cognitive speediness (*S*), and processing/decision speed (*T*). Below each of the abilities in the second stratum are many "level factors" and/or "speed factors"—each different, depending on the second-level stratum to which they are linked. For example, three factors linked to *Gf* are general reasoning, quantitative reasoning, and Piagetian reasoning. A speed factor linked to *Gf* is speed of reasoning. Four factors linked to *Gc* are language development, comprehension, spelling ability, and communication ability. Two speed factors linked to *Gc* are oral fluency and writing ability. The three-stratum theory is a **hierarchical model,** meaning that all of the abilities listed in a stratum are subsumed by or incorporated in the strata above.

Desire for a comprehensive, agreed-upon conceptualization of human cognitive abilities has led some researchers to try to extract elements of existing models to create a new, more complete model. Using factor analysis as well as other statistical tools, these researchers have attempted to modify and reconfigure existing models to better fit empirical evidence. One such modification that has gained increasing attention blends the Cattell-Horn theory with Carroll's three-stratum theory. Although this blending was not initiated by Cattell or Horn or Carroll, it is nonetheless referred to as the Cattell-Horn-Carroll **(CHC) model** of cognitive abilities.

The Cattell-Horn and Carroll models are similar in several respects, among them the designation of broad abilities (second-stratum level in Carroll's theory) that subsume several narrow abilities (first-stratum level in Carroll's theory). Still, any prospective integration of the Cattell-Horn and Carroll models must somehow account for the differences between these two models. One difference involves the existence of a general intellectual (*g*) factor. For Carroll, *g* is the third-stratum factor, subsuming *Gf, Gc,* and the remaining six other broad, second-stratum abilities. By contrast, *g* has no place in the Cattell-Horn model. Another difference between the two models concerns whether or not abilities labeled "quantitative knowledge" and "reading/writing ability" should each be considered a distinct, broad ability as they are in the Cattell-Horn model. For Carroll, all of these abilities are first-stratum, narrow abilities. Other differences between the two models include the notation, the specific definitions of abilities, and the grouping of narrow factors related to memory.

An integration of the Cattell-Horn and Carroll models was proposed by Kevin S. McGrew (1997). On the basis of additional factor-analytic work, McGrew and Flanagan (1998) subsequently modified McGrew's initial CHC model. In its current form, the McGrew-Flanagan CHC model features ten "broad-stratum" abilities and over seventy "narrow-stratum" abilities, with each broad-stratum ability subsuming two or more narrow-stratum abilities. The ten broad-stratum abilities, with their "code names" in parentheses, are labeled as follows: fluid intelligence (*Gf*), crystallized intelligence (*Gc*), quantitative knowledge (*Gq*), reading/writing ability (*Grw*), short-term memory (*Gsm*), visual processing (*Gv*), auditory processing (*Ga*), long-term storage and retrieval (*Glr*), processing speed (*Gs*), and decision/reaction time or speed (*Gt*).

The McGrew-Flanagan CHC model makes no provision for the general intellectual ability factor (*g*). To understand the reason for this omission, it is important to understand why the authors undertook to create the model in the first place. The model was the product of efforts designed to improve the practice of psychological assessment in education (sometimes referred to as **psychoeducational assessment**) by identifying tests from different batteries that could be used to provide a comprehensive assessment of a student's abilities. Having identified key abilities, the authors made recommendations for **cross-battery assessment** of students, or assessment that employs tests from different test batteries and entails interpretation of data from specified subtests to provide a comprehensive assessment. According to these authors,

g was not employed in their CHC model because it lacked utility in psychoeducational evaluations. They explained:

> The exclusion of *g* does not mean that the integrated model does not subscribe to a separate general human ability or that *g* does not exist. Rather, it was omitted by McGrew (1997) (and is similarly omitted in the current integrated model) since it has little practical relevance to cross-battery assessment and interpretation. (McGrew & Flanagan, 1998, p. 14)

Other differences between the Cattell-Horn and Carroll models were resolved more on the basis of factor-analytic studies than judgments regarding practical relevance to cross-battery assessment. The abilities labeled "quantitative knowledge" and "reading/writing" were conceived of as distinct broad abilities, much as they were by Horn and Cattell. McGrew and Flanagan drew heavily on Carroll's (1993) writings for definitions of many of the broad and narrow abilities listed and also for the codes for these abilities.

McGrew (2009) called on intelligence researchers to adopt CHC as a consensus model, thus allowing for a common nomenclature and theoretical framework. Toward that end, he established an online archive of over 460 correlation matrices that formed the basis of Carroll's factor-analytic work.[2] This resource was designed to allow researchers to test the CHC model using confirmatory factor analysis, a more powerful statistical technique than the exploratory factor analysis employed by Carroll.

JUST THINK . . .

Do you agree that *g* has little practical relevance in educational settings?

At the very least, CHC theory as formulated by McGrew and Flanagan has great value from a heuristic standpoint. It compels practitioners and researchers alike to think about exactly how many human abilities really need to be measured and about how narrow or how broad an approach is optimal in terms of being clinically useful. Further, it stimulates researchers to revisit other existing theories that may be ripe for reexamination by means of statistical methods like factor analysis. The best features of such theories might then be combined with the goal of developing a clinically useful and actionable model of human abilities.

Another multifactor theory of intelligence we will mention was proposed by psychometrics pioneer, E. L. Thorndike. According to Thorndike (Thorndike et al., 1909; 1921), intelligence can be conceived in terms of three clusters of ability: social intelligence (dealing with people), concrete intelligence (dealing with objects), and abstract intelligence (dealing with verbal and mathematical symbols). Thorndike also incorporated a general mental ability factor (*g*) into the theory, defining it as the total number of modifiable neural connections or "bonds" available in the brain. For Thorndike, one's ability to learn is determined by the number and speed of the bonds that can be marshaled. No major test of intelligence was ever developed based on Thorndike's multifactor theory. And so, to all would-be or future developers of the next great intelligence test: *This is your moment!* Complete the *Just Think* exercise before reading on.

JUST THINK . . .

Outline notes for your very own version of a "Thorndike Test of Intelligence." How will test items be grouped? What types of items would be found in each grouping? What types of summary scores might be reported for each testtaker? What types of interpretations would be made from the test data?

The information-processing view Another approach to conceptualizing intelligence derives from the work of the Russian neuropsychologist Aleksandr Luria (1966a, 1966b, 1970, 1973, 1980). This approach focuses on the mechanisms by which

2. The archived data is available through the Woodcock-Muñoz Foundation's Human Cognitive Abilities (WMF HCA) project at *http://www.iapsych.com/wmfhcaarchive/map.htm*

information is processed—*how* information is processed, rather than *what* is processed. Two basic types of information-processing styles, simultaneous and successive, have been distinguished (Das et al., 1975; Luria, 1966a, 1966b). In **simultaneous** (or **parallel**) **processing,** information is integrated all at one time. In **successive** (or **sequential**) **processing,** each bit of information is individually processed in sequence. As its name implies, sequential processing is logical and analytic in nature; piece by piece and one piece after the other, information is arranged and rearranged so that it makes sense. In trying to anticipate who the murderer is while watching television shows like *Law & Order, Criminal Minds,* or *Elementary,* for example, one's thinking could be characterized as *sequential.* The viewer constantly integrates bits of information that will lead to a solution of the "whodunnit?" problem. Memorizing a telephone number or learning the spelling of a new word is typical of the types of tasks that involve acquisition of information through successive processing.

By contrast, *simultaneous* processing may be described as "synthesized." Information is integrated and synthesized at once and as a whole. As you stand before and appreciate a painting in an art museum, the information conveyed by the painting is processed in a manner that, at least for most of us, could reasonably be described as simultaneous. Of course, art critics and connoisseurs may be exceptions to this general rule. In general, tasks that involve the simultaneous mental representations of images or information involve simultaneous processing. Map reading is another task that is typical of such processing.

The strong influence of an information-processing perspective is also evident in the work of others (Das, 1972; Das et al., 1975; Naglieri, 1989, 1990; Naglieri & Das, 1988) who have developed what is referred to as the **PASS model** of intellectual functioning. Here, PASS is an acronym for planning, attention, simultaneous, and successive. In this model, *planning* refers to strategy development for problem solving; *attention* (also referred to as *arousal*) refers to receptivity to information; and *simultaneous* and *successive* refer to the type of information processing employed.

Measuring Intelligence

The measurement of intelligence entails sampling an examinee's performance on different types of tests and tasks as a function of developmental level. At all developmental levels, the intellectual assessment process also provides a standardized situation from which the examinee's approach to the various tasks can be closely observed. It therefore provides an opportunity for an assessment that in itself can have great utility in settings as diverse as schools, the military, and business organizations.

Some Tasks Used to Measure Intelligence

In infancy (the period from birth through 18 months), intellectual assessment consists primarily of measuring sensorimotor development. This includes, for example, the measurement of nonverbal motor responses such as turning over, lifting the head, sitting up, following a moving object with the eyes, imitating gestures, and reaching for a group of objects (Figure 9–4). The examiner who attempts to assess the intellectual and related abilities of infants must be skillful in establishing and maintaining rapport with examinees who do not yet know the meaning of words like *cooperation* and *patience.* Typically, measures of infant intelligence rely to a great degree on information obtained from a structured interview with the examinee's parents, guardians, or other caretakers. For school psychologists and others who have occasion to assess young children, enlisting the participation of parents or other caregivers can, practically

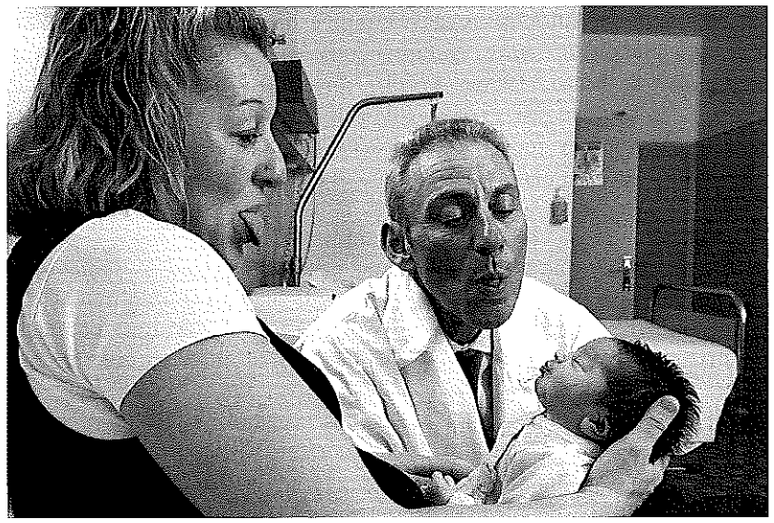


Figure 9–4
Imitation and Cognitive Development

Researchers such as Susan Fenstemacher of the University of Vermont and Kimberly Saudino of Boston University (not pictured here) have explored links among imitation, mental development, temperament, and genetics. Interested readers are referred to their study published in the September–October (2016) issue of Infancy.

speaking, be challenging in its own right. Just ask the consulting psychologist profiled in this chapter's *Meet an Assessment Professional.*

The focus in evaluation of the older child shifts to verbal and performance abilities. More specifically, the child may be called on to perform tasks designed to yield a measure of general fund of information, vocabulary, social judgment, language, reasoning, numerical concepts, auditory and visual memory, attention, concentration, and spatial visualization. The administration of many of the items may be preceded, as prescribed by the test manual, with teaching items designed to provide the examinee with practice in what is required by a particular test item.

According to Wechsler (1958), adult intelligence scales should tap abilities such as retention of general information, quantitative reasoning, expressive language and memory, and social judgment. The types of tasks used to reach these measurement objectives on the Wechsler scale for adults are the same as many of the tasks used on the Wechsler scales for children, although the content of specific items may vary. For a general description of some past and present items, see Table 9–1.

Note that tests of intelligence are seldom administered to adults for purposes of educational placement. Rather, they may be given to obtain clinically relevant information or some measure of learning potential and skill acquisition. Data from the administration of an adult intelligence test may be used to evaluate the faculties of an impaired individual (or one suspected of being senile, traumatized, or otherwise impaired) for the purpose of judging that person's competency to make important decisions (such as those regarding a will, a contract, or other legal matter). Insurance companies rely on such data to make determinations regarding disability. Data from adult intelligence tests may also be used to help make decisions about vocational and career decisions and transitions.

JUST THINK . . .

How else might data from adult intelligence tests be used?

MEET AN ASSESSMENT PROFESSIONAL

Meet Dr. Rebecca Anderson

In my opinion, one of the most important components of an evaluation is generating a report that is reader friendly and provides useful information for parents and teachers who are directly working with the child. A key component of the evaluation is the summary, which should provide a concise picture of the child's strengths and areas of difficulty. Moreover, the recommendations section is a critical element of the report and should provide useful information on ways to support the child's social/emotional and educational success. I try to give recommendations that are accessible to staff and provide tangible tools and suggestions that can be implemented both at home and at school. I often list additional resources (such as books, internet sites, handouts) specific to the child's area of deficiency.

Rebecca Anderson, Ph.D., Independent Practice, Consulting School Psychologist

Realistically, when working within the schools, there are strict timelines, which prohibit extensive evaluations. I think one of the biggest challenges relates to the time and effort that goes into a student's evaluation. Schools are often under budget restrictions and want things done quickly. As a rule of thumb, I conduct more thorough evaluations on students who are receiving an initial evaluation in order to determine the nature of the presenting problem. Less time is required for re-evaluation of students who are already receiving specialized support services. An additional obstacle in the assessment process is accessing parents and staff. I make several efforts to contact parents. If unsuccessful, I note that parental information was unavailable at the time of the evaluation. Ideally, school psychologists would be given ample time and resources and have access to parents and all relevant school personnel, but the reality is that we do the best we can with the time allotted and available resources.

Some Tests Used to Measure Intelligence

As evidenced by reference volumes such as *Tests in Print,* many different intelligence tests exist.[3] From the test user's standpoint, several considerations figure into a test's appeal:

- The theory (if any) on which the test is based
- The ease with which the test can be administered
- The ease with which the test can be scored
- The ease with which results can be interpreted for a particular purpose

3. One objective in this and succeeding chapters is not to in any way duplicate the information that can be found in such reference works. Rather, our more modest objective is to supplement discussion of measurement in a particular area with a brief description or overview of sample tests. In each chapter, only a few of the many tests available for the specified measurement purposes are described. The rationale for selecting these illustrative tests had to do with factors such as historical significance, contemporary popularity, or novelty in contrast to other available tools of assessment. Readers are asked not to draw any conclusions about the value of any particular test on the basis of its inclusion in or omission from our discussion.

Table 9–1
Sample Items Used to Measure Intelligence

Subtest	Description
Information	*In what continent is Brazil?* Questions such as these, which are wide-ranging and tap general knowledge, learning, and memory, are asked. Interests, education, cultural background, and reading skills are some influencing factors in the score achieved.
Comprehension	In general, these questions tap social comprehension, the ability to organize and apply knowledge, and what is colloquially referred to as "common sense." An illustrative question is *Why should children be cautious in speaking to strangers?*
Similarities	*How are a pen and a pencil alike?* This is the general type of question that appears in this subtest. Pairs of words are presented to the examinee, and the task is to determine how they are alike. The ability to analyze relationships and engage in logical, abstract thinking are two cognitive abilities tapped by this type of test.
Arithmetic	Arithmetic problems are presented and solved verbally. At lower levels, the task may involve simple counting. Learning of arithmetic, alertness and concentration, and short-term auditory memory are some of the intellectual abilities tapped by this test.
Vocabulary	The task is to define words. This test is thought to be a good measure of general intelligence, although education and cultural opportunity clearly contribute to success on it.
Receptive Vocabulary	The task is to select from four pictures what the examiner has said aloud. This tests taps auditory discrimination and processing, auditory memory, and the integration of visual perception and auditory input.
Picture Naming	The task is to name a picture displayed in a book of stimulus pictures. This test taps expressive language and word retrieval ability.
Digit Span	The examiner verbally presents a series of numbers, and the examinee's task is to repeat the numbers in the same sequence or backward. This subtest taps auditory short-term memory, encoding, and attention.
Letter-Number Sequencing	Letters and numbers are orally presented in a mixed-up order. The task is to repeat the list with numbers in ascending order and letters in alphabetical order. Success on this subtest requires attention, sequencing ability, mental manipulation, and processing speed.
Picture Completion	The subject's task here is to identify what important part is missing from a picture. For example, the testtaker might be shown a picture of a chair with one leg missing. This subtest draws on visual perception abilities, alertness, memory, concentration, attention to detail, and ability to differentiate essential from nonessential detail. Because respondents may point to the missing part, this test provides a good nonverbal estimate of intelligence. However, successful performance on a test such as this still tends to be highly influenced by cultural factors.
Picture Arrangement	In the genre of a comic-strip panel, this subtest requires the testtaker to re-sort a scrambled set of cards with pictures on them into a story that makes sense. Because the testtaker must understand the whole story before a successful re-sorting will occur, this subtest is thought to tap the ability to comprehend or "size up" a situation. Additionally, attention, concentration, and ability to see temporal and cause-and-effect relationships are tapped.
Block Design	A design with colored blocks is illustrated either with blocks themselves or with a picture of the finished design, and the examinee's task is to reproduce the design. This test draws on perceptual-motor skills, psychomotor speed, and the ability to analyze and synthesize. Factors that may influence performance on this test include the examinee's color vision, frustration tolerance, and flexibility or rigidity in problem solving.
Object Assembly	The task here is to assemble, as quickly as possible, a cut-up picture of a familiar object. Some of the abilities called on here include pattern recognition, assembly skills, and psychomotor speed. Useful qualitative information pertinent to the examinee's work habits may also be obtained here by careful observation of the approach to the task. For example, does the examinee give up easily or persist in the face of difficulty?
Coding	If you were given the dot-and-dash equivalents of several letters in Morse code and then had to write out letters in Morse code as quickly as you could, you would be completing a coding task. The Wechsler coding task involves using a code from a printed key. The test is thought to draw on factors such as attention, learning ability, psychomotor speed, and concentration ability.
Symbol Search	The task is to visually scan two groups of symbols, one search group and one target group, and determine whether the target symbol appears in the search group. The test is presumed to tap cognitive processing speed.
Matrix Reasoning	A nonverbal analogy-like task involving an incomplete matrix designed to tap perceptual organizing abilities and reasoning.
Word Reasoning	The task is to identify the common concept being described with a series of clues. This test taps verbal abstraction ability and the ability to generate alternative concepts.
Picture Concepts	The task is to select one picture from two or three rows of pictures to form a group with a common characteristic. It is designed to tap the ability to abstract as well as categorical reasoning ability.
Cancellation	The task is to scan either a structured or an unstructured arrangement of visual stimuli and mark targeted images within a specified time limit. This subtest taps visual selective attention and related abilities.

- The adequacy and appropriateness of the norms
- The acceptability of the published reliability and validity indices
- The test's utility in terms of costs versus benefits

Historically, some tests seem to have been developed more as a matter of necessity than anything else. In the early 1900s, for example, Alfred Binet was charged with the responsibility of developing a test to screen for children with developmental disabilities in the Paris schools. Binet collaborated with Theodore Simon to create the world's first formal test of intelligence in 1905. Adaptations and translations of Binet's work soon appeared in many countries throughout the world. The original Binet-Simon Scale was in use in the United States as early as 1908 (Goddard, 1908, 1910). By 1912 a modified version had been published that extended the age range of the test downward to 3 months (Kuhlmann, 1912). However, it was the work of Lewis Madison Terman at Stanford University that culminated in the ancestor of what we know now as the Stanford-Binet Intelligence Scale.

In what follows, we briefly set the Stanford-Binet in historical context, and describe several aspects of the test in its current form.

The Stanford-Binet Intelligence Scales: Fifth Edition (SB5) The history of the current version of the Stanford-Binet Intelligence Scales can be traced to Stanford University, and the 1916 publication of an English translation of the Binet-Simon test authored by Lewis Terman (see Figure 9–5).

The result of years of research, Terman's translation and "extension" of the Binet-Simon test featured newly developed test items, and a new methodological approach that included normative studies. Although there were other English translations available, none were as

Figure 9–5
Lewis Madison Terman (1877–1956)

Born on a farm in Indiana, Terman was the 12th of 14 children in the family. After stints at being a teacher and then a school principal, Terman decided to pursue a career in psychology. In 1903 he was awarded a Masters degree. This was followed, two years later, by a doctorate from Clark University. After a few years of teaching child study at Los Angeles State Normal School (a California State teaching college), Terman received an appointment as Assistant Professor in the Education Department at Stanford University. By 1916, largely owing to his revision and refinement of Binet's test, Terman became a prominent figure in the world of psychological testing and assessment. During the first world war, Terman and other leading psychologists were called upon to help the armed forces develop measures that could be used to quickly screen thousands of recruits. Among measurement professionals, Terman is perhaps best remembered for his pioneering innovations in the area of test construction, particularly with regard to standardization. For the larger community, Terman's great contributions to the world of measurement seem to have been overshadowed by his strong, increasingly unpopular views regarding the hereditary nature of intelligence. For example, based on the belief that intelligence is an inherited trait, Terman saw intelligence tests as a tool to identify gifted children, which, in turn could be used as a social tool to identify the best—that is, the most intelligent—leaders (Minton, 2000).

methodologically advanced as Terman's. The publication of the Stanford-Binet had the effect of stimulating a worldwide appetite for intelligence tests (Minton, 1988).

Although the first edition of the **Stanford-Binet** was certainly not without major flaws (such as lack of representativeness of the standardization sample), it also contained some important innovations. It was the first published intelligence test to provide organized and detailed administration and scoring instructions. It was also the first American test to employ the concept of IQ. And it was the first test to introduce the concept of an **alternate item,** an item to be substituted for a regular item under specified conditions (such as the situation in which the examiner failed to properly administer the regular item).

In 1926, Lewis Terman began a project to revise the Stanford-Binet with his former student and subsequent colleague, Maud Merrill (see Figure 9–6). The project would take 11 years to complete. Innovations in the 1937 scale included the development of two equivalent forms, labeled *L* (for Lewis) and *M* (for Maud, according to Becker, 2003), as well as new types of tasks for use with preschool-level and adult-level testtakers.[4] The manual contained many examples to aid the examiner in scoring. The test authors went to then-unprecedented lengths to achieve an adequate standardization sample (Flanagan, 1938), and the test was praised for its technical achievement in the areas of validity and especially reliability. A serious criticism of the test remained: lack of representation of minority groups during the test's development.

Another revision of the Stanford-Binet was well under way at the time of Terman's death at age 79 in 1956. This edition of the Stanford-Binet, the 1960 revision, consisted of only a single form (labeled *L-M*) and included the items considered to be the best from the two forms of the 1937 test, with no new items added to the test. A major innovation, however, was the use of the *deviation IQ* tables in place of the ratio IQ tables. Earlier versions of the Stanford-Binet had employed the *ratio IQ,* which was based on the concept of **mental age** (the age level at which an individual appears to be functioning intellectually as indicated by the level of items

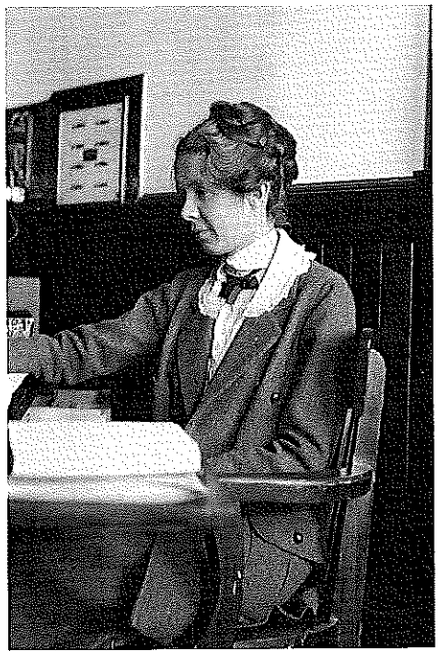


Figure 9–6
Maud Amanda Merrill (1888–1978)

After earning a BA degree from Oberlin College in Minnesota, Merrill was accepted by the Education Department of Stanford University for Masters-level study with Lewis Terman, then a professor in the educational psychology program. Merrill earned a Masters degree in Education in 1920 (Seagoe, 1967) and in 1923 went on to complete a doctorate in psychology, also under Lewis Terman (who had since been promoted to head of the Psychology Department). In her long and distinguished career, Merrill was recognized not only for her expertise on the Stanford-Binet and its administration, but for her expertise in the area of juvenile delinquency (Sears, 1979).

4. L. M. Terman left no clue to what initials would have been used for Forms L and M if his co-author's name had not begun with the letter *M*.

responded to correctly). The **ratio IQ** is the ratio of the testtaker's mental age divided by his or her chronological age, multiplied by 100 to eliminate decimals. As illustrated by the formula for its computation, those were the days, now long gone, when an **IQ** (for **intelligence quotient**) really was a quotient:

$$\text{ratio IQ} = \frac{\text{mental age}}{\text{chronological age}} \times 100$$

A child whose mental age and chronological age were equal would thus have an IQ of 100. Beginning with the third edition of the Stanford-Binet, the deviation IQ was used in place of the ratio IQ. The **deviation IQ** reflects a comparison of the performance of the individual with the performance of others of the same age in the standardization sample. Essentially, test performance is converted into a standard score with a mean of 100 and a standard deviation of 16. If an individual performs at the same level as the average person of the same age, the deviation IQ is 100. If performance is a standard deviation above the mean for the examinee's age group, the deviation IQ is 116.

A third revision of the Stanford-Binet was published in 1972. As with previous revisions, the quality of the standardization sample was criticized. Specifically, the manual was vague about the number of minority individuals in the standardization sample, stating only that a "substantial portion" of Black and Spanish-surnamed individuals was included. The 1972 norms may also have overrepresented the West, as well as large urban communities (Waddell, 1980).

JUST THINK . . .

The term *IQ* is an abbreviation for "intelligence quotient." Despite the fact that modern expressions of intelligence are no longer quotients, the term *IQ* is very much a part of the public's vocabulary. If what is popularly characterized as "IQ" was to be called by something that is more technically accurate, what would "IQ" be called?

The fourth edition of the Stanford-Binet Intelligence Scale (SB:FE; Thorndike et al., 1986) represented a significant departure from previous versions of the Stanford-Binet in theoretical organization, test organization, test administration, test scoring, and test interpretation. Previously, different items were grouped by age and the test was referred to as an **age scale.** The Stanford-Binet: Fourth Edition (SB:FE) was a *point scale.* In contrast to an age scale, a **point scale** is a test organized into subtests by category of item, not by age at which most testtakers are presumed capable of responding in the way that is keyed as correct. The SB:FE manual contained an explicit exposition of the theoretical model of intelligence that guided the revision. The model was one based on the Cattell-Horn (Horn & Cattell, 1966) model of intelligence. A *test composite*—formerly described as a deviation IQ score—could also be obtained. In general, a **test composite** may be defined as a test score or index derived from the combination of, and/or a mathematical transformation of, one or more subtest scores.

The fifth edition of the Stanford-Binet (SB5; Roid, 2003a) was designed for administration to assessees as young as 2 and as old as 85 (or older). The test yields a number of composite scores, including a Full Scale IQ derived from the administration of ten subtests. Subtest scores all have a mean of 10 and a standard deviation of 3. Other composite scores are an Abbreviated Battery IQ score, a Verbal IQ score, and a Nonverbal IQ score. All composite scores have a mean set at 100 and a standard deviation of 15. In addition, the test yields five Factor Index scores corresponding to each of the five factors that the test is presumed to measure (see Table 9–2).

JUST THINK . . .

We live in a society where ability to express oneself in language is highly prized. Should verbal self-expression skills be given more weight on any measure of general ability or intelligence?

The SB5 was based on the Cattell-Horn-Carroll (CHC) theory of intellectual abilities. In fact, according to Roid (2003c), a factor analysis of the early Forms L and M showed that "the CHC factors were clearly recognizable in the early editions of the Binet scales" (Roid et al., 1997, p. 8). The SB5 measures five CHC factors by different types of tasks and subtests at different levels. The five CHC factor names (with abbreviations) alongside their SB5 equivalents are summarized in Table 9–2.

Table 9–2
CHC and Corresponding SB5 Factors

CHC Factor Name	SB5 Factor Name	Brief Definition	Sample SB5 Subtest
Fluid Intelligence (*Gf*)	Fluid Reasoning (FR)	Novel problem solving; understanding of relationships that are not culturally bound	Object Series/Matrices (nonverbal) Verbal Analogies (verbal)
Crystallized Knowledge (*Gc*)	Knowledge (KN)	Skills and knowledge acquired by formal and informal education	Picture Absurdities (nonverbal) Vocabulary (verbal)
Quantitative Knowledge (*Gq*)	Quantitative Reasoning (QR)	Knowledge of mathematical thinking including number concepts, estimation, problem solving, and measurement	Verbal Quantitative Reasoning (verbal) Nonverbal Quantitative Reasoning (nonverbal)
Visual Processing (*Gv*)	Visual-Spatial Processing (VS)	Ability to see patterns and relationships and spatial orientation as well as the gestalt among diverse visual stimuli	Position and Direction (verbal) Form Board (nonverbal)
Short-Term Memory (*Gsm*)	Working Memory (WM)	Cognitive process of temporarily storing and then transforming or sorting information in memory	Memory for Sentences (verbal) Delayed Response (nonverbal)

Also provided in that table is a brief definition of the cognitive ability being measured by the SB5 as well as illustrative SB5 verbal and nonverbal subtests designed to measure that ability.

In designing the SB5, an attempt was made to strike an equal balance between tasks requiring facility with language (both expressive and receptive) and tasks that minimize demands on facility with language. In the latter category are subtests that use pictorial items with brief vocal directions administered by the examiner. The examinee response to such items may be made in the form of nonvocal pointing, gesturing, or manipulating.

After about five years in development and extensive item analysis to address possible objections on the grounds of gender, racial/ethnic, cultural, or religious bias, the final standardization edition of the test was developed. Some 500 examiners from all 50 states were trained to administer the test. Examinees in the norming sample were 4,800 subjects from age 2 to over 85. The sample was nationally representative according to year-2000 U.S. Census data stratified with regard to age, race/ethnicity, geographic region, and socioeconomic level. No accommodations were made for persons with special needs in the standardization sample, although such accommodations were made in separate studies. Persons were excluded from the standardization sample (although included in separate validity studies) if they had limited English proficiency, severe medical conditions, severe sensory or communication deficits, or severe emotional/behavior disturbance (Roid, 2003c).

To determine the reliability of the SB5 Full Scale IQ with the norming sample, an internal-consistency reliability formula designed for the sum of multiple tests (Nunnally, 1967, p. 229) was employed. The calculated coefficients for the SB5 Full Scale IQ were consistently high (.97 to .98) across age groups, as was the reliability for the Abbreviated Battery IQ (average of .91). Test-retest reliability coefficients reported in the manual were also high. The test-retest interval was only 5 to 8 days—shorter by some 20 to 25 days than the interval employed on other, comparable tests. Inter-scorer reliability coefficients reported in the SB5 Technical Manual ranged from .74 to .97 with an overall median of .90. Items showing especially poor inter-scorer agreement had been deleted during the test development process.

Content-related evidence of validity for SB5 items was established in various ways, ranging from expert input to empirical item analysis. Criterion-related evidence was presented in the form of both concurrent and predictive data. For the concurrent studies, Roid (2003c) studied correlations between the SB5 and the SB:FE as well as between the SB5 and all three of the then-current major Wechsler batteries (WPPSI-R, WISC-III, and WAIS-III). The correlations were high when comparing the SB5 to the SB:FE and, perhaps as expected, generally less so

when comparing to the Wechsler tests. Roid (2003c) attributed the difference in part to the varying extents to which the SB5 and the Wechsler tests were presumed to tap *g*. To establish evidence for predictive validity, correlations with measures of achievement (the Woodcock Johnson III Test of Achievement and the Wechsler Individual Achievement Test, among other tests) were employed and the detailed findings reported in the manual. Roid (2003c) presented a number of factor-analytic studies in support of the construct validity of the SB5. However, exactly how many factors best account for what the test is measuring has been a matter of some debate. Some believe as little as one factor, *g,* best describes what the test measures (Canivez, 2008; DiStefano & Dombrowski, 2006). One study of high-achieving third-graders supported a model with 4 factors (Williams et al., 2010). Using a clinical population in her study, another researcher concluded that "the five factor model on which the SB5 was constructed does not reliably hold true across clinical samples." With regard to her clinical sample, she concluded, "Roid's findings were not generalizable" (Chase, 2005, p. 64). At the very least, it can be said that questions have been raised regarding the utility of the SB-5's five factor model, especially with regard to its applicability to clinical populations.

With regard to the "nuts-and-bolts" of test administration, after the examiner has established a rapport with the testtaker, the examination formally begins with an item from what is called a *routing test.* A **routing test** may be defined as a task used to direct or route the examinee to a particular level of questions. A purpose of the routing test, then, is to direct an examinee to test items that have a high probability of being at an optimal level of difficulty. There are two routing tests on the SB5, each of which may be referred to by either their activity names (Object Series/Matrices and Vocabulary) or their factor-related names (Nonverbal Fluid Reasoning and Verbal Knowledge). By the way, these same two subtests—and only these—are administered for the purpose of obtaining the Abbreviated Battery IQ score.

The routing tests, as well as many of the other subtests, contain **teaching items,** which are designed to illustrate the task required and assure the examiner that the examinee understands. Qualitative aspects of an examinee's performance on teaching items may be recorded as examiner observations on the test protocol. However, performance on teaching items is not formally scored, and performance on such items in no way enters into calculations of any other scores.

Some of the ways that the items of a subtest in intelligence and other ability tests are described by assessment professionals have parallels in your home. For example, there is the *floor.* In intelligence testing parlance, the term **floor** refers to the lowest level of the items on a subtest. So, for example, if the items on a particular subtest run the gamut of ability from *developmentally delayed* at one end of the spectrum to *intellectually gifted* at the other, then the lowest-level item at the former end would be considered the *floor* of the subtest. The highest-level item of the subtest is the **ceiling.** On the Binet, another useful term is *basal level,* which is used to describe a subtest with reference to a specific testtaker's performance. Many Binet subtests have rules for establishing a **basal level,** or a base-level criterion that must be met for testing on the subtest to continue. For example, a rule for establishing a basal level might be "Examinee answers two consecutive items correctly." If and when examinees fail a certain number of items in a row, a **ceiling level** is said to have been reached and testing is discontinued.[5]

For each subtest on the SB5, there are explicit rules for where to *start,* where to *reverse,* and where to *stop* (or *discontinue*). For example, an examiner might start at the examinee's estimated

5. Experienced clinicians who have had occasion to test the limits of an examinee will tell you that this assumption is not always correct. **Testing the limits** is a procedure that involves administering test items beyond the level at which the test manual dictates discontinuance. The procedure may be employed when an examiner has reason to believe that an examinee can respond correctly to items at the higher level. On a standardized ability test such as the SB:FE, the discontinue guidelines must be respected, at least in terms of scoring. Testtakers do not earn formal credit for passing the more difficult items. Rather, the examiner would simply note on the protocol that testing the limits was conducted with regard to a particular subtest and then record the findings.

JUST THINK . . .

In what way(s) might an examiner misuse or abuse the obligation to prompt examinees? How could such misuse or abuse be prevented?

present ability level. The examiner might reverse if the examinee scores 0 on the first two items from the start point. The examiner would discontinue testing (stop) after a certain number of item failures after reversing. The manual also provides explicit rules for prompting examinees. If a vague or ambiguous response is given on some verbal items in subtests such as Vocabulary, Verbal Absurdities, or Verbal Analogies, the examiner is encouraged to give the examinee a prompt such as "Tell me more."

Although a few of the subtests are timed, most of the SB5 items are not. The test was constructed this way to accommodate testtakers with special needs and to fit the item response theory model used to calibrate the difficulty of items. Let's also point out that the SB5 has a test administration protocol that could be characterized as *adaptive* in nature.

The SB5 is exemplary in terms of what is called adaptive testing, or testing individually tailored to the testtaker. Other terms used to refer to adaptive testing include tailored testing, sequential testing, branched testing, and response-contingent testing. As employed in intelligence tests, adaptive testing might entail beginning a subtest with a question in the middle range of difficulty. If the testtaker responds correctly to the item, an item of greater difficulty is posed next. If the testtaker responds incorrectly to the item, an item of lesser difficulty is posed. Computerized adaptive testing is in essence designed "to mimic automatically what a wise examiner would do" (Wainer, 1990, p. 10).

Adaptive testing helps ensure that the early test or subtest items are not so difficult as to frustrate the testtaker and not so easy as to lull the testtaker into a false sense of security or a state of mind in which the task will not be taken seriously enough. Three other advantages of beginning an intelligence test or subtest at an optimal level of difficulty are that (1) it allows the test user to collect the maximum amount of information in the minimum amount of time, (2) it facilitates rapport, and (3) it minimizes the potential for examinee fatigue from being administered too many items.

In terms of scoring and interpretation, test manual contains explicit directions for administering, scoring, and interpreting the test in addition to numerous examples of correct and incorrect responses useful in the scoring of individual items. Scores on the individual items of the various subtests are tallied to yield raw scores on each of the various subtests. The scorer then employs tables found in the manual to convert each of the raw subtest scores into a standard score. From these standard scores, composite scores are derived.

When scored by a knowledge test user, an administration of the SB5 may yield much more than a number for a Full Scale IQ and related composite scores: The test may yield a wealth of valuable information regarding the testtaker's strengths and weaknesses with respect to cognitive functioning. This information may be used by clinical and academic professionals in interventions designed to make a meaningful difference in the quality of examinees' lives.

Various methods of profile analysis have been described for use with all major tests of cognitive ability (see, for example, Kaufman & Lichtenberger, 1999). These methods tend to have in common the identification of significant differences between subtest, composite, or other types of index scores as well as a detailed analysis of the factors analyzing those differences. In identifying these significant differences, the test user relies not only on statistical calculations (or tables, if available) but also on the normative data described in the test manual. Large differences between the scores under analysis should be uncommon or infrequent. The SB5 Technical Manual contains various tables designed to assist the test user in analysis. For example, one such table is "Differences Between SB5 IQ Scores and Between SB5 Factor Index Scores Required for Statistical Significance at .05 Level by Age."

In addition to formal scoring and analysis of significant difference scores, the occasion of an individually administered test affords the examiner an opportunity for behavioral observation. More specifically, the assessor is alert to the assessee's **extra-test behavior.** The way

the examinee copes with frustration; how the examinee reacts to items considered very easy; the amount of support the examinee seems to require; the general approach to the task; how anxious, fatigued, cooperative, distractible, or compulsive the examinee appears to be—these are the types of behavioral observations that will supplement formal scores. The SB5 record form includes a checklist form of notable examinee behaviors. Included is a brief, yes–no questionnaire with items such as *Examinee's English usage was adequate for testing* and *Examinee was adequately cooperative.* There is also space to record notes and observations regarding the examinee's physical appearance, mood, and activity level, current medications, and related variables. Examiners may also note specific observations during the assessment. For example, when administering Memory for Sentences, there is usually no need to record an examinee's verbatim response. However, if the examinee produced unusual elaborations on the stimulus sentences, good judgment on the part of the examiner dictates that verbatim responses be recorded. Unusual responses on this subtest may also cue the examiner to possible hearing or speech problems.

A long-standing custom with regard to Stanford-Binet Full Scale scores is to convert them into nominal categories designated by certain cutoff boundaries for quick reference. Through the years, these categories have had different names. For the SB5, here are the cutoff boundaries with their corresponding nominal categories:

Measured IQ Range	Category
145–160	Very gifted or highly advanced
130–144	Gifted or very advanced
120–129	Superior
110–119	High average
90–109	Average
80–89	Low average
70–79	Borderline impaired or delayed
55–69	Mildly impaired or delayed
40–54	Moderately impaired or delayed

With reference to this list, Roid (2003b) cautioned that "the important concern is to describe the examinee's skills and abilities in detail, going beyond the label itself" (p. 150). The primary value of such labels is as a shorthand reference in some psychological reports. For example, in a summary statement at the end of a detailed SB5 report, a school psychologist might write, "In summary, Theodore presents as a well-groomed, engaging, and witty fifth-grader who is functioning in the high average range of intellectual ability."

JUST THINK . . .

Not that very long ago, *moron,* a word with pejorative connotations, was one of the categories in use. What, if anything, can test developers do to guard against the use of classification categories with pejorative connotations?

The next revision of the Stanford-Binet will contain not only changes in item content, but changes that will almost certainly relate to its standardization, administration, scoring, and interpretation. Students of psychological testing and assessment would do well to acquaint themselves with these and related issues (such as issues related to the test's psychometric soundness or theoretical basis), using appropriate resources for more information about the test. For now, let's briefly overview some other tests that have been widely used to measure intelligence.

The Wechsler tests In the early 1930s, psychologist David Wechsler's employer, Bellevue Hospital in Manhattan, needed an instrument for evaluating the intellectual capacity of its multilingual, multinational, and multicultural clients (see Figure 9–7). Dissatisfied with existing intelligence tests, Wechsler began to experiment. The eventual result was a test of his own,

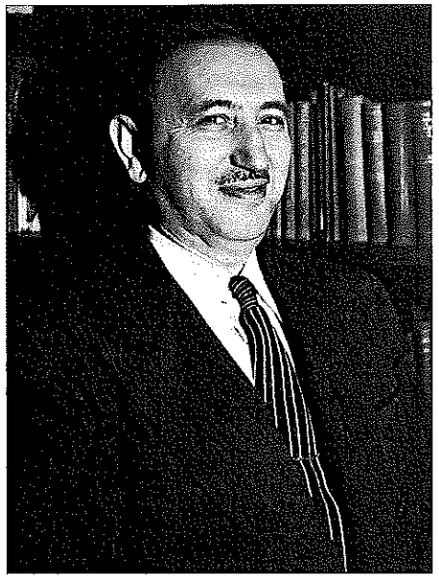


Figure 9–7
David Wechsler (1896–1981)

Born in Romania, David Wechsler came to New York City six years later with his parents and six older siblings. He completed his bachelor's degree in 1916 at City College (New York) and obtained a master's degree at Columbia University the following year. While awaiting induction into the Army at a base in Long Island, Wechsler came in contact with the renowned historian of psychology, E. G. Boring. Wechsler assisted Boring by evaluating the data from one of the first large-scale administrations of a group intelligence test (the Army Alpha test) as the nation geared up for World War I. Wechsler was subsequently assigned to an Army base in Fort Logan, Texas, where his primary duty was administering individual intelligence tests such as the newly published Stanford-Binet Intelligence Scale. Discharged from the Army in 1919, Wechsler spent two years studying in Europe, where he had the opportunity to study with Charles Spearman and Karl Pearson, two brilliant English statisticians known primarily for their work in the area of correlation. Upon his return to New York City, he took a position as a staff psychologist with the Bureau of Child Guidance. In 1935, Wechsler earned a Ph.D. from Columbia University. His dissertation was entitled "The Measurement of Emotional Reactions." By 1932, Wechsler was appointed Chief Psychologist at Bellevue Psychiatric Hospital, a position he held until 1967. An Army private during the first world war, Wechsler served as a measurement consultant to the Armed Forces during the second world war, and as a consultant to the Veterans Administration after the war. Wechsler also assisted in the set-up of a clinic and mental health program for Holocaust survivors in Cyprus in 1947 (Saxon, 1981), and, along with Abraham Maslow, assisted in efforts to launch the Department of Psychology at Hebrew University in Israel ("History of the Department," 2016).

published in 1939. This new test, now referred to as the Wechsler-Bellevue 1 (W-B 1), borrowed from existing tests in format though not in content.

Unlike the most popular individually administered intelligence test of the time, the Stanford-Binet, the W-B 1 was a point scale, not an age scale. The items were classified by subtests rather than by age. The test was organized into six verbal subtests and five performance subtests, and all the items in each test were arranged in order of increasing difficulty. An equivalent alternate form of the test, the W-B 2, was created in 1942 but was never thoroughly standardized (Rapaport et al., 1968). Unless a specific reference is made to the W-B 2, references here (and in the literature in general) to the Wechsler-Bellevue (or the W-B) refer only to the Wechsler-Bellevue 1.

Research comparing the W-B to other intelligence tests of the day suggested that the W-B measured something comparable to what other intelligence tests measured. Still, the test suffered from some problems: (1) The standardization sample was rather restricted; (2) some subtests lacked sufficient inter-item reliability; (3) some of the subtests were made up of items that were too easy; and (4) the scoring criteria for certain items were too ambiguous. Sixteen years after the publication of the W-B, a new Wechsler scale for adults was published: the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955).

Like the W-B, the WAIS was organized into Verbal and Performance scales. Scoring yielded a Verbal IQ, a Performance IQ, and a Full Scale IQ. As a result of many improvements

over its W-B predecessor, the WAIS would quickly become the standard against which other adult tests were compared. A revision of the WAIS, the WAIS-R, was published in 1981 shortly after Wechsler's death in May of that same year. In addition to new norms and updated materials, the WAIS-R test administration manual mandated the alternate administration of verbal and performance tests. In 1997 the third edition of the test (the WAIS-III) was published.

The WAIS-III contained updated and more user-friendly materials. In some cases, test materials were made physically larger to facilitate viewing by older adults. Some items were added to each of the subtests that extended the test's floor in order to make the test more useful for evaluating people with extreme intellectual deficits. Extensive research was designed to detect and eliminate items that may have contained cultural bias. Norms were expanded to include testtakers in the age range of 74 to 89. The test was co-normed with the Wechsler Memory Scale-Third Edition (WMS-III), thus facilitating comparisons of memory with other indices of intellectual functioning when both the WAIS-III and the WMS-III were administered. The WAIS-III yielded a Full Scale (composite) IQ as well as four Index Scores—Verbal Comprehension, Perceptual Organization, Working Memory, and Processing Speed—used for more in-depth interpretation of findings.

JUST THINK . . .

Why is it important to demonstrate that a new version of an intelligence test is measuring much the same thing as a previous version of the test? Why might it be desirable for the test to measure something that was *not* measured by the previous version of the test?

At this writing, the WAIS-IV is the current Wechsler adult scale. It is made up of subtests that are designated either as *core* or *supplemental.* A **core subtest** is one that is administered to obtain a composite score. Under usual circumstances, a **supplemental subtest** (also sometimes referred to as an **optional subtest**) is used for purposes such as providing additional clinical information or extending the number of abilities or processes sampled. There are, however, situations in which a supplemental subtest can be used *in place of* a core subtest. The latter types of situation arise when, for some reason, the use of a score on a particular core subtest would be questionable. So, for example, a supplemental subtest might be substituted for a core subtest if:

- the examiner incorrectly administered a core subtest
- the assessee had been inappropriately exposed to the subtest items prior to their administration
- the assessee evidenced a physical limitation that affected the assessee's ability to effectively respond to the items of a particular subtest

The WAIS-IV contains ten core subtests (Block Design, Similarities, Digit Span, Matrix Reasoning, Vocabulary, Arithmetic, Symbol Search, Visual Puzzles, Information, and Coding) and five supplemental subtests (Letter-Number Sequencing, Figure Weights, Comprehension, Cancellation, and Picture Completion). Longtime users of previous versions of the Wechsler series of adult tests will note the absence of four subtests (Picture Arrangement, Object Assembly, Coding Recall, and Coding Copy-Digit Symbol) and the addition of three new subtests (Visual Puzzles, Figure Weights, and Cancellation). Visual Puzzles and Figure Weights are both timed subtests scored on the WAIS-IV Perceptual Reasoning Scale. In Visual Puzzles, the assessee's task is to identify the parts that went into making a stimulus design. In Figure Weights, the assessee's task is to determine what needs to be added to balance a two-sided scale—one that is reminiscent of the "blind justice" type of scale. In Cancellation, a timed subtest used in calculating the Processing Speed Index, the assessee's task is to draw lines through targeted pairs of colored shapes (while not drawing lines through nontargeted shapes presented as distractors).

Improvements in the WAIS-IV over earlier versions of the test include more explicit administration instructions as well as the expanded use of demonstration and sample

items—this in an effort to provide assessees with practice in doing what is required, in addition to feedback on their performance. Practice items (or teaching items, as they are also called) are presumed to pay dividends in terms of ensuring that low scores are actually due to a deficit of some sort and not simply to a misunderstanding of directions. As is now customary in the development of most tests of cognitive ability, all of the test items were thoroughly reviewed to root out any possible cultural bias. The WAIS-IV also represents an improvement over its predecessor in terms of its "floor" and "ceiling." The floor of an intelligence test is the lowest level of intelligence the test purports to measure. The WAIS-III had a Full Scale IQ floor of 45; the WAIS-IV has a Full Scale IQ floor of 40. The ceiling of an intelligence test is the highest level of intelligence the test purports to measure. The WAIS-III had a Full Scale IQ ceiling of 155; the WAIS-IV has a Full Scale IQ ceiling of 160. If interest in measuring such extremes in intelligence grows, we can expect to see comparable "home improvements" (in the floors and ceilings) in future versions of this and comparable tests.

Because of longer life expectancies, normative data was extended to include information for testtakers up to age 90 years, 11 months. Other changes in the WAIS-IV as compared to the previous edition of this test reflect greater sensitivity to the needs of older adults. These improvements include:

- enlargement of the images in the Picture Completion, Symbol Search, and Coding subtests
- the recommended nonadministration of certain supplemental tests that tap short-term memory, hand-eye coordination, and/or motor speed for testtakers above the age of 69 (this to reduce testing time and to minimize testtaker frustration)
- an average reduction in overall test administration time from 80 to 67 minutes (accomplished primarily by shortening the number of items the testtaker must fail before a subtest is discontinued)

In a bygone era, testtakers' subtest scores on Wechsler tests were used to calculate a Verbal IQ, a Performance IQ, and a Full Scale IQ; that is not the case with the WAIS-IV. As with its predecessor, the WAIS-III, factor-analytic methods were used to help identify the factors that the test seemed to be loading on. The developers of the WAIS-IV deemed the subtests to be loading on four factors: Verbal Comprehension, Working Memory, Perceptual Reasoning, and Processing Speed.[6] Subtests that loaded heavily on any one of these factors were grouped together, and scores on these subtests were used to calculate corresponding index scores. Subtests that loaded less on a particular factor were designated as supplemental with regard to the measurement of that factor (see Table 9–3). As a result, scoring of subtests yields four index scores: a Verbal Comprehension Index, a Working Memory Index, a Perceptual Reasoning Index, and a Processing Speed Index. There is also a fifth index score, the General Ability Index (GAI), which is a kind of "composite of two composites." It is calculated using the Verbal Comprehension and Perceptual Reasoning indexes. The GAI is useful to clinicians as an overall index of intellectual ability.

Another composite score that has clinical application is the Cognitive Proficiency Index (CPI). Comprised of the Working Memory Index and the Processing Speed Index, the CPI is used to identify problems related to working memory or processing speed (Dumont & Willis, 2001). Some researchers have suggested that it can be used in conjunction with the GAI as an aid to better understanding and identifying various learning disabilities (Weiss et al., 2010). Like the GAI and the Full Scale IQ (FSIQ), the CPI was calibrated to have a mean of 100 and a standard deviation of 15.

The WAIS-IV standardization sample consisted of 2,200 adults from the age of 16 to 90 years, 11 months. The sample was stratified on the basis of 2005 U.S. Census data with regard to variables such as age, sex, race/ethnicity, educational level, and geographic region.

6. The WAIS-IV factor called "Perceptual Reasoning" is the same factor that was called "Perceptual Organization" on the WAIS-III.

Table 9–3
WAIS-IV Subtests Grouped According to Indexes

Verbal Comprehension Scale	Perceptual Reasoning Scale	Working Memory Scale	Processing Speed Scale
Similarities[a]	Block Design[a]	Digit Span[a]	Symbol Search[a]
Vocabulary[a]	Matrix Reasoning[a]	Arithmetic[a]	Coding[a]
Information[a]	Visual Puzzles[a]	Letter-Number Sequencing (ages 16–69)[b]	Cancellation (ages 16–69)[b]
Comprehension[b]	Picture Completion[b]		
	Figure Weights (ages 16–69)[b]		

[a]Core subtest.
[b]Supplemental subtest.

Consistent with census data, there were more females than males in the older age bands. As compared to the WAIS-III standardization sample, the WAIS-IV sample is older, more diverse, and has an improved standard of living.

Following a Wechsler tradition, most subtest raw scores for each age group were converted to percentiles and then to a scale with a mean of 10 and a standard deviation of 3. Another Wechsler tradition, beginning with the WAIS-R, called for scaled scores for each subtest to be based on the performance of a "normal" (or, at least, nondiagnosed and nonimpaired) reference group of testtakers 20–34 years old. According to Tulsky et al. (1997), this was done as a consequence of David Wechsler's conviction that "optimal performance tended to occur at these ages" (p. 40). However, the practice was found to contribute to a number of problems in WAIS-R test interpretation, especially with older testtakers (Ivnik et al., 1992; Ryan et al., 1990; Tulsky et al., 1997). Beginning with the WAIS-III and continuing with the WAIS-IV, the practice of deriving norms on a hypothesized "optimal performance" reference group was abandoned. Scores obtained by the testtaker's same-age normative group would serve as the basis for the testtaker's scaled score.[7]

The manual for the WAIS-IV (Coalson & Raiford, 2008) presents data from a number of studies attesting to the reliability, validity, and overall psychometric soundness of the test. For example, high internal consistency reliability estimates were found for all subtests and composite scores for which an estimate of internal consistency is appropriate.[8]

The validity of the WAIS-IV was established by a number of means such as concurrent validity studies and convergent and discriminative validity studies. Additionally, qualitative studies were conducted on the problem-solving strategies testtakers used in responding to questions in order to confirm that they were the same processes targeted for assessment. Independent researchers have noted that although there is comparability between WAIS-IV and SB5 scores in the middle range of intelligence, some discrepancies exist between scores achieved on these tests at the extreme ends of the distribution. For example, in one study, individuals known to be intellectually disabled were found to earn WAIS full scale scores that were roughly 16 points higher than those earned on the SB5 (Silverman et al., 2010).

JUST THINK . . .

Give some thought to your own problem-solving processes. Answer the question "What is the square root of 81?" Now, answer the question "What did you have for dinner last evening?" How are the processes of thought you used to respond to these two questions different? For example, did one of the questions evoke more mental imagery than the other question?

7. However, such reference group scores (derived from the performance of adults from age 20 through age 34 years, 11 months) are still published in the WAIS-IV manual. Presumably, these norms are there for research purposes—or for examiners who seek to determine how an individual testtaker's performance compares with adults in this age group.

8. An estimate of internal consistency would not be appropriate for speeded subtests, such as those subtests used to calculate the Processing Speed Index.

The enthusiasm with which the professional community received the Wechsler adult scale prompted a "brand extension" of sorts downward. The result would be a series of Wechsler intelligence tests for children including the Wechsler Intelligence Scale for Children (WISC) first published in 1949 (currently in its fifth edition), and the Wechsler Pre-School and Primary Scale of Intelligence (WPPSI) first published in 1967 (currently in its fourth edition).

A general description of the various types of tasks measured in current as well as past revisions of these tests is presented in Table 9–1. Additionally, taking full advantage of the benefits of computerized test administration, some of the subtests on some of the newer Wechsler revisions (such as the WISC-V) have been specially re-designed for computerized administration.

Traditionally, whether it was the Wechsler adult scale, the child scale, or the preschool scale, an examiner familiar with one Wechsler test would not have a great deal of difficulty navigating any other Wechsler test. Although this is probably still true, the Wechsler tests have shown a clear trend away from such uniformity. For example, there was a time when all Wechsler scales yielded, among other possible composite scores, a Full Scale IQ (a measure of general intelligence), a Verbal IQ (calculated on the basis of scores on subtests categorized as verbal), and a Performance IQ (calculated on the basis of scores on subtests categorized as nonverbal). All of that changed in 2003 with the publication of the fourth edition of the children's scale, a test that dispensed with the long-standing Wechsler dichotomy of Verbal and Performance subtests.

Regardless of the changes instituted to date, there remains a great deal of commonality between the scales. The Wechsler tests are all point scales that yield deviation IQs with a mean of 100 (interpreted as average) and a standard deviation of 15. On each of the Wechsler tests, a testtaker's performance is compared with scores earned by others in that age group. The tests have in common clearly written manuals that provide descriptions of each of the subtests, including the rationale for their inclusion. The manuals also contain clear, explicit directions for administering subtests as well as a number of standard prompts for dealing with a variety of questions, comments, or other contingencies. There are similar starting, stopping, and discontinue guidelines and explicit scoring instructions with clear examples. For test interpretation, all the Wechsler manuals come with myriad statistical charts that can prove very useful when it comes time for the assessor to make recommendations on the basis of the assessment. In addition, a number of aftermarket publications authored by various assessment professionals are available to supplement guidelines presented in the test manuals.

In general, the Wechsler tests have been evaluated favorably from a psychometric standpoint. Although the coefficients of reliability will vary as a function of the specific type of reliability assessed, reported reliability estimates for the Wechsler tests in various categories (internal consistency, test-retest reliability, inter-scorer reliability) tend to be satisfactory and, in many cases, more than satisfactory. Wechsler manuals also typically contain a great deal of information on validity studies, usually in the form of correlational studies or factor-analytic studies.

Short forms of intelligence tests The term **short form** refers to a test that has been abbreviated in length, typically to reduce the time needed for test administration, scoring, and interpretation. Sometimes, particularly when the testtaker is believed to have an atypically short attention span or other problems that would make administration of the complete test impossible, a sampling of representative subtests is administered. Arguments for such use of Wechsler scales have been made with reference to testtakers from the general population (Kaufman et al., 1991), the elderly (Paolo & Ryan, 1991), and others (Benedict et al., 1992; Boone, 1991; Grossman et al., 1993; Hayes, 1999; Randolph et al., 1993; Ryan & Ward, 1999; Schoop et al., 2001; Sweet et al., 1990).

Short forms of intelligence tests are nothing new. In fact, they have been around almost as long as the long forms. Soon after the Binet-Simon reached the United States, a short form of it was proposed (Doll, 1917). Today, school psychologists with long waiting lists for assessment appointments, forensic psychologists working in an overburdened criminal justice

system, and health insurers seeking to pay less for assessment services are some of the groups to whom the short form appeals.

In 1958, David Wechsler endorsed the use of short forms but only for screening purposes. Years later, perhaps in response to the potential for abuse of short forms, he took a much dimmer view of reducing the number of subtests just to save time. He advised those claiming that they did not have the time to administer the entire test to "find the time" (Wechsler, 1967, p. 37).

Some literature reviews on the validity of short forms have tended to support Wechsler's admonition to "find the time." Watkins (1986) concluded that short forms may be used for screening purposes only, not to make placement or educational decisions. From a historical perspective, Smith, McCarthy, and Anderson (2000) characterized views on the transfer of validity from the parent form to the short form as "overly optimistic." In contrast to some critics who have called for the abolishment of short forms altogether, Smith et al. (2000) argued that the standards for the validity of a short form must be high. They suggested a series of procedures to be used in the development of valid short forms. Silverstein (1990) provided an incisive review of the history of short forms, focusing on four issues: (1) how to abbreviate the original test; (2) how to select subjects; (3) how to estimate scores on the original test; and (4) the criteria to apply when comparing the short form with the original. Ryan and Ward (1999) advised that anytime a short form is used, the score should be reported on the official record with the abbreviation "Est" next to it, indicating that the reported value is only an estimate.

From a psychometric standpoint, the validity of a test is affected by and is somewhat dependent on the test's reliability. Changes in a test that lessen its reliability may also lessen its validity. Reducing the number of items in a test typically reduces the test's reliability and hence its validity. For that reason, decisions made on the basis of data derived from administrations of a test's short form must, in general, be made with caution (Nagle & Bell, 1993). In fact, when data from the administration of a short form clearly suggest the need for intervention or placement, the best practice may be to "find the time" to administer the full form of the test.

Against a backdrop in which many practitioners view short forms as desirable and many psychometricians urge caution in their use, the Wechsler Abbreviated Scale of Intelligence (WASI) was published in 1999. The WASI was designed to answer the need for a short instrument to screen intellectual ability in testtakers from 6 to 89 years of age. The test comes in a two-subtest form (consisting of Vocabulary and Block Design) that takes about 15 minutes to administer and a four-subtest form that takes about 30 minutes to administer. The four subtests (Vocabulary, Block Design, Similarities, and Matrix Reasoning) are WISC- and WAIStype subtests that had high correlations with Full Scale IQ on those tests and are thought to tap a wide range of cognitive abilities. The WASI yields measures of Verbal IQ, Performance IQ, and Full Scale IQ. Consistent with many other intelligence tests, the Full Scale IQ was set at 100 with a standard deviation of 15. The WASI was standardized with 2,245 cases including 1,100 children and 1,145 adults. The manual presents evidence for satisfactory psychometric soundness, although some reviewers of this test were not completely satisfied with the way the validity research was conducted and reported (Keith et al., 2001). However, other reviewers have found that the psychometric qualities of the WASI, as well as its overall usefulness, far exceed those of comparable, brief measures of intelligence (Lindskog & Smith, 2001).

A revision of the WASI referred to, logically enough, as the WASI-2 was published in 2011. The test developers had as their goal an increase in linkage and usability with other Wechsler tests, making the test materials more user friendly, and increasing the psychometric soundness of the test. In general, the WASI-2 test developers seem to have accomplished what they set out to do (Irby & Floyd, 2013). Still, users of an abbreviated measure of intelligence are strongly cautioned that reduced clinical accuracy as compared to the use of a full-length test may be expected to result (McCrimmon & Smith, 2013).

Group tests of intelligence The Stanford revision of the Binet-Simon test was published in 1916, and only one year later, many psychologists were compelled to start thinking about how such a test could be adapted for group administration. To understand why, consider a brief historical look at testing in the military.

On April 6, 1917, the United States entered World War I. On April 7, the president of the American Psychological Association, Robert M. Yerkes, began efforts to mobilize psychologists to help in the war effort. By late May, the APA committee that would develop group tests for the military had their first meeting. There was little debate among the participants about the nature of intelligence, only a clear sense of urgency about developing instruments for the military to identify both the "unfit" and those of "exceptionally superior ability."

Whereas the development of a major intelligence or ability test today might take three to five years, the committee had two tests ready in a matter of weeks and a final form of those tests ready for the printer on July 7. One test became known as the **Army Alpha test.** This test would be administered to Army recruits who could read. It contained tasks such as general information questions, analogies, and scrambled sentences to reassemble. The other test was the **Army Beta test,** designed for administration to foreign-born recruits with poor knowledge of English or to illiterate recruits (defined as "someone who could not read a newspaper or write a letter home"). It contained tasks such as mazes, coding, and picture completion (wherein the examinee's task was to draw in the missing element of the picture). Both tests were soon administered in army camps by teams of officers and enlisted men. By 1919 nearly 2 million recruits had been tested, 8,000 of whom had been recommended for immediate discharge on the basis of the test results. Other recruits had been assigned to various units in the Army based on their Alpha or Beta test results. For example, recruits who scored in the low but acceptable range were likely to draw duty that involved digging ditches or similar kinds of assignments.

If one dream drove the development of the Army Alpha and Beta tests, it was for the Army, other organizations, and society as a whole to run smoothly and efficiently as a result of the proper allocation of human resources—all thanks to tests. Some psychometric scrutiny of the Alpha and Beta tests supported their use. The tests were reliable enough, and they seemed to correlate acceptably with external criteria such as Stanford-Binet Full Scale IQ scores and officers' ratings of men on "practical soldier value." Yerkes (1921) provided this explanation of what he thought the test actually measured:

> The tests give a reliable index of a man's ability to learn, to think quickly and accurately, and to comprehend instructions. They do not measure loyalty, bravery, dependability, or the emotional traits that make a man "carry on." A man's value to the service is measured by his intelligence plus other necessary qualifications. (p. 424)

An original objective of the Alpha and Beta tests was to measure the ability to be a good soldier. However, after the war, that objective seemed to get lost in the shuffle as the tests were used in various aspects of civilian life to measure general intelligence. An Army Alpha or Beta test was much easier to obtain, administer, and interpret than a Stanford-Binet test, and it was also much cheaper. Thousands of unused Alpha and Beta booklets became government surplus that almost anyone could buy. The tests were administered, scored, and interpreted by many who lacked the background and training to use them properly. The utopian vision of a society in which individuals contributed according to their abilities as determined by tests would never materialize. To the contrary, the misuse of tests soured many members of the public and the profession on the use of tests, particularly group tests.

The military's interest in psychological testing during the 1920s and 1930s was minimal. It was only when the threat of a second world war loomed that interest in group intelligence testing reemerged; this led to development of the Army General Classification Test (AGCT). During the course of World War II, the AGCT would be administered to more than 12 million recruits. Other, more specialized tests were also developed by military psychologists. An

assessment unit discretely named the Office of Strategic Services (OSS) developed innovative measures for selecting spies and secret agents to work abroad. By the way, the OSS was a predecessor to today's Central Intelligence Agency (CIA).

Today, group tests are still administered to prospective recruits, primarily for screening purposes. In general, we may define a **screening tool** as an instrument or procedure used to identify a particular trait or constellation of traits at a gross or imprecise level. Data derived from the process of screening may be explored in more depth by more individualized methods of assessment. Various types of screening instruments are used in many different settings. For example, in the following chapter we see how screening tools such as behavior checklists are used in preschool settings to identify young children to be evaluated with more individualized, in-depth procedures.

JUST THINK . . .

James Bond aside, what qualities do you think a real secret agent needs to have? How might you measure these qualities in an applicant?

In the military, the long tradition of using data from screening tools as an aid to duty and training assignments continues to this day. Such data also serve to mold the nature of training experiences. For example, data from group testing have indicated a downward trend in the mean intelligence level of recruits since the inception of an all-volunteer army. In response to such findings, the military has developed new weapons training programs that incorporate, for example, simpler vocabulary in programmed instruction.

Included among many group tests used today by the armed forces are the Officer Qualifying Test (a 115-item multiple-choice test used by the U.S. Navy as an admissions test to Officer Candidate School), the Airman Qualifying Exam (a 200-item multiple-choice test given to all U.S. Air Force volunteers), and the Armed Services Vocational Aptitude Battery (ASVAB). The ASVAB is administered to prospective new recruits in all the armed services. It is also made available to high-school students and other young adults who seek guidance and counseling about their future education and career plans.

Annually, hundreds of thousands of people take the ASVAB, making it perhaps the most widely used multiple aptitude test in the United States. It is administered by school counselors and at various walk-in centers at no cost to the testtaker. In the context of a career exploration program, the ASVAB is designed to help testtakers learn about their interests, abilities, and personal preferences in relation to career opportunities in military and civilian settings. Illustrative items from each of the ten subtests are presented in this chapter's *Everyday Psychometrics.*

Through the years, various forms of the ASVAB have been produced, some for exclusive use in schools and some for exclusive use in the military. A set of 100 selected items included in the subtests of Arithmetic Reasoning, Numerical Operations, Word Knowledge, and Paragraph Comprehension make up a measure within the ASVAB called the Armed Forces Qualification Test (AFQT). The AFQT is a measure of general ability used in the selection of recruits. The different armed services employ different cutoff scores in making accept/reject determinations for service, which are based also on such considerations as their preset quotas for particular demographic groups. In addition to the AFQT score, ten aptitude areas are also tapped on the ASVAB, including general technical, general mechanics, electrical, motor-mechanics, science, combat operations, and skill-technical. These are combined to assess aptitude in five separate career areas, including clerical, electronics, mechanics, skill-technical (medical, computers), and combat operations.

The test battery is continually reviewed and improved on the basis of data regarding how predictive scores are of actual performance in various occupations and military training programs. The ASVAB has been found to predict success in computer programming and computer operating roles (Besetsny et al., 1993), multi-tasking in Navy sailors (Hambrick et al., 2011), and grades in military technical schools across a variety of fields (Earles & Ree, 1992; Ree & Earles, 1990). In one study, the ASVAB adequately predicted grades in three United States Air Force courses

The Armed Services Vocational Aptitude Battery (ASVAB): A Test You Can Take

If you would like firsthand experience in taking an ability test that can be useful in vocational guidance, do what about 900,000 other people do each year and take the Armed Services Vocational Aptitude Battery (ASVAB). Uncle Sam makes this test available to you free of charge—along with other elements of a career guidance package, including a workbook and other printed materials and test scoring and interpretation. Although one objective is to get testtakers "into boots" (that is, into the military), taking the test entails no obligation of military service. For more information about how you can take the ASVAB, contact your school's counseling office or a military recruiter. Meanwhile, you may wish to warm up with the following ten sample items representing each of the ten ASVAB subtests.

I. General Science
Included here are general science questions, including questions from the areas of biology and physics.
1. An eclipse of the sun throws the shadow of the
   a. moon on the sun.
   b. moon on the earth.
   c. earth on the sun.
   d. earth on the moon.

II. Arithmetic Reasoning
The task here is to solve arithmetic problems. Testtakers are permitted to use (government-supplied) scratch paper.
2. It costs $0.50 per square yard to waterproof canvas. What will it cost to waterproof a canvas truck that is 15' × 24'?
   a. $6.67
   b. $18.00
   c. $20.00
   d. $180.00

III. Word Knowledge
Which of four possible definitions best defines the underlined word?
3. Rudiments most nearly means
   a. politics.
   b. minute details.
   c. promotion opportunities.
   d. basic methods and procedures.

IV. Paragraph Comprehension
A test of reading comprehension and reasoning.
4. Twenty-five percent of all household burglaries can be attributed to unlocked windows or doors. Crime is the result of opportunity plus desire. To prevent crime, it is each individual's responsibility to
   a. provide the desire.
   b. provide the opportunity.
   c. prevent the desire.
   d. prevent the opportunity.

V. Numerical Operations
This speeded test contains simple arithmetic problems that the testtaker must solve quickly; it is one of two speeded tests on the ASVAB.
5. 6 — 5 =
   a. 1
   b. 4
   c. 2
   d. 3

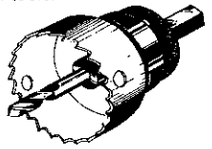
VI. Coding Speed
This subtest contains coding items that measure perceptual/motor speed, among other factors.
KEY

green . . . 2715	man . . . 3451	salt . . . 4586
hat . . . 1413	room . . . 2864	tree . . . 5927

		a.	b.	c.	d.	e.
6.	room	1413	2715	2864	3451	4586

VII. Auto and Shop Information

This test assesses knowledge of automobile shop practice and the use of tools.



7. What tool is shown above?
   a. hole saw
   b. keyhole saw
   c. counter saw
   d. grinding saw

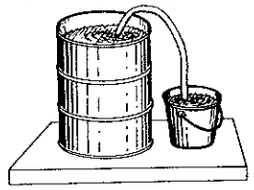
VIII. Mathematics Knowledge

This is a test of ability to solve problems using high-school-level mathematics. Use of scratch paper is permitted.

8. If $3X = -5$, then $X =$
   a. -2
   b. $-5/3$
   c. $-3/5$
   d. $3/5$

IX. Mechanical Comprehension

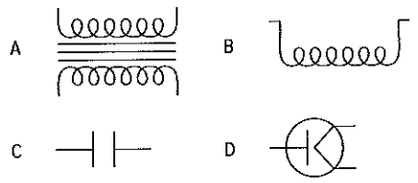
Knowledge and understanding of general mechanical and physical principles are probed by this test.



9. Liquid is being transferred from the barrel to the bucket by
   a. capillary action.
   b. gravitational forces.
   c. fluid pressure in the hose.
   d. water pressure in the barrel.

X. Electronics Information

Here, knowledge of electrical, radio, and electronics information is assessed.



10. Which of the above is the symbol for a transformer?
    a. A
    b. B
    c. C
    d. D

Answer Key

1. b	6. c
2. c	7. a
3. d	8. b
4. d	9. b
5. Why are you looking this one up?	10. a

offered to sensor operators (Carretta et al., 2015).[9] A review of validity studies supports the construct, content, and criterion-related validity of the ASVAB as a device to guide training and selection decisions (Welsh et al., 1990). In general, the test has been deemed quite useful for selection and placement decisions regarding personnel in the armed forces (Chan et al., 1999).

Beyond their applications for military purposes, group tests of intelligence are extensively used in schools and related educational settings. Perhaps no more than a decade or two ago, approximately two-thirds of all school districts in the United States used group intelligence tests on a routine basis to screen 90% of their students. The other 10% were administered individual intelligence tests. Litigation and legislation surrounding the routine use of group intelligence tests have altered this picture somewhat. Still, the group intelligence test, now also referred to as a *school ability test,* is by no means extinct. In many states, legal mandates prohibit the use of group intelligence data alone for class assignment purposes. However, group intelligence test data can, when combined with other data, be extremely useful in developing a profile of a child's intellectual assets.

9. Sensor operators are enlisted aviators who provide a variety of assistance to operators of unmanned, remotely-piloted aircraft.

Group intelligence test results provide school personnel with valuable information for instruction-related activities and increased understanding of the individual pupil. One primary function of data from a group intelligence test is to alert educators to students who might profit from more extensive assessment with individually administered ability tests. The individually administered intelligence test, along with other tests, may point the way to placement in a special class, a program for the gifted, or some other program. Group intelligence test data can also help a school district plan educational goals for all children.

Group intelligence tests in the schools are used in special forms as early as the kindergarten level. The tests are administered to groups of 10 to 15 children, each of whom receives a test booklet that includes printed pictures and diagrams. For the most part, simple motor responses are required to answer items. Oversized alternatives in the form of pictures in a multiple-choice test might appear on the pages, and it is the child's job to circle or place an *X* on the picture that represents the correct answer to the item presented orally by the examiner. During such testing in small groups, the testtakers will be carefully monitored to make certain they are following the directions.

The California Test of Mental Maturity, the Kuhlmann-Anderson Intelligence Tests, the Henmon-Nelson Tests of Mental Ability, and the Cognitive Abilities Test are some of the many group intelligence tests available for use in school settings. The first group intelligence test to be used in U.S. schools was the Otis-Lennon School Ability Test, formerly the Otis Mental Ability Test. In its current edition, the test is designed to measure abstract thinking and reasoning ability and to assist in school evaluation and placement decision-making. This nationally standardized test yields Verbal and Nonverbal score indexes as well as an overall School Ability Index (SAI).

JUST THINK. . .

How has the dynamics of what has traditionally been referred to as "group testing" changed as a result of the administration of tests to groups of testtakers using personal computers?

In general, group tests are useful screening tools when large numbers of examinees must be evaluated either simultaneously or within a limited time frame. More specific advantages—and disadvantages—of traditional group testing are listed in Table 9–4. We qualify group testing with *traditional* because more contemporary forms of group testing, especially testing with all testtakers seated at a computer station, might more aptly be termed *individual assessment simultaneously administered in a group* rather than *group testing*.

Other measures of intellectual abilities Widely used measures of general intelligence sample only a small realm of the many human abilities that may be conceived of as contributing to an individual's intelligence. There are many known intellectual abilities and talents that are not—or are only indirectly—assessed by popular intelligence tests. There are, for example, tests available to measure very specific abilities such as critical thinking, music, or art appreciation. There is also an evolving knowledge base regarding what are called *cognitive styles*. A **cognitive style** is a psychological dimension that characterizes the consistency with which one acquires and processes information (Ausburn & Ausburn, 1978; Messick, 1976). Examples of cognitive styles include Witkin et al.'s (1977) field dependence versus field independence dimension, the reflection versus impulsivity dimension (Messer, 1976), and the visualizer versus verbalizer dimension (Kirby et al., 1988; Paivio, 1971).

Interestingly, although most intelligence tests do not measure creativity, tests designed to measure creativity may well measure variables related to intelligence. For example, some component abilities of creativity are thought to be originality in problem solving, originality in perception, and originality in abstraction. To the extent that tests of intelligence tap these components, measures of creativity may also be thought of as tools for assessing intelligence. A number of tests and test batteries are available to measure creativity in children and adults. In fact, some universities, such as the University of Georgia and the State University College of New York at Buffalo, maintain libraries containing several hundred of these tests. What types of tasks are featured on these tests? And what do these tests really measure?

Table 9–4
The Pros and Cons of Traditional Group Testing

Advantages of Group Tests	Disadvantages of Group Tests
Large numbers of testtakers can be tested at one time, offering efficient use of time and resources.	All testtakers, regardless of ability, typically must start on the same item, end on the same item, and be exposed to every item on the test. Opportunity for adaptive testing is minimized.
Testtakers work independently at their own pace.	Testtakers must be able to work independently and understand what is expected of them, with little or no opportunity for questions or clarification once testing has begun.
Test items are typically in a format easily scored by computer or machine.	Test items may not be in more innovative formats or any format involving examiner manipulation of materials or examiner–examinee interaction.
The test administrator need not be highly trained, as task may require little beyond reading instructions, keeping time, and supervising testtakers.	Opportunity for assessor observation of testtaker's extra-test behavior is lost.
Test administrator may have less effect on the examinee's score than a test administrator in a one-on-one situation.	Opportunity for learning about assessee through assessor–assessee interaction is lost.
Group testing is less expensive than individual testing on a per-testtaker basis.	The information from a group test may not be as detailed and actionable as information from an individual test administration.
Group testing has proven value for screening purposes.	Instruments designed expressly for screening are occasionally used for making momentous decisions.
Group tests may be normed on large numbers of people more easily than an individual test.	In any test-taking situation, testtakers are assumed to be motivated to perform and follow directions. The opportunity to verify these assumptions may be minimized in large-scale testing programs. The testtaker who "marches to the beat of a different drummer" is at a greater risk of obtaining a score that does not accurately approximate his or her hypothetical true score.
Group tests work well with people who can read, follow directions, grip a pencil, and do not require a great deal of assistance.	Group tests may not work very well with people who cannot read, who cannot grip a pencil (such as very young children), who "march to the beat of a different drummer," or who have exceptional needs or requirements.

Four terms common to many measures of creativity are *originality, fluency, flexibility,* and *elaboration. Originality* refers to the ability to produce something that is innovative or nonobvious. It may be something abstract like an idea or something tangible and visible like artwork or a poem. *Fluency* refers to the ease with which responses are reproduced and is usually measured by the total number of responses produced. For example, an item in a test of word fluency might be *In the next thirty seconds, name as many words as you can that begin with the letter w. Flexibility* refers to the variety of ideas presented and the ability to shift from one approach to another. *Elaboration* refers to the richness of detail in a verbal explanation or pictorial display.

A criticism frequently leveled at group standardized intelligence tests (as well as at other ability and achievement tests) is that evaluation of test performance is too heavily focused on whether the answer is correct. The heavy emphasis on correct response leaves little room for the evaluation of processes such as originality, fluency, flexibility, and elaboration. Stated another way, on most achievement tests the thought process typically required is *convergent thinking.* **Convergent thinking** is a deductive reasoning process that entails recall and consideration of facts as well as a series of logical judgments to narrow down solutions and eventually arrive at one solution. In his structure-of-intellect model, Guilford (1967) drew a distinction between the intellectual processes of *convergent* and *divergent* thinking. **Divergent thinking** is a reasoning process in which thought is free to move in many different directions, making several solutions possible. Divergent thinking requires flexibility of thought, originality, and imagination. There is much less emphasis on recall of facts than in convergent thinking. Guilford's model has served to focus research attention not only on the products but also on the process of creative thought.

Guilford (1954) described several tasks designed to measure creativity, such as Consequences ("Imagine what would happen if . . .") and Unusual Uses (e.g., "Name as many uses as you can think of for a rubber band"). Included in Guilford et al.'s (1974) test battery, the Structure-of-Intellect Abilities, are verbally oriented tasks (such as Word Fluency) and nonverbally oriented tasks (such as Sketches).

A number of other tests are available to tap various aspects of creativity. For example, based on the work of Mednick (1962), the Remote Associates Test (RAT) presents the testtaker with three words; the task is to find a fourth word associated with the other three. The Torrance (1966, 1987a, 1987b) Tests of Creative Thinking consist of word-based, picture-based, and sound-based test materials. In a subtest of different sounds, for example, the examinee's task is to respond with the thoughts that each sound conjures up. Each subtest is designed to measure various characteristics deemed important in the process of creative thought.

JUST THINK . . .

Based on this brief description of the RAT and the Torrance Tests, demonstrate your own creativity by creating a new RAT or Torrance Test item that is unmistakably one from the twenty-first century.

It is interesting that many tests of creativity do not fare well when evaluated by traditional psychometric procedures. For example, the test-retest reliability estimates for some of these tests tend to border on the unacceptable range. Some have wondered aloud whether tests of creativity should be judged by different standards from other tests. After all, creativity may differ from other abilities in that it may be highly susceptible to emotional or physical health, motivation, and related factors—even more so than other abilities. This fact would explain tenuous reliability and validity estimates.

JUST THINK . . .

Should tests of creativity be held to different psychometric standards from other ability tests?

As you read about various human abilities and how they all might be related to that intangible construct *intelligence,* you may have said to yourself, "Why doesn't anyone create a test that measures all these diverse aspects of intelligence?"

Although no one has undertaken that ambitious project, in recent years test packages have been developed to test not only intelligence but also related abilities in educational settings. These test packages, called *psychoeducational batteries,* are discussed in the chapter that follows. For now, let's conclude our introduction to intelligence (and intelligent) testing and assessment with a brief discussion of some important issues associated with such measurement.

Issues in the Assessment of Intelligence

Measured intelligence may vary as a result of factors related to the measurement process. Just a few of the many factors that can affect measured intelligence are a test author's definition of intelligence, the diligence of the examiner, the amount of feedback the examiner gives the examinee (Vygotsky, 1978), the amount of previous practice or coaching the examinee has had, and the competence of the person interpreting the test data. There are many other factors that can cause measured intelligence to vary. In what follows, we briefly discuss the role of culture in measured intelligence, as well as a phenomenon that has come to be called the "Flynn effect."

Culture and Measured Intelligence

A culture provides specific models for thinking, acting, and feeling. Culture enables people to survive both physically and socially and to master and control the world around them (Chinoy, 1967). Because values may differ radically between cultural and subcultural groups, people from different cultural groups can have radically different views about what constitutes intelligence (Super, 1983; Wober, 1974). Because different cultural groups value and promote

different types of abilities and pursuits, testtakers from different cultural groups can be expected to bring to a test situation differential levels of ability, achievement, and motivation. These differential levels may even find expression in measured perception and perceptual motor skills.

Consider, for example, an experiment conducted with children who were members of a rural community in eastern Zambia. Serpell (1979) tested Zambian and English research subjects on a task involving the reconstruction of models using pencil and paper, clay, or wire. The English children did best on the paper-and-pencil reconstructions because those were the materials with which they were most familiar. By contrast, the Zambian children did best using wire because that was the medium with which they were most familiar. Both groups of children did about equally well using clay. Any conclusions about the subjects' ability to reconstruct models would have to be qualified with regard to the particular instrument used. This point could be generalized with regard to the use of most any instrument of evaluation or assessment; is it really tapping the ability it purports to tap, or is it tapping something else—especially when used with culturally different subjects or testtakers?

Items on a test of intelligence tend to reflect the culture of the society where the test is employed. To the extent that a score on such a test reflects the degree to which testtakers have been integrated into the society and the culture, it would be expected that members of subcultures (as well as others who, for whatever reason, choose not to identify themselves with the mainstream society) would score lower. In fact, Blacks (Baughman & Dahlstrom, 1968; Dreger & Miller, 1960; Lesser et al., 1965; Shuey, 1966), Hispanics (Gerry, 1973; Holland, 1960; Lesser et al., 1965; Mercer, 1976; Murray, 2007; Simpson, 1970), and Native Americans (Cundick, 1976) tend to score lower on intelligence tests than Whites or Asians (Flynn, 1991). These findings are controversial on many counts—ranging from the great diversity of the people who are grouped under each of these categories, to sampling differences (Zuckerman, 1990), as well as related definitional issues (Daley & Onwuegbuzie, 2011; Sternberg et al., 2005). The meaningfulness of such findings can be questioned further when claims of genetic difference are made owing to the difficulty of separating the effects of genes from effects of the environment. For an authoritative and readable account of the complex issues involved in making such separations see Neisser et al., 1996.

As Gu, He, and Xuqun (2017) have observed, cultural differences with respect to the conceptualization of intelligence extend to culturally appropriate ways of expressing intelligence. In the West, we may be culturally accustomed to expressions of intelligence in the form of writing, speech, debate, and the like. By contrast, in the East, where modesty is culturally valued, such overt demonstrations of one's intellectual prowess may be culturally discouraged. Gu et al. (2017) explained that,

> a component of intelligence in the East has to do with the ability to exhibit culturally appropriate restraint in display of ability. Lao Zi, the philosopher who founded Taoism, states in his work Tao Te Ching, "Whereas the force of words is soon spent, far better is it to keep what is in the heart." This wisdom informs the extent to which a general demeanor of caution and moderation is not only culturally preferable, but seen as the more "intelligent" option. So, all other things being equal, comparing the generally silent person to the generally talkative person, the former may be viewed as the more "intelligent" in the East, while the latter may be viewed as the more "intelligent" in the West.

Alfred Binet shared with many others the desire to develop a measure of intelligence as untainted as possible by factors such as prior education and economic advantages. The Binet-Simon test was designed to separate "natural intelligence from instruction" by "disregarding, insofar as possible, the degree of instruction which the subject possesses" (Binet & Simon, 1908/1961, p. 93). This desire to create what might be termed a **culture-free intelligence test** has resurfaced with various degrees of fervor throughout history. One assumption inherent in the development of such tests is that if cultural factors can be controlled then differences between cultural groups will be lessened. A related assumption is that the effect of culture can be controlled through the elimination

JUST THINK . . .

Is it possible to create a culture-free test of intelligence? Is it desirable to create one?

of verbal items and the exclusive reliance on nonverbal, performance items. Nonverbal items were thought to represent the best available means for determining the cognitive ability of minority group children and adults. However logical this assumption may seem on its face, it has not been borne out in practice (see, for example, Cole & Hunter, 1971; McGurk, 1975).

Exclusively nonverbal tests of intelligence have not lived up to the high expectations of their developers. They have not been found to have the same high level of predictive validity as more verbally loaded tests. This may be due to the fact that nonverbal items do not sample the same psychological processes as do the more verbally loaded, conventional tests of intelligence. Whatever the reason, nonverbal tests tend not to be very good at predicting success in various academic and business settings. Perhaps this is so because such settings require at least some verbal facility.

The idea of developing a truly culture-free test has had great intuitive appeal but has proven to be a practical impossibility. All tests of intelligence reflect, to a greater or lesser degree, the culture in which they were devised and will be used. Stated another way, intelligence tests differ in the extent to which they are *culture-loaded.*

Culture loading may be defined as the extent to which a test incorporates the vocabulary, concepts, traditions, knowledge, and feelings associated with a particular culture. A test item such as "Name three words for snow" is a highly culture-loaded item—one that draws heavily from the Eskimo culture, where many words exist for snow. Testtakers from Brooklyn would be hard put to come up with more than one word for snow (well, maybe two, if you count *slush*).

Soon after it became evident that no test could legitimately be called "culture free," a number of tests referred to as *culture fair* began to be published. We may define a **culture-fair intelligence test** as a test or assessment process designed to minimize the influence of culture with regard to various aspects of the evaluation procedures, such as administration instructions, item content, responses required of testtakers, and interpretations made from the resulting data. Table 9–5 lists techniques used to reduce the culture loading of tests. Note that—in contrast

Table 9–5
Ways of Reducing the Culture Loading of Tests

Culture Loaded	Culture Loading Reduced
Paper-and-pencil tasks	Performance tests
Printed instructions	Oral instructions
Oral instructions	Pantomime instructions
No preliminary practice	Preliminary practice items
Reading required	Purely pictorial
Pictorial (objects)	Abstract figural
Written response	Oral response
Separate answer sheet	Answers written on test itself
Language	Nonlanguage
Speed tests	Power tests
Verbal content	Nonverbal content
Specific factual knowledge	Abstract reasoning
Scholastic skills	Nonscholastic skills
Recall of past-learned information	Solving novel problems
Content graded from familiar to rote	All item content highly familiar
Difficulty based on rarity of content	Difficulty based on complexity of relation education

Source: Jensen (1980).

to the factor-analytic concept of *factor loading,* which can be quantified—the *culture loading* of a test tends to involve more of a subjective, qualitative, nonnumerical judgment.

The rationale for culture-fair test items was to include only those tasks that seemed to reflect experiences, knowledge, and skills common to all different cultures. In addition, all the tasks were designed to be motivating to all groups (Samuda, 1982). An attempt was made to minimize the importance of factors such as verbal skills thought to be responsible for the lower mean scores of various minority groups. Therefore, the culture-fair tests tended to be nonverbal and to have simple, clear directions administered orally by the examiner. The nonverbal tasks typically consisted of assembling, classifying, selecting, or manipulating objects and drawing or identifying geometric designs. Some sample items from the Cattell Culture Fair Test are illustrated in Figure 9–8.

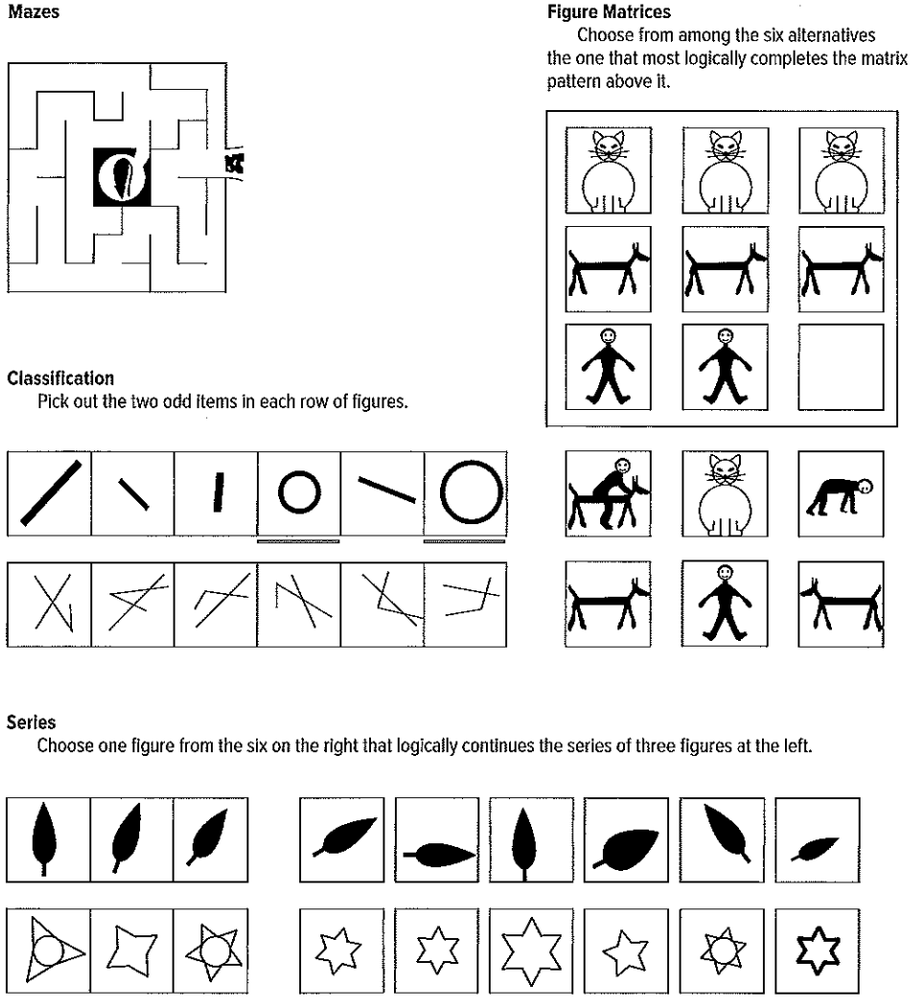


Figure 9–8
Sample "Culture-Fair" and "Culture-Loaded" Items

What types of test items are thought to be "culture-fair"—or at least more culture-fair than other, more culture-loaded items? The items reprinted below from the Culture Fair Test of Intelligence (Cattell, 1940) are a sample. As you look at them, think about how culture-fair they really are. Items from the Culture Fair Test of Intelligence (Cattell, 1940)

Reducing culture loading of intelligence tests seems to lead to a parallel decrease in the value of the test. Culture-fair tests have been found to lack the hallmark of traditional tests of intelligence: predictive validity. Not only that, minority group members still tended to score lower on these tests than did majority group members. Various subcultural characteristics have been presumed to penalize unfairly some minority group members who take intelligence tests that are culturally loaded with American White, middle-class values. Some have argued, for example, that Americans living in urban ghettos share common beliefs and values that are quite different from those of mainstream America. Included among these common beliefs and values, for example, are a "live for today" orientation and a reliance on slang in verbal communication. Native Americans also share a common subculture with core values that may negatively influence their measured intelligence. Central to these values is the belief that individuals should be judged in terms of their relative contribution to the group, not in terms of their individual accomplishments. Native Americans also value their relatively unhurried, present time–oriented lifestyle (Foerster & Little Soldier, 1974).

Frustrated by their seeming inability to develop culture-fair equivalents of traditional intelligence tests, some test developers attempted to develop equivalents of traditional intelligence tests that were culture-specific. Expressly developed for members of a particular cultural group or subculture, such tests were thought to be able to yield a more valid measure of mental development. One culture-specific intelligence test developed expressly for use with African-Americans was the Black Intelligence Test of Cultural Homogeneity (Williams, 1975), a 100-item multiple-choice test. Keeping in mind that many of the items on this test are now dated, here are three samples:[10]

1. *Mother's Day* means
   a. Black independence day.
   b. a day when mothers are honored.
   c. a day the welfare checks come in.
   d. every first Sunday in church.
2. *Blood* means
   a. a vampire.
   b. a dependent individual.
   c. an injured person.
   d. a brother of color.
3. The following are popular brand names. Which one does not belong?
   a. Murray's
   b. Dixie Peach
   c. Royal Crown
   d. Preparation H

As you read the previous items, you may be asking yourself, "Is this really an intelligence test? Should I be taking this seriously?" If you were thinking such questions, you are in good company. At the time, many psychologists probably asked themselves the same questions. In fact, a parody of the BITCH (the test's acronym) was published in the May 1974 issue of *Psychology Today* (p. 101) and was called the "S.O.B. (Son of the Original BITCH) Test." However, the Williams (1975) test was purported to be a genuine culture-specific test of intelligence standardized on 100 Black high-school students in the St. Louis area. Williams was awarded $153,000 by the National Institute of Mental Health to develop the BITCH.

10. The answers keyed correct are as follows: 1(c), 2(d), and 3(d).

In what was probably one of the few published studies designed to explore the test's validity, the Wechsler Adult Intelligence Scale (WAIS) and the BITCH were both administered to Black ($n = 17$) and White ($n = 116$) applicants for a job with the Portland, Oregon, police department. The Black subjects performed much better on the test than did the White subjects, with a mean score that exceeded the White mean score by 2.83 standard deviations. The White mean IQ as measured by the WAIS exceeded the Black mean IQ by about 1.5 standard deviations. None of the correlations between the BITCH score and any of the following variables for either the Black or the White testtakers differed significantly from zero: WAIS Verbal IQ, WAIS Performance IQ, WAIS Full Scale IQ, and years of education. Even though the Black sample in this study had an average of more than 2.5 years of college education, and even though their overall mean on the WAIS was about 20 points higher than for Blacks in general, their scores on the BITCH fell below the average of the standardization sample (high-school pupils ranging in age from 16 to 18). What, then, is the BITCH measuring? The study authors, Matarazzo and Wiens (1977), concluded that the test was measuring a variable that could be characterized as *streetwiseness.* This variable, also known by other names (such as "street smarts" or "street efficacy"), has since received serious attention from researchers (see Figure 9–9).

Many of the tests designed to be culture-specific did yield higher mean scores for the minority group for which they were specifically designed. Still, they lacked predictive validity and provided little useful, practical information. The knowledge required to score high on all of the culture-specific and culture-reduced tests has not been seen as relevant for educational purposes within our pluralistic society. Such tests have low predictive validity for the criterion of success in academic as well as vocational settings.

At various phases in the life history of the development of an intelligence test, a number of approaches to reduce cultural bias may be employed. Panels of experts may evaluate the potential bias inherent in a newly developed test, and those items judged to be biased may be eliminated. The test may be devised so that relatively few verbal instructions are needed to administer it or to demonstrate how to respond. Related efforts can be made to minimize any possible language bias. A tryout or pilot testing with ethnically mixed samples of testtakers may be undertaken. If differences in scores emerge solely as a function of ethnic group membership, individual items may be studied further for possible bias.

Major tests of intelligence have undergone a great deal of scrutiny for bias in many investigations. Procedures range from analysis of individual items to analysis of the test's predictive validity. Only when it can be reasonably concluded that a test is as free as it can be of systematic bias is it made available for use. Of course, even if a test is free of bias, other potential sources of bias still exist. These sources include the criterion for referral for assessment, the conduct of the assessment, the scoring of items (particularly those items that are somewhat

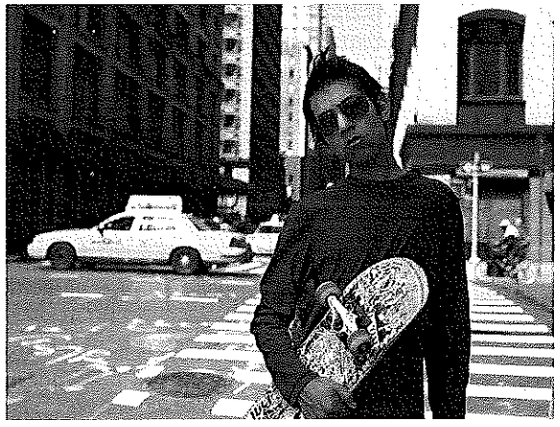


Figure 9–9
"Street Smarts"

A person who "knows his (or her) way around the streets" is referred to as "streetwise" or as possessing "street smarts." This characteristic—which has absolutely nothing to do with map-reading ability—was characterized by Sharkey (2006) as street efficacy (or "the perceived ability to avoid violent confrontations and to be safe in one's neighborhood"). Question: Is this characteristic a personality trait, an aspect of intelligence, or something of a "hybrid"?

subjective), and, finally, the interpretation of the findings. Potentially, there are also less obvious sources of systematic bias in scores on intelligence tests. One such source has come to be known as "the Flynn Effect."

The Flynn Effect

James R. Flynn, while at the Department of Political Studies at the University of Otago in Dunedin, New Zealand, published findings that caused those who study and use intelligence tests in the United States to take notice. In his article entitled "The Mean IQ of Americans: Massive Gains 1932 to 1978," Flynn (1984) presented compelling evidence of what might be termed *intelligence inflation.* He found that measured intelligence seems to rise on average, year by year, starting with the year for which the test is normed. The rise in measured IQ is not accompanied by any academic dividend and so is not thought to be due to any actual rise in "true intelligence." The phenomenon has since been well documented not only in the United States but in other countries as well (Flynn, 1988, 2007). The **Flynn effect** is thus a shorthand reference to the progressive rise in intelligence test scores that is expected to occur on a normed test intelligence from the date when the test was first normed. According to Flynn (2000), the exact amount of the rise in IQ will vary as a function of several factors, such as how culture-specific the items are and whether the measure used is one of fluid or crystallized intelligence.

Beyond being a phenomenon of academic interest, the Flynn effect has wide-ranging, real-world implications and consequences. Flynn (2000) sarcastically advised examiners who want the children they test to be eligible for special services to use the most recently normed version of an intelligence test. On the other hand, examiners who want the children they test to escape the stigma of any labeling were advised to use "the oldest test they can get away with," which should, according to Flynn, allow for at least 10 points' leeway in measured intelligence. At the very least, examiners who use intelligence tests to make important decisions need to be aware of a possible Flynn effect, especially at the beginning or near the end of the test's norming cycle (Kanaya et al., 2003).

JUST THINK . . .

What is your opinion regarding the ethics of Flynn's advice to psychologists and educators who examine children for placement in special classes?

There are numerous other, everyday potential consequences of the Flynn effect ranging from eligibility for special services at school to eligibility for social security benefits. One potential consequence of the Flynn effect has to do with an issue of no less importance than whether one will live or die. Soon after the U.S. Supreme Court ruled it illegal to execute a person who suffers from mental retardation (*Atkins v. Virginia,* 2002), many criminal defense attorneys started familiarizing themselves with the Flynn effect, and investigating whether or not clients accused of capital crimes had been evaluated with an older test—one that spuriously inflated measured intelligence, thereby making such defendants eligible for execution (Fletcher et al., 2010). As might be expected, the ethics of such defense tactics have been questioned, especially because there seems to be sufficient variability in the Flynn effect leading researchers to conclude that not everyone's scores are affected in the same way (Hagan et al., 2010; Zhou et al., 2010).

From a less applied, and more academic perspective, consideration of the Flynn effect can be used to shed light on theories, and to help support or disprove them. For example, Cattell (1971) wrote that fluid intelligence (a product of heredity) formed the basis for crystallized intelligence (a product of learning and the environment). If Cattell was correct, we might expect generational gains in IQ to be due to increased crystallized intelligence—this as a result of factors such as improvements in education, greater educational opportunities for people, and greater cognitive demands in the workplace (Colom et al., 2007). However, according to Flynn (2009), most of the observed increases in IQ have been in the

JUST THINK . . .

In your opinion, are generational gains in measured intelligence due more to factors related more to heredity, environment, or some combination of both?

realm of fluid intelligence. Some research has been designed to address this issue (Rindermann et al., 2010) but the results have been equivocal, with partial support for both Cattell and Flynn.

The Construct Validity of Tests of Intelligence

The evaluation of a test's construct validity proceeds on the assumption that one knows in advance exactly what the test is supposed to measure. For intelligence tests, it is essential to understand how the test developer defined intelligence. If, for example, *intelligence* were defined in a particular intelligence test as Spearman's *g*, then we would expect factor analysis of this test to yield a single large common factor. Such a factor would indicate that the different questions or tasks on the test largely reflected the same underlying characteristic (intelligence, or *g*). By contrast, if intelligence were defined by a test developer in accordance with Guilford's theory, then no one factor would be expected to dominate. Instead, one would anticipate many different factors reflecting a diverse set of abilities. Recall that, from Guilford's perspective, there is no single underlying intelligence for the different test items to reflect. This means that there would be no basis for a large common factor.

In a sense, a compromise between Spearman and Guilford is Thorndike. Thorndike's theory of intelligence leads us to look for one central factor reflecting *g* along with three additional factors representing social, concrete, and abstract intelligences. In this case, an analysis of the test's construct validity would ideally suggest that testtakers' responses to specific items reflected in part a general intelligence but also different types of intelligence: social, concrete, and abstract.

A Perspective

So many decades after the publication of the 1921 symposium, professionals still debate the nature of intelligence and how it should be measured. In the wake of the controversial book *The Bell Curve* (Herrnstein & Murray, 1994), the American Psychological Association commissioned a panel to write a report on intelligence that would carry psychology's official imprimatur. The panel's report reflected wide disagreement with regard to the definition of intelligence but noted that "such disagreements are not cause for dismay. Scientific research rarely begins with fully agreed definitions, though it may eventually lead to them" (Neisser et al., 1996, p. 77).

Another issue that is not going to go away concerns group differences in measured intelligence. Human beings certainly do differ in size, shape, and color, and it is thus reasonable to consider that there is also a physical basis for differences in intellectual ability, so discerning where and how nature can be differentiated from nurture is a laudable academic pursuit. Still, such differentiation remains not only a complex business but one potentially fraught with social, political, and even legal consequences. Claims about group differences can and have been used as political and social tools to oppress religious, ethnic, or other minority group members. This is most unfortunate because, as Jensen (1980) observed, variance attributable to group differences is far less than variance attributable to individual differences. Echoing this sentiment is the view that "what matters for the next person you meet (to the extent that test scores matter at all) is that person's own particular score, not the mean of some reference group to which he or she happens to belong" (Neisser et al., 1996, p. 90).

JUST THINK . . .

In a "real-life" competitive job market, what part—if any—does the "mean of the reference group" play in employment decisions?

The relationship between intelligence and a wide range of social outcomes has been well documented. Scores on intelligence tests, especially when used with other indicators, have value in predicting outcomes such as school performance, years of education, and even social status and income. Measured intelligence is negatively correlated with socially undesirable

outcomes such as juvenile crime. For these and related reasons, we would do well to concentrate research attention on the environmental end of the heredity–environment spectrum. We need to find ways of effectively boosting measured intelligence through environmental interventions, the better to engender hope and optimism.

Unfairly maligned by some and unduly worshipped by others, intelligence has endured—and will continue to endure—as a key construct in psychology and psychological assessment. For this reason, professionals who administer intelligence tests have a great responsibility, one for which thorough preparation is a necessity.

Self-Assessment

Test your understanding of elements of this chapter by seeing if you can explain each of the following terms, expressions, and abbreviations:

accommodation
adaptive testing
AFQT
alerting response
alternate item
Army Alpha test
Army Beta test
assimilation
ASVAB
basal level
Binet, Alfred
ceiling effect
ceiling level
CHC model
cognitive style
convergent thinking
cross-battery assessment
crystallized intelligence
culture-fair intelligence test
culture-free intelligence test
culture loading
deviation IQ
divergent thinking
emotional intelligence
extra-test behavior
factor-analytic theories (of intelligence)
floor
fluid intelligence
Flynn effect
g (factor of intelligence)
Gf and *Gc*
giftedness
group factors
hierarchical model
information-processing theories (of intelligence)
intelligence
interactionism
interpersonal intelligence
intrapersonal intelligence
IQ (intelligence quotient)
maintained abilities
mental age
nominating technique
optional subtest
parallel processing
PASS model
point scale
predeterminism
preformationism
psychoeducational assessment
RAT
ratio IQ
routing test
schema
schemata
screening tool
sequential processing
s factor (of intelligence)
short form
simultaneous processing
Stanford-Binet
successful intelligence
successive processing
supplemental subtest
teaching item
temperament
Terman, Lewis
Termites
testing the limits
three-stratum theory of cognitive abilities
two-factor theory of intelligence
Verbal, Perceptual, and Image Rotation (VPR) model
vulnerable abilities
WAIS
WASI
Wechsler, David
Wechsler-Bellevue
WISC
WPPSI