

How Intelligent Is

by Robert J. Sternberg

A typical American adolescent spends more than 5,000 hours in high school and several thousand more hours studying in the library and at home. But for those students who wish to go on to college, much of their fate is determined in the three or so hours it takes to complete the Scholastic Assessment Test (SAT) or the American College Test (ACT). Four years later they may find themselves in a similar position when they apply to graduate, medical, law or business school.

The stakes are high. In their 1994 book *The Bell Curve*, Richard J. Herrnstein and Charles Murray pointed out a correlation between scores on such tests and a variety of measures of success, such as occupational attainment. They suggested that the U.S. is developing a “cognitive elite”—consisting of high-ability people in prestigious, lucrative jobs—and a larger population of low-ability people in dead-end, low-wage positions. They suggested an invisible hand of nature at work.

But to a large extent, the hand is neither invisible nor natural. We have decided as a society that people who score well on these high-stakes tests will be granted admission to the best schools and, by extension, to the best access routes to success. People have used other criteria, of course: caste at birth, membership in governmental party, religious affiliation. A society can use whatever it wishes—even height, so that very soon people in prestigious occupations would be tall. (Oddly enough, to some extent Americans and many people in other societies already use this criterion.) Why have the U.S. and other countries chosen to use ability tests as a basis to open and close the access gates?

Are they really the measures that should be used? The answers lie in how intelligence testing began.

A Brief History of Testing

Sir Francis Galton, a cousin of Charles Darwin, made the first scientific attempt to measure intelligence. Between 1884 and 1890 Galton ran a service at the South Kensington Museum in London, where, for a small fee, people could have their intelligence checked. The only problem was that Galton’s tests were ill chosen. For example, he contrived a whistle that would tell him the highest pitch a person could perceive. Another test used several cases of gun cartridges filled with layers of either shot, wool or wadding. The cases were identical in appearance and differed only in weight. The test was to pick up the cartridges and then to discriminate the lighter from the heavier. Yet another test was of sensitivity to the smell of roses.

James McKeen Cattell, a psychologist at Columbia University, was so impressed with Galton’s work that in 1890 he devised similar tests to be used in the U.S. Unfortunately for him, a student of his, Clark Wissler, decided to see whether scores on such tests were actually meaningful. In particular, he wanted to know if the scores were related either to one another or to college grades. The answer to both questions proved to be no—so if the tests didn’t predict school performance or even each other, of what use were they? Understandably, interest in Galton’s and Cattell’s tests waned.

A Frenchman, Alfred Binet, got off to a better start. Commissioned to devise a

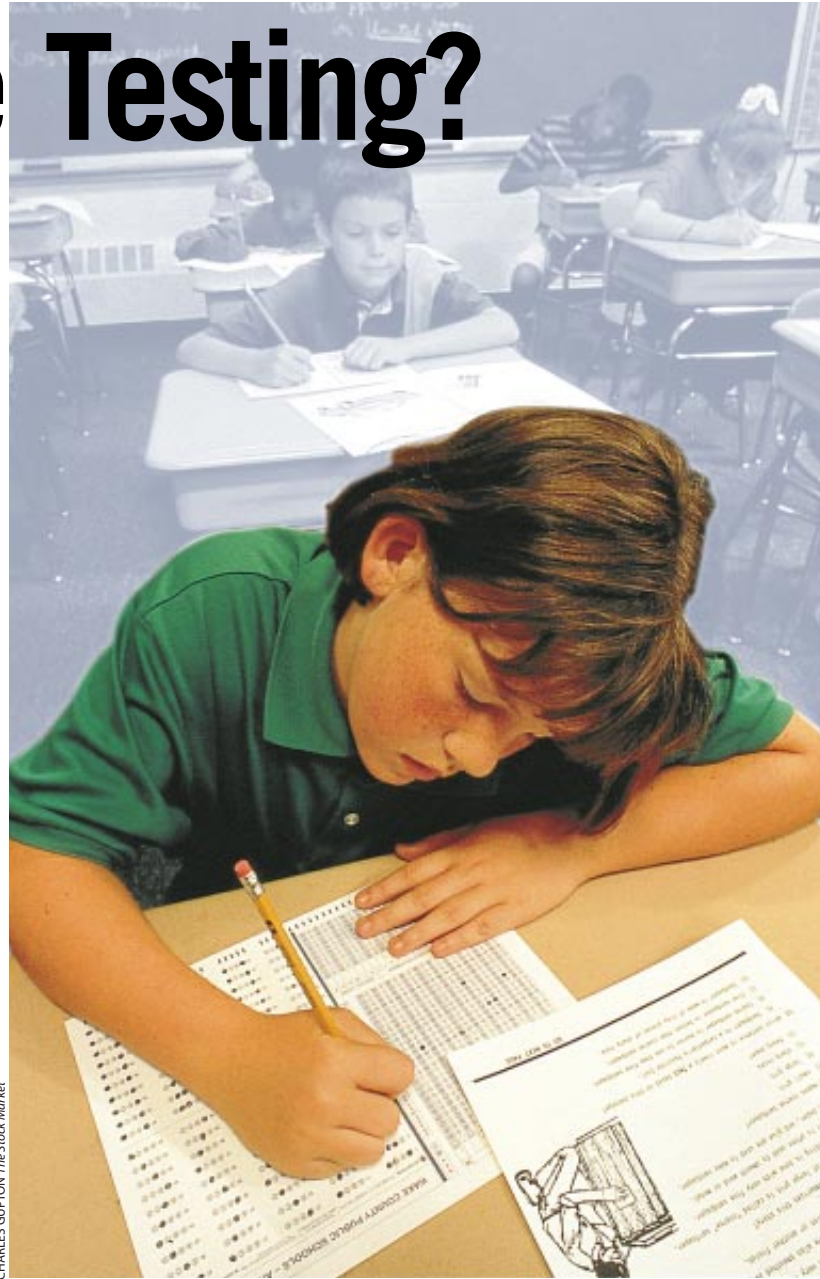
Intelligence Testing?

means to predict school performance, he cast around for test items. Together with his colleague Theodore Simon, he developed a test of intelligence, published in 1905, that measured things such as vocabulary (“What does *misanthrope* mean?”), comprehension (“Why do people sometimes borrow money?”) and verbal relations (“What do an orange, an apple and a pear have in common?”). Binet’s tests of judgment were so successful at predicting school performance that a variant of them, called the Stanford-Binet Intelligence Scale (fourth edition), is still in use today. (Louis Terman of Stanford University popularized the test in the U.S.—hence the name.) A competing test series, the Wechsler Intelligence Scales, measures similar kinds of skills.

It is critical to keep in mind that Binet’s mission was linked to school performance and, especially, to distinguishing children who were genuinely mentally retarded from those who had behavior problems but who were able to think just fine. The result was that the tests were designed, and continue to be designed, in ways that at their best predict school performance.

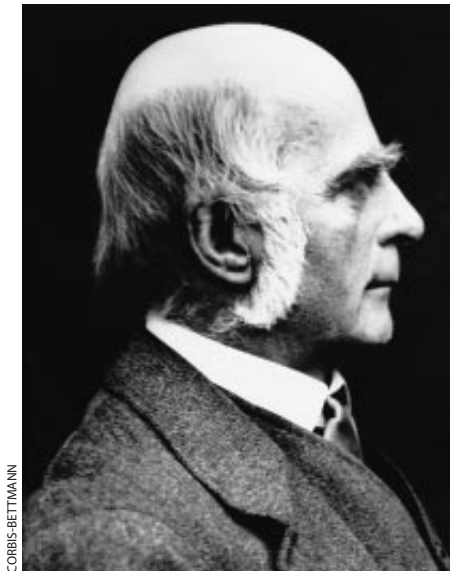
During World War I, intelligence testing really took off: psychologists were asked to develop a method to screen soldiers. That led to the Army Alpha (a verbal test) and Beta (a performance test with pantomimed directions instead of words), which were administered in groups. (Psychologists can now choose between group or individually administered tests, although the individual tests generally give more reliable scores.) In 1926 a new test was introduced, the forerunner to today’s SAT. Devised by Carl C. Brigham of Princeton University, the test provided verbal and mathematical scores.

Shortly thereafter, a series of tests evolved, which today are used to measure various kinds of achievements and abilities, including IQ (intelligence quotient), “scholastic aptitude,” “academic aptitude” and related constructs. Although the names of these tests vary, scores on all of them tend to correlate highly



CHARLES GUPTON The Stock Market

Conventional measures, such as SATs and IQ tests, miss critical abilities essential to academic and professional success



CORBIS-BETTMANN

SIR FRANCIS GALTON made the first scientific attempt to measure intelligence. His tests included determining the pitch of whistles and the weight of gun cartridges. They were not particularly useful.



CORBIS-BETTMANN

ALFRED BINET developed the examination that is the forerunner of the modern IQ test. He devised questions that probed vocabulary, comprehension and verbal abilities to predict school performance.

with one another, so for the purposes of this article I will refer to them loosely as conventional tests of intelligence.

What Tests Predict

Typically, conventional intelligence tests correlate about 0.4 to 0.6 (on a 0 to 1 scale) with school grades, which statistically speaking is a respectable level of correlation. A test that predicts performance with a correlation of 0.5, however,

accounts for only about 25 percent of the variation in individual performances, leaving 75 percent of the variation unexplained. (In statistics, the variation is the square of the correlation, so in this case, $0.5^2 = 0.25$.) Thus, there has to be much more to school performance than IQ.

The predictive validity of the tests declines when they are used to forecast outcomes in later life, such as job performance, salary or even obtaining a job in the first place. Generally, the correla-

tions are only a bit over 0.3, meaning that the tests account for roughly 10 percent of variation in people's performance. That means 90 percent of the variation is unexplained. Moreover, IQ prediction becomes less effective once populations, situations or tasks change. For instance, Fred Fiedler of the University of Washington found that IQ positively predicts leadership success under conditions of low stress. But in high-stress situations, the tests negatively predict success. Some intelligence tests, including both the Stanford-Binet and Wechsler, can yield multiple scores. But can prediction be improved?

Curiously, whereas many kinds of technologies, such as computers and communications, have moved forward in leaps and bounds in the U.S. and around the world, intelligence testing remains almost a lone exception. The content of intelligence tests differs little from that used at the turn of the century. Edwin E. Ghiselli, an American industrial psychologist, wrote an article in 1966 bemoaning how little the predictive value of intelligence tests had improved in 40 years. More than 30 years later the situation remains unchanged.

Improving Prediction

We can do better. In research with Michael Ferrari of the University of Pittsburgh, Pamela R. Clinkenbeard of the University of Wisconsin-Whitewater and Elena L. Grigorenko of Yale University, I showed that a test that measured not only the conventional memory and analytical abilities but also creative and prac-

QUESTIONS REPRESENTATIVE OF IQ and other standardized tests include mathematical deduction and computation, spatial visualization and verbal analogies.

1. The same mathematical rules apply within each row to produce the numbers in the circles. The upper row, for instance, might mean multiplication, whereas the lower row means subtraction. Deduce the rules for the items below and write the answer in the circle.

A.

B.

2. Two of the shapes represent mirror images of the same shape. Underline that pair.

Example:

A.

B.

ANSWERS: 1A. 5; 1B. 3; 2A. A and C; 2B. B and D

Courtesy of *Self-Scoring IQ Tests*, by Victor Serebriakoff and Barnes & Noble and Robinson Publishing

INTELLIGENCE TESTING by Galton took place between 1884 and 1890 at the South Kensington Museum in London.

tical thinking abilities could improve prediction of course grades for high school students in an introductory psychology course. (A direct comparison of correlations between this test and conventional tests is not possible because of the restricted sample, which consisted of high-ability students selected by their schools.)

In these broader tests, individuals had to solve mathematical problems with newly defined operators (for example, X glick $Y = X + Y$ if $X < Y$, and $X - Y$ if $X \geq Y$), which require a more flexible kind of thinking. And they were asked to plan routes on maps and to solve problems related to personal predicaments, which require a more everyday, practical kind of thinking. Here is one example:

The following question gives you information about the situation involving a high school student. Read the question carefully. Choose the answer that provides the best solution, given the specific situation and desired outcomes.

John's family moved to Iowa from Arizona during his junior year in high school. He enrolled as a new student in the local high school two months ago but still has not made friends and feels bored and lonely. One of his favorite activities is writing stories. What is likely to be the most effective solution to this problem?

- A. Volunteer to work on the school newspaper staff
- B. Spend more time at home writing

- C. Try to convince his parents to move back to Arizona
- D. Invite a friend from Arizona to visit during Christmas break

Best answer: A

Creativity can similarly be measured. For example, in another study, Todd Lubart, now at René Descartes University-Paris V, and I asked individuals to perform several creative tasks. They had to write short stories based on bizarre titles such as *The Octopus's Sneakers* or 3853, draw pictures of topics such as the earth seen from an insect's point of view or the end of time, come up with exciting advertisements for bow ties, doorknobs

or other mundane products, and solve quasiscientific problems, such as how someone might find among us extraterrestrial aliens seeking to escape detection. The research found that creative intelligence was relatively domain-specific—that is, people who are creative in one area are not necessarily creative in another—and that creative performance is only weakly to moderately correlated with the scores of conventional measures of IQ.

The implications for such testing extend to teaching. The achievement of students taught in a way that allowed them to make the most of their distinctive pattern of abilities was significantly higher than that of students who were taught in the conventional way, empha-



ARCHIVES OF THE HISTORY OF AMERICAN PSYCHOLOGY, UNIVERSITY OF AKRON

3. Underline the analogous shape.

Example: **A** is to **A** as **B** is to **B** **B** **B**

A. is to as is to
 A B C

B. is to as is to
 A B C D E

4. Complete each analogy by underlining two words from those in the parentheses.

- A. dog is to puppy as (pig, cat, kitten)
- B. circle is to globe as (triangle, square, solid, cube)

5. Underline the two words whose meanings do not belong with the others.

- A. shark, sea lion, cod, whale, flounder
- B. baize, paper, felt, cloth, tinfoil
- C. sword, arrow, dagger, bullet, club

Answers: 3A. B; 3B. E; 4A. Cat, kitten; 4B. Square, cube; 5A. Sea lion, whale (others are fish); 5B. Cloth, tinfoil (others are made of compressed fibers); 5C. Arrow, bullet (others are used by the hand)

sizing memory. Indeed, further research done by Bruce Torff of Hofstra University, Grigorenko and me has shown that the achievements of all students improve, on average, when they are taught to think analytically, creatively and practically about the material they learn, even if they are tested only for memory performance.

Interestingly, whereas individuals higher in conventional (memory and analytical) abilities tended to be primarily white, middle- to upper-middle-class and in “better” schools, students higher in creative and practical abilities tended to be racially, socioeconomically and educationally more diverse, and group differences were not significant. Group differences in conventional test scores—which are common and tend to favor white students—therefore may be in part a function of the narrow range of abilities that standard tests favor.

Tests can also be designed to improve prediction of job performance. Richard K. Wagner of Florida State University and I have shown that tests of practical intelligence in the workplace can predict job performance as well as or better than IQ tests do, even though these tests do not correlate with IQ. In such a test, managers might be told that they have a number of tasks to get done in the next three weeks but do not have time to do them all and so must set priorities. We have devised similar tests for salespeople, students and, most recently, military leaders (in a collaborative effort with psychologists at the U.S. Military Academy at West Point). Such tests do not replace conventional intelligence tests, which also predict job performance, but rather supplement them.

A Question of Culture

Cultural prerogatives also affect scores on conventional tests. Grigorenko and I, in collaboration with Kate Nokes and Ruth Prince of the University of Oxford, Wenzel Geissler of the Danish Bilharziasis Laboratory in Copenhagen, Frederick Okatcha of Kenyatta University in Nairobi and Don Bundy of the University of Cambridge, designed a test of indigenous intelligence for Kenyan children in a rural village. The test required them to perform a task that is adaptive for them:



KPELLE OF WESTERN AFRICA illustrate the shortcoming of translating Western IQ tests for different cultures. The Kpelle would sort items based on functionality—such as “apple” with “eat”—whereas standard tests seek to sort based on category—“apple” with “orange.”

recognizing how to use natural herbal medicines to fight illnesses. Children in the village knew the names of many such medicines and in fact treated themselves once a week on average. (Western children, of course, would know none of them.) The children also took conventional IQ tests.

Scores on the indigenous intelligence test correlated significantly but negatively with vocabulary scores on the Western tests. In other words, children who did better on the indigenous tests actually did worse on the Western tests, and vice versa. The reason may be that parents tend to value indigenous education or Westernized education but not both, and they convey those particular values to their children.

People from different cultures may also interpret the test items differently. In 1971 Michael Cole, now at the University of California at San Diego, and his colleagues studied the Kpelle, who live in western Africa. Cole’s team found that what the Kpelle considered to be a

smart answer to a sorting problem, Westerners considered to be stupid, and vice versa. For instance, given the names of categories such as fruits and vegetables, the Kpelle would sort functionally (for instance, “apple” with “eat”), whereas Westerners would sort categorically (“apple” with “orange,” nested under the word “fruit”).

Westerners do it the way they learn in school, but the Kpelle do it the way they (and Westerners) are more likely to do it in everyday life. People are more likely to think about eating an apple than about sorting an apple into abstract taxonomic categories.

Right now conventional Western tests appear in translated form throughout the world. But the research results necessarily raise the question of whether simply translating Western tests for other cultures makes much sense.

Toward a Better Test

If we can do better in testing than we currently do, then, getting back to the original question posed at the beginning of the article, how have we gotten to where we are? Several factors have conspired to lead us as a society to weigh conventional test scores heavily:

1. *The appearance of precision.* Test scores look so precise that institutions and the people in them often accord them more weight than they probably deserve.
2. *The similarity factor.* A fundamental principle of interpersonal attraction is that people tend to be attracted to those who are similar to them. This principle applies not only in intimate relationships but in work relationships as well. People in positions of power look for others like themselves; because they needed high test scores to get where they are, they tend to seek others who have high test scores.
3. *The publication factor.* Ratings of institutions, such as those published annually in *U.S. News and World Report*, create intense competition among colleges and universities to rank near the top. The institutions cannot control all the factors that go into the ranking. But test scores are relatively easier to control than, say, scholarly publications of faculty, so institutions start to weigh test scores more heavily to prop up their rat-

