Looking Beyond the SAT: Psychological Research Examines the Merit and Future of the High-Stakes Test

May 18, 2004

The SAT, as everyone who ever spent a Saturday hunched over a No. 2 pencil and a machine-readable answer sheet already knows, has deeply marked not only generations of adolescents but American society at large. Scores earned on the nation's premier college admissions test have long served as a kind of surrogate for merit, both personal and institutional. *The Washington Post* cited George W. Bush's 1206 and Al Gore's 1355 40 years after the fact; the coveted top spots in the annual *US News & World Report* college rankings go to colleges with astronomical average SATs; and the Supreme Court has pondered whether admitting minority students with lower standardized test scores violates the unconstitutional rights of rejected whites.

In the nearly 80 years since the first SAT, this highest of high-stakes admissions tests has both reacted to and molded changes in American education and society. It has also responded, but more gradually, to evolving scientific thinking about how to measure the abilities needed to succeed in college. But now, with America facing complex educational challenges and possessing far greater scientific knowledge, the College Board, the consortium of educational institutions that owns the SAT, is exploring newer ways of predicting which applicants can succeed.

In the 1940s, Harvard president James Conant and his assistant, Henry Chauncey, spearheaded the use of standardized testing to break the grip of the Northeastern social elite, and the exclusive New England prep schools they attended, on the nation's most prestigious colleges. Over the following decades, machine-scored meritocracy opened to high achievers from public schools, other regions, and other backgrounds prestigious campuses and positions of leadership in society that had been dominated by scions of privilege. In more recent decades, the nation has sought to expand educational opportunity even farther, to racial and ethnic minorities formerly largely excluded.

But critics have come to view the SAT, last extensively revised in 1990, as more of an obstacle than a vehicle of change. SAT scores show "persistent and substantial differences among subgroups, not just ethnic and racial groups but also language minorities, people with disabilities," according to APS Charter Member Wayne Camara, College Board's vice president for research. Those gaps, he said, "have really only minimally closed over the last three decades."

Science, furthermore, has moved past the "aptitude- and trait-based theories" of intelligence on which "the design of the SAT was predicated" in the 1930s and 40s, said APS Charter Member Howard T. Everson, the College Board's vice president for academic initiatives and its chief research scientist. "Those notions of intelligence and intellectual development are not prominent any more."

Instead, through advances in cognitive psychology, "we've come to learn more about cognitive ability and how it's developed." The SAT, Everson continued, is clearly "a measure of developed abilities" rather than of fixed aptitudes, as once thought. Intellectual development is now known to develop as "a

function of socioeconomic background, educational background and other opportunities." Contrary to the "problematic" older construct, we now "know that [abilities] can change in terms of experience, context, and education." What's needed, therefore, are testing approaches "in line with that theoretical understanding." A revised SAT, still intended to measure verbal and mathematical ability, is scheduled for next year.

But the College Board also wants "to explore different measures of college success to supplement SAT," Camara said. "Are there other factors that we can identify that would identify students … who have different strengths [and] would be successful in college, but [whom] grades, admissions tests, SATs, do not necessarily pick up?" Colleges, of course, have long used information other than the SAT to try to capture factors that cognitive tests ignore. But letters of reference and lists of extracurricular activities "are very subjective and … very difficult to compare," Camara said. Though "we can't make everything standardized," there may be "ways to improve the reliability and the objectivity of these measures of the other factors." To find out, the College Board has commissioned two leading researchers, APS Charter and Fellow Members Neal Schmitt, Michigan State University, and Robert Sternberg, Yale University, to explore innovative measures of the abilities needed for college and "to root these new constructs in theory," Everson said.

A *New York Times* headline has dubbed this effort "SAT-III." At this point, "that's certainly not what we have in mind," Camara said. "We don't know how [or] if this will ever result in a product or a new assessment battery." Instead, the project seeks to bring "greater attention to different predictors, other than the SAT and high school grades."

Long experience has proven the SAT an accurate predictor of grades in early college. "But there's much more to success in college, most of us would argue, than just college grades" and the cognitive abilities that the SAT uses to predict them, Camara said. Being able to "persist, graduating in four or five years … is an accomplishment in and of itself. … There are a lot of factors that probably go into college success." Camara's background in industrial psychology suggests that college success should be viewed "just like we tend to view job success, as multidimensional, as having various factors that load on it."

Schmitt and Sternberg, though starting from quite different theoretical bases, "are actually doing complementary work" for the College Board, Camara explained. Schmitt "is developing measures … very similar to the kind … used in employment," which examine an individual's biographical data and ability to make situational judgments, both well-recognized predictors of job success widely used in conjunction with cognitive testing. In the college arena, such measures might "be helpful in predicting graduation, persistence, and other, similar outcomes," Camara said.

Sternberg, on the other hand, is developing instruments based on his triarchic theory of successful intelligence, which Everson called "one of the more emerging and widely recognized theories" in the field. The theory holds that individuals combine three forms of intelligence – analytical, practical, and creative – to succeed in complex endeavors. Schmitt's and Sternberg's teams have both issued reports on preliminary work, but their projects still have several years yet to run. But one crucial finding is already clear, Camara said. These measures do not show "the kind of subgroup differences that we find on cognitive measures."

"The basic objective [of] employment testing," Schmitt explained, "is to get information that allows us

to predict subsequent job performance. … In the case of college students, we're trying to predict … success at the school they go to." Though collegiate excellence has traditionally been measured by grade point average, academic achievement is only a part, and not necessarily the most important part, of a successful college experience. "When you ask folks, including college presidents, what they believe success in college is, they have very broad definitions," Camara said. "They often say someone who has worked, someone who appreciates the value of learning, someone who understands and has … mastered skills and also appreciates diversity, gets along with folks, has a lifelong interest in learning. They identify a number of different factors."

Before they could measure factors that predict a broader concept of college success, however, Schmitt and his team first had to define what college success is. They began by asking, "What do universities say they want to produce in a student?" and then "decided to take them at their word." This led them to the Web sites of 35 universities and colleges representing a broad range of size, geographic location, public versus private status, and prestige as rated by *US News & World Report*. They searched each site for explicit statements of the institution's mission or educational goals. From the 23 schools providing such information, they extracted 173 phrases that each expressed a single goal.

Three researchers separately sorted the phrases into categories. Then, working together, they agreed on another sorting, from which 12 dimensions emerged. Several additional checks of the chosen categories followed, including comparisons with the industrial and organizational, vocational, and educational psychology literature, and a second sorting by the same researchers.

In addition to such obvious qualities as mastery of fact and ideas, the distilled essence of these universities' statements of purpose – and, presumably, of the qualities of their successful students – includes artistic appreciation, tolerance of diversity, leadership, interpersonal skills, responsible community involvement, adaptability, an interest in continual learning, mental and physical health, orientation to a career goal, perseverance, and ethical standards. "We built our whole model around" the proposition that the 12 dimensions, taken together, "are success in college," Schmitt said. Next, the team constructed two instruments to measure the 12 dimensions in students. On the assumption that "what predicts success is probably similar past experience," one instrument considers biographical information. The other evaluates individuals' situational judgment about problematic or challenging incidents typical of student life and ordinary daily life.

"Biodata is simply a set of questions in multiple choice format that asks individuals to tell us … their background experiences and interests," Schmitt explained. Particular questions hone in on particular dimensions from the list of 12. Students are asked, for example, "How many leadership positions have you held while you were in high school? How many volunteer organizations did you join? What kinds of activities did you engage in in those volunteer activities? How many books have you read the last year outside of class? What kinds of social-citizen … activities have you engaged in?" Individual items are generally scored from one to five. Adding across the items produces the total score.

Unlike the SAT, which tests for predetermined knowledge or skills, the biodata instrument does not seek particular experiences. Rather, it looks for examples of personal qualities that might be expressed in many different ways in the lives of students from a wide range of cultural and personal circumstances, including, for example, the young person who can't run for class office or captain the varsity team because he or she must hold down a job or help care for a disabled family member.

“I had to go home and milk the cows,” Schmitt recalled. Able to test some 300 personal qualities, the instrument affords “a variety of different ways in which one could evidence leadership. … We can capture their leadership in whatever domain it’s been evidenced. … There’s no right or wrong way. … Part of the success of this measure is that we generate items that are broad enough to allow a person who has such an odd background to excel.”

The situational judgment questions also plumb personal qualities rather than specific information or skills. Items present incidents from college life – a conflict between the need to study and the desire to have fun, for example, or difficulties among classmates working together on an assigned project – and an array of possible responses. The student indicates which action or approach is the best solution and which is the worst.

“Lots of data” show that subgroup differences “on these measures” are smaller than on the SAT, Schmitt said. “We’re measuring something different … social responsibility, motivation, leadership. They’re measuring academic knowledge.” But because of these instruments’ very open-endedness, faking, exaggeration, even outright lying are among “the major problems associated with [using them] in high stakes testing,” he continued. Measures that “seem to have some effect” on the propensity to dissemble include a forthright warning statement that lying, if discovered, disqualifies the applicant. Asking for elaboration – not only how many languages a student speaks or books he or she has read, but which languages and books – also tends to lower scores. In addition, the biodata instrument includes what “amounts to a lie scale,” Schmitt said. It asks how often students have done something “that we know they can’t have done … For example, ‘Have you written software programs in alternate basic language?’… Answering yes means you’ve lied. It doesn’t exist. We put in five or six items like that.”

Coaching is another bane of all high-stakes testing. “Our answer,” Schmitt said, is to “provide the coaching as part of the tool, actually tell them up front … what we’re trying to measure and put everyone on an equal footing, not just those who got coached. … We’ll describe the dimensions that we’re trying to measure. … Our philosophy is that if we provide coaching to everyone as part of the directions, … they will all be on the same footing. There won’t be any motivation to provide [outside] coaching.”

SAT – Sternberg’s Active Theory

Sternberg’s approach also seeks to measure multiple capacities, which he terms analytical, practical, and creative intelligence. In his theory, these are “not separate intelligences,” he explained. Rather, “they’re related, but not hierarchically.” A set of “information processing components” such as “mentally representing information [and] monitoring your progress … are common to everybody” and are applied across a variety of situations. When “they are applied to fairly abstract but somewhat conventional kinds of problems” – studying an academic subject, for example – “that’s analytical ability.” Applying them to “everyday life, … to adapt to, shape, and select environments,” constitutes practical intelligence. The difference between these two is the “distinction between solving a theorem and writing a grant,” he explained.

Creative intelligence involves applying the mental processes to “generate ideas and products that are novel, high in quality, and appropriate to the task at hand. It could be anything” – an artistic work, a novel use of a tool, an original sales approach. “In my theory the three kinds of processing are highly interactive. … The components are all the same. Everybody has to monitor their problem-solving and evaluate it. … You need creative skills to come up with ideas, you need analytical skills to know if

they're good ideas, and you need practical skills to make the ideas work." The specific skills that demonstrate intelligence, however, vary widely among individuals. "What's practical for kids that we study in rural Kenya is totally impractical here," Sternberg noted. The mental components are "not a part of the brain. … Rather, [what's] practical depends on where you are, … on the context in which you live."

Sternberg's theory therefore conceives intelligence – and intelligence tests – as covering a broad range of forms and content. The mental processes, not the particular academic skills valued in upper-middle-class American high schools, constitute intelligence. "Kids with different backgrounds develop different profiles of skills," he emphasized. "The more skills you test, the more different kinds of socializations you value. … All tests are culturally loaded. Only the extreme right-wing argue for a culture free test. … There is no such thing."

Even a given test, in fact, "does not necessarily measure the same things for different people," Sternberg said. The mental capacities needed to answer questions differ with the individual test-taker's experience. For a poor child from the inner city, "the SAT is more novel than for a kid from Scarsdale. [It] is not really, in terms of my theory, measuring the same thing for the two kids, [particularly] not in terms of its level of novelty," because the suburban student has extensive training in the tested skills that the inner-city child lacks. For those two students, "it's a different test. … Most people just say, 'The test is what's on a piece of paper' " and must therefore be the same for everyone. "But what a test measures is in the interaction between the person and the item. It doesn't inhere in the item, it inheres in the interaction."

In a society attempting to expand opportunity to formerly excluded groups, therefore, "what you do is … include subtests that have … skills developed in a wider range of cultures. The answer is not to reduce culture, which [you] can't do anyway. It's to tap into a wider variety of cultures. It's to say we're not going to just look at white upper middle class Scarsdale culture. We're going to look at other ones, too."

The instrument developed by Sternberg's Rainbow Project – named for humanity's wide spectrum of intelligence – contains tests for all three forms of intelligence. The analytical section involves verbal, quantitative, and figural skills. The practical skills section measures situational judgment by having students view short videos or read descriptions of problematic situations in everyday life and in college life, then rate six proposed solutions from very good to very bad. A creative skills test asks the student to write or tell stories based on a title or a series of pictures and to write a caption to a New Yorker cartoon.

Sternberg claimed two advantages for his instrument over the SAT alone. "One is reduced ethnic group difference and the other is … doubled prediction of college performance" as measured by GPA. Testing for a broader range of intelligence usefully serves the cause of diversity, Sternberg believes. "One way of achieving diversity is through affirmative action. This is a way of achieving diversity that is merit-based. So it can supplement affirmative action [or] it can displace it. But it's another way of achieving the same goals."

But will schools, colleges, students, and parents accept such radical changes from the venerable test that has defined college ability for generations? Will the vast social and personal investment in the SAT permit a new, broader definition of testable merit? "People who have high SAT scores … may call it dumbing down," Sternberg said, "just as the wealthy elite in this country often felt that when lower social classes were allowed into schools like Yale, that was … pandering to the low classes." But fairness

is in the eye of the beholder. Despite the SAT's longstanding familiarity and prestige, Schmitt has data showing that minority students consider the test less relevant to college success and less fair than do white students. Minority students, on the other hand, "were more favorably disposed to our measures than the white group was," Schmitt added.

"In every society the elite think that whatever criteria they were chosen by are the ones God prefers," Sternberg said. But, "occasionally people do the right thing. The [original] SAT was some people doing the right thing, oddly enough. At that time they said, 'Your last name and what prep school you go to is not a sufficient basis.' … There were people in the power structure who were willing to change the way admissions were done, because they thought it was the right thing to do. I'm personally not totally cynical that you can't do things in a way that would actually result in some changes." Testing a wider range of abilities defines merit "in a way that matters for societal success. It's not … arbitrary, [just] saying 'Let's do a new merit system.' [It tests] in terms of what matters for society."