From *SIAM News*, Volume 45, Number 4, May 2012

What If the SAT Were Optional?

Uneducated Guesses: Using Evidence to Uncover Misguided Education Policies. *By Howard Wainer, Princeton University Press, Princeton, New Jersey, 2011, 200 pages, $24.95.*

BOOK REVIEW
By James Case

Howard Wainer's latest book is well described by its subtitle. Persuaded that many educational policies in the U.S. are misguided, and that even a passing glance at the facts will confirm as much, Wainer confronts a series of actual or proposed policies with hard data. Not surprisingly, in view of his 21 years as principal research scientist at the Educational Testing Service, many of the policies he addresses—and much of the data he employs—have to do with the administration and interpretation of standardized tests.

A growing number of educators favor a reduced role for the SAT in college admissions; some have proposed making it optional. At present, most selective institutions of higher learning encourage or require applicants to submit SAT scores, although many will accept rival ACT scores in their stead. An SAT score consists of a pair of numbers V, M satisfying $200 \le V, M \le 800$, which purport to measure the student's verbal (V) and mathematical (M) aptitude for college work. Often, the total $V + M$ is all that is reported, although the combination $2V + M$ is a somewhat more accurate predictor of a student's freshman GPA.

So what would happen if the requirement were relaxed? It turns out that Bowdoin College actually made the SAT optional some years ago, allowing Wainer to make a direct assessment of the consequences, using data pertaining to the entering class of 1999. His results are particularly valid because, as luck would have it, all the Bowdoin applicants for admission in the fall of 1999 who elected not to submit their SAT scores to the school did take the SAT, presumably because they were applying simultaneously to schools that required that test and no other. Wainer emphasizes that opportunities to make such direct comparisons are rare and should be fully exploited when they do occur.

Because the students who chose not to submit their SAT scores to Bowdoin did take the test, the decision to withhold their scores was presumably strategic, based on a conviction that their scores would not improve their chances of success. And because the ETS knew what the missing scores were, staff members were able to test the hypothesis that such was indeed the case. As might be expected, the 28% who entered Bowdoin that fall without submitting their SAT scores had performed less well on the test than their classmates. Their scores ran about 120 points lower, on average, and their freshman GPAs ran lower as well, averaging slightly less than 2.8 against something more than 3.2.

To further test the hypothesis that applicants who choose not to submit SAT scores are likely to perform less well in college than those who do, Wainer considered five schools whose entering classes had combined SAT scores ($V + M$) closely comparable to Bowdoin's 1288 and were willing to accept ACT in place of SAT scores. In the fall of 1999, the incoming freshman classes at those schools—Colby College, Barnard College, Northwestern University, Carnegie Mellon University, and the Georgia Institute of Technology—had average SAT scores between 1278 and 1338.

Each of the five schools accepted a substantial number of applicants who had submitted ACT in lieu of SAT scores, though they had taken both tests. Again, because the ETS knew the missing scores, a direct comparison could be made. Wainer summarizes the findings in a rather elaborate graph suggesting that students who withheld their SAT scores did significantly less well (50–125 points lower) than their classmates on the test, and compiled substantially lower freshman GPAs (about a fifth of a letter grade lower than those who submitted SAT scores). Making the SAT optional, Wainer concludes, seems to all but guarantee that it will be the lower-scoring students who withhold their SAT scores, and that those students will perform less well in their first year.

The correlation between aptitude scores and GPAs tends to diminish after the freshman year, in part because those who do poorly as freshmen tend to abandon more difficult curricula—in the physical sciences, engineering, and ancient languages—in favor of less demanding ones. As a proxy for the correlation between scholastic aptitude and classroom performance, Wainer chooses to examine the connection between students' scores on the Preliminary SAT with their success on various Advanced Placement tests administered by the ETS. There may be more telling ways to study that correlation, but the abundance of data at Wainer's fingertips argues for doing it his way.

The Preliminary SAT/National Merit Scholarship Qualifying Test (PSAT/NMSQT) is a program co-sponsored by the College Board and National Merit Scholarship Corporation. Each year more than 1.5 million high school juniors take the test, which offers useful practice for the SAT, along with an opportunity to compete for National Merit Scholarship money. A student's score again consists of a pair of numbers V, M, this time satisfying $20 \le V, M \le 80$. The test, says Wainer, is "made up of relatively easy retired SAT questions, and is dirt cheap to construct and administer."

Like most colleges and universities, Wainer defines success on an AP test as a score of 3 or better out of a possible 5, and describes the correlation between PSAT scores and scores on various tests. There is *no meaningful correlation* between PSAT scores and performance on the AP tests in German, Spanish, or Studio Art. He speculates that scores on the AP Spanish test are corrupted by the number of native Spanish speakers who take it merely for an easy grade, and that scores on the French and German tests would be similarly corrupted if taken by more native speakers.

The most revealing data Wainer could find on the matter, compiled by an ETS colleague in 1997,* are summarized in a three-column table. The middle column contains 26 numbers ranging from a low of 22.0 to a high of 63.2; the numbers represent the PSAT scores (either V or M) at which 50% of the students taking a particular AP test earned scores of 3 or more. If a listed number represents a V score, the name of the test for which it is predictive appears in the left-hand column. If the number represents an M score, the name of the associated test appears in the column on the right. The table reveals, for instance, that a V score of 42.4 confers a 50% chance of success on the European History test, while an M score of 63.2 is required to confer a similar probability of success on the Physics C test (electricity and magnetism). Curiously, *either* a V score *or* an M score of at least 48.2 suffices for an even chance of success on the Biology test. The numbers confirm what everyone already knew: Some subjects require more aptitude than others, and the aptitudes required for success on the quantitative tests are significantly rarer than the ones required for success in most other fields.

*W.J. Camara, *The relationship of PSAT/NMSQT scores and AP examination grades*, Research Note RN–02, The College Board, Office of Research and Development, New York, 1997.

Wainer devotes special attention to the M scores required for a probability p of success on the Calculus AB test, where p ranges from small positive values to almost 1. Because tens of thousands of students take the test each year, the probabilities in question can be estimated with considerable accuracy. A number of them are plotted in Figure 1. On the whole, the level of agreement between model probabilities and observed frequencies seems remarkable.

Wainer has little to say about elementary and junior high schools, in which the bulk of educational failures occur. As a result, the book is only indirectly about fixing the American educational system as a whole. An exception occurs in Chapter 9, in which Wainer describes attempts to estimate teacher effectiveness from student test scores. Over the last decade, a variety of procedures for doing so have been devised and adopted by various states. The term "value-added models" has come to denote them all.

One of the most widely used VAMs goes by the name Tennessee Value-Added Assessment System. If s_k denotes a given student's score in the kth year of testing, the system assumes in particular that

$$s_1 = \mu_1 + \theta_1 + \varepsilon_1 \text{ and } s_2 = \mu_2 + \theta_1 + \theta_2 + \varepsilon_2,$$

where μ_k denotes the district average in year k, θ_k represents a contribution from the student's teacher for that year, and ε_k is a random error term. The increment $s_2 - s_1$, conveniently independent of θ_1, can be interpreted as the "value added by teacher #2."

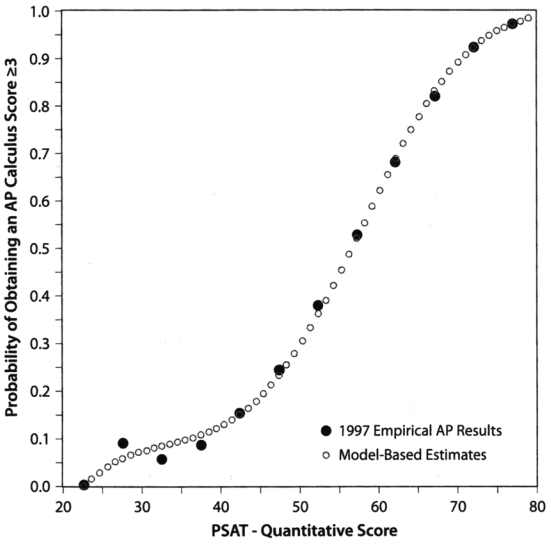

Figure 1. *Comparison of AP Calculus passing rates and model-based estimates. Data source: Camara, 1997.*

Currently available VAMs provide extremely simple answers to inordinately complex questions, Wainer argues, and can hardly be expected to do so reliably. He sees three main difficulties to be overcome by future generations of VAMs: the difficulty of establishing causality, the inevitability of missing data, and the variability of the tests themselves over time. As educators Albert Beaton and Rebecca Zwick have put it, "If you want to measure change, you should not change the measure." Wainer discusses each challenge in turn, and alludes to others as well.

All in all, Wainer has written an entertaining and insightful book about a subject few mathematical scientists have approached systematically.

James Case writes from Baltimore, Maryland.