ETS®

Listening. Learning. Leading.®

R & D Connections

No. 20 • September 2012

Why Bother with Research When We Have Common Sense?

By Brent Bridgeman[1]

Common sense: *Merriam-Webster®* defines common sense as "sound and prudent judgment based on a simple perception of the situation or facts," but common sense is often based on relatively narrow and simplistic assumptions about the world, as well as prejudice, which can lead to erroneous conclusions.

Tests can contribute to high-stakes decisions, and may help to open (or close) doors to a desired education or profession. At stake are personal interests and potential futures. We should therefore not be surprised that the testing and scoring process is surrounded by a range of ideas and suggestions, many claiming to be based on common sense … but much of what goes as common sense is actually based on common misunderstandings, urban legends, or outright nonsense.

If the simplistic assumptions underlying the common sense claims were right, then there would be no need for lengthy and/or expensive research studies. But are these claims correct? Let's take a look at six common claims about tests and compare them to what we have learned from research.

1. *More time for a test will lead to higher scores.* This claim states that students would get much higher SAT® scores if they were given substantially more time to complete the exam. More time would be especially helpful to students with weak scores.
2. *Typing on a computer gives higher scores than writing by hand.* This claim states that essays written on computers have a more professional look than those written by hand and therefore will get higher scores.
3. *SAT scores are weak predictors of college success.* This claim states that SAT scores are of very limited value as they only predict first-year grades and at best predict less than 10% of the variance in grades, once high school grades are taken into account.
4. *Admissions tests are biased against African-American and Hispanic students.* This claim states that the SAT test is biased against African-American and Hispanic students, and that they will therefore do better in college than the SAT scores predict.
5. *Don't change the answer.* This claim states that students should stick with their first response on a multiple-choice test, because changing answers usually results in a lower score.
6. *Machines cannot evaluate essays.* This claim states that machines can evaluate trivial aspects of essays such as the number of words written or certain types of

[1] Editor's note: *Brent Bridgeman is a Distinguished Research Scientist in the Center for Foundational and Validity Research in the Research & Development division of ETS.*

grammatical errors, but they cannot assess the quality of thinking or creativity in an essay. A machine will therefore never be able to predict future performance on a human-scored essay test as well as a human can.

These claims may sound convincing, but research shows that every one of them is false. Let's go through them one by one.

1. *Does Extra Time During the SAT Test Result in Substantially Higher Scores?*

Students with disabilities can receive extra time on the SAT test, and their performance may improve significantly with extra time. The focus of the research described here is to understand the benefits of extra time, if any, for students without a diagnosed disability.

One study suggests that even a generous extension of time limits during the SAT test — 1.5 times the ordinary time per question — produces only modest gains in test scores (Bridgeman, Trapani, & Curley, 2004). This study was conducted in the context of a regular SAT test administration, as this kind of research is most valid when students are highly motivated and doing their best. The testing time was manipulated by reducing the number of test questions in one of the regularly-timed sections. Standard directions for the SAT test inform students that one section on the test "helps us ensure the test is fair," and that "Questions in the unscored section are not factored into your SAT score," but students do not know which section is the unscored section. Hence, it is in students' interest to do their best on all sections.

Three versions of a 30-minute verbal section of the test were created in the study. One had the typical 35 questions, a second had 27, and a third had 23. The questions in the smallest group were also included in the sections with 27 and 35 question respectively, so that everyone would be compared based on the same questions. Scores from these 23 questions were then matched to the 200–800 SAT scale so that the differences would be easier to interpret on the familiar scale. The result of this study was that the extra time added less than 10 points.

A similar approach was used in the math section of the test, which on average boosted the result by about 20 points following an addition of a substantial amount of extra time per question. This may sound like a large number, but we must remember that the scale goes from 200 to 800 in 10-point increments. Hence, the time adjustment only added about two correct answers on a 60-question test.

This may be true for an average student, but couldn't a student who struggles under the standard time limit benefit from getting extra time?

This sounds plausible, but once again, we find that common sense can be incorrect. Research shows that extra time was of no benefit whatsoever for students who scored 400 or lower on the SAT test. Indeed, for two of the four math sections studied, adding extra time lowered the scores for these students.

> *"Research shows that extra time was of no benefit whatsoever for students who scored 400 or lower on the SAT test. Indeed, for two of the four math sections studied, adding extra time lowered the scores for these students."*

But why would extra time hurt some students?

The answer has to do with how the SAT test is designed and scored. The most difficult questions are usually placed near the end of the test, which means that low-performing students often run out of time before they get a chance to answer them. With extra time, they attempt to answer these questions, but tend to get them wrong. The way the SAT test is scored, it is better to leave a question blank than to provide an incorrect answer, so extra time can actually result in a lower score.

The research we discussed above relates to the version of the SAT test that existed in 2003. The current version of the SAT test has more generous time limits, so extra time would likely have even less impact now.

2. Do Essays Written on a Computer Get Higher Scores Than Those Written by Hand?

Part of the answer can be found in an article published in a scientific journal in 1994: *Will They Think Less of My Handwritten Essay If Others Word Process Theirs? Effects on Essay Scores of Intermingling Handwritten and Word-Processed Essays* (Powers, Fowles, Farnum, & Ramsey, 1992).

The article describes how examinees who participated in a pilot study for a set of teacher assessments wrote essays both by hand and with a word processor. The essays originally written with a word processor were rewritten by hand and the original handwritten essays were converted to word-processed versions. These essays were then intermingled and scored. The result went against what many would consider common sense: The handwritten versions consistently received higher scores than the word-processed versions, and this was true regardless of the mode in which the essay was originally written.

We are not certain why the handwritten essays received higher scores, but one reasonable hypothesis is that raters tend to be more forgiving of minor errors in a handwritten essay. They don’t expect a handwritten essay to be perfect, while they expect word-processed essays to be essentially error free. Another reason could be that it is harder to detect certain errors in a handwritten essay. Sloppy handwriting can hide minor spelling errors or a misplaced comma, while such errors are very apparent in a word-processed essay. As with many surprising findings, the result looks perfectly predictable in hindsight, and this is why we need research in order to appropriately focus our hindsight. When raters were made aware of this tendency to be more lenient with handwritten essays, they were able to greatly reduce, although not totally eliminate, the discrepancy between the scores given to handwritten and word-processed essays.

3. Are SAT Scores Practically Worthless for Predicting How Well a Student Will Do in College?

Research from ETS and elsewhere has shown that SAT scores predict more than just the grade point average (GPA) in the first year of college. It predicts the four- or five-year cumulative GPA as well as the freshman GPA (Bridgeman, Pollack, & Burton, 2008; Mattern & Patterson, 2011). Furthermore, SAT scores predict more than just GPAs. It predicts the likelihood of a student dropping out of college (Mattern & Patterson, 2011) as well as his or her chances of graduating (Godfrey & Matos-Elefonte, 2010).

But even if SAT scores predict all of these outcomes, some would argue that it is still of very limited value, because its contribution to the prediction is so small. Critics claim that SAT scores are essentially of no value, because they explain less than 10% of the variance (variability between high and low numbers) in grades over and above what can be predicted from high school grades alone. It may be that 10% of anything sounds trivial, but few really understand what this means — even among trained social scientists.

Let us visualize it as a percentage of people instead of a percentage of a variance, which is what Bridgeman, Pollack, and Burton (2004) did in a study. They constructed a sample of more than 16,000 students from a diverse group of 41 colleges, which were placed into four categories based on the average SAT scores of their incoming freshmen. The students were then divided into categories based on three separate indicators:

- the rigor of the curriculum in the high schools the freshmen had attended;
- their high school grade point average (HSGPA); and
- the SAT score (combined verbal and math[2]).

The study defined college students whose college GPA after four years of college was 3.5 or higher as highly successful.

All three indicators were strong predictors of a high degree of success in college when considered separately, *but how valuable would the SAT scores be once the rigor of the high school curriculum and the HSGPA were taken into consideration?*

The full report slices the data in many different ways, but all we need to do here is to look at the results from just one slice. Let's look at a narrow slice of students from colleges at a similar level (with average SAT scores in the 1,110 to 1,195 range, a solid college-prep high school curriculum, but without many AP® courses, and a HSGPA above 3.7), and ask whether the SAT scores could provide valuable additional information beyond what we can expect from their level and grade point average. It turns out that the SAT score is a good predictor for a student's success at college. Only 6% of the students in the 800–1,000 SAT score range at this group of colleges would be highly successful — i.e., reach an average GPA of at least 3.5 after four years at college. However, 59% of students with a SAT score of at least 1,400 would be highly successful. From this we can conclude that even though SAT scores are criticized for explaining less than 10% of the variance in college grades, a difference of 53% in the proportion of highly successful students predicted by the SAT scores is clearly important.

4. *Do Minorities Perform Better in College Than Predicted by Their SAT Scores?*

There is no evidence that the SAT test is biased against minorities. It is certainly true that there are mean score differences among various groups, and that scores from White and Asian-American students generally tend to be higher than scores from African-American and Hispanic students. But score differences are not the same as bias. We would, for example, not conclude that tape measures are biased against women even though men on average measure taller than women. Neither would

[2] This study was completed before the writing score was added to the SAT test.

we call a thermometer biased if it indicated that people with the flu had a higher temperature than those who didn't have it. Likewise, we should not say that a test is biased because it showed that students from education-friendly environments (e.g., who attended top schools with highly qualified teachers, had more books at home, and rarely moved from one school to another) tend to get higher scores in reading, writing, and mathematics than less-fortunate students.[3] On the contrary — a test that failed to reflect these educational realities would be suspect.

A key question for a test used for college admissions is whether African-American and Hispanic students tend to do better or worse in college than their SAT scores would predict. There have been hundreds of studies — some with very large samples — and they tend to indicate that these minority groups are likely to do worse than predicted by their SAT scores. For example, Mattern et al. (2008) studied more than 150,000 students from 110 colleges. They used GPA as a criterion, and predicted these outcomes either from SAT scores alone or the combination of SAT scores and high school grades. They followed the typical procedure in this kind of work by making a prediction based on all students at a particular college and then checking whether the minority groups got higher or lower grades than had been predicted. Although the predictions are made separately for each college, the results are accumulated across all 110 colleges. The findings were consistent with previous large-scale studies (e.g., Bridgeman, McCamley-Jenkins, & Ervin, 2000; Ramist, Lewis, & McCamley-Jenkins, 1994), and showed that African-American and Hispanic students did worse in college than would be predicted from SAT scores alone, from the high school grades alone, or from a combination of SAT scores and high school grades.

We should not forget that this kind of research, indeed most research, reports results in terms of averages. Many individuals will defy these averages and do much better (or worse) in college than would be predicted from just looking at their high school grades or SAT scores.

5. Is the First Answer Always the Best?

Many students believe that it is best to stick with the first answer when given a chance to review and change answers on a multiple-choice test. Many test takers remember with regret that they changed an answer that turned out to have been correct, but it is easy to forget the times when they changed an incorrect answer to a correct one. This has more to do with psychology than the facts of test taking.

"Many test takers remember with regret that they changed an answer that turned out to have been correct, but it is easy to forget the times when they changed an incorrect answer to a correct one. This has more to do with psychology than the facts of test taking."

Evidence from research shows overwhelmingly that students are much more likely to get a higher score when given the opportunity to review and change answers, compared to a situation where they can't change their answers (e.g., Benjamin, Cavell, & Shallenberger, 1987; McMorris, DeMers, & Schwarz, 1987).

The clear message to test takers is that *they should carefully review their answers at the end of a test, and not be afraid to correct them* if the initial answer appears to be wrong.

[3] For a discussion of 16 factors, ranging from low birth weight to teacher quality, that are correlated with academic performance and that tend to differ between White students and African-American students, see *Parsing the Achievement Gap II* (**http://www.ets.org/research/policy_research_reports/pic-parsingii**) by Barton & Coley (2009).

6. Can a Machine Score Essays as Well as a Human Can?

Computers can score many aspects of an essay. ETS's *e-rater*® computer essay-scoring system looks at dozens of specific features in the following broad categories:[4]

Grammar – Based on rates of errors such as fragments, run-on sentences, garbled sentences, subject-verb agreement errors, ill-formed verbs, pronoun errors, missing possessives, and wrong or missing words

Usage – Based on rates of errors such as wrong or missing articles, confused words, wrong form of words, faulty comparisons, and preposition errors

Mechanics – Based on rates of spelling, capitalization, and punctuation errors

Style – Based on rates of cases such as overly repetitious words, inappropriate use of words and phrases, sentences beginning with coordinated conjunctions, very long and short sentences, and passive-voice sentences

Organization and development – Based on the number of discourse elements (e.g., main ideas, supporting elements, concluding paragraph) and the number of words per element

Vocabulary – Based on frequencies of essay words in a large corpus of text

Word length – Average word length

These features cover important aspects of an essay-writing assessment, but they are certainly not the whole story. The machine — a computer — cannot evaluate the logic or persuasiveness of an argument. It could deliver a high score for an entirely illogical essay if it is grammatically sound and uses sophisticated vocabulary. It would therefore seem reasonable to assume that a machine can never agree with a human rater as well as one human rater could agree with another. But research evidence shows that it is possible. It is a common finding in large-scale essay assessment programs that a machine and a human agree more often on an assigned essay score than do two humans (Bridgeman, Trapani, & Attali, 2012).

Two factors are likely to account for the higher rate of agreement between humans and machines.

First, the machine is extremely consistent in what it is doing, and it never gets tired or skips over sentences accidently. Given the same essay, it will produce exactly the same score every time (which is not true for human raters).

Second, although there is no necessary connection between what the machine can and cannot evaluate (e.g., between grammar and the logic of an argument), these things tend to be highly related. It is rare to find an extremely well-written essay from the standpoint of grammar, usage, mechanics, style (and other aspects the machine can evaluate) that is very illogical in its argumentation. Because of this, the machine and human scores tend to be highly related.

On rare occasions, human and machine scores will disagree substantially on an essay, just as two human raters may disagree on an essay score. In such cases, an additional human rating would be obtained. However, when evaluating writing

> *"However, when evaluating writing quality, it is generally better to rely on a combination of human and machine ratings than on a single human rater. Such a mix of human and machine raters is actually better than relying on two humans."*

[4] The following is adapted from Attali & Powers, 2008.

quality, it is generally better to rely on a combination of human and machine ratings than on a single human rater. Such a mix of human and machine raters is actually better than relying on two humans. This was confirmed in a study of English-language learners taking the *TOEFL iBT*® test, a test of English-language proficiency in which the writing score is based on two essays. Scores for two essays from one administration of the test (Time 1) were used in the study to predict scores from a different pair of essays at a later administration (Time 2) (Bridgeman, Trapani, & Williamson, 2011). Essays scored at Time 1 were scored either by a combination of human raters and the computerized *e-rater* scoring engine, or by two humans. For the second administration, two human raters each scored a second pair of essays from an examinee. Each examinee's writing score (criterion) was based on a combination of the four scores. This criterion was predicted more accurately from essay scores at Time 1 based on both human and machine raters, compared to when two human raters were used at Time 1. It appears that humans and machines have different strengths in evaluating essays, and that the combination of them provides better measurement than if we rely on either one of them alone.

"What we need instead is rigorous research that takes us beyond presumptions and hunches that on closer investigation often turn out to be false."

Conclusion

It is easy to jump to conclusions about tests and scoring relying on intuition, common sense, or urban legends, but it is likewise easy to jump to the wrong conclusions. It doesn't matter if it seems that "everybody knows" something. What we need instead is rigorous research that takes us beyond presumptions and hunches that on closer investigation often turn out to be false. In hindsight, it is easy to see that more time to answer questions is of little help to students who have no idea of how to answer them. It will seem obvious that (a) raters are more lenient when reading handwritten essays, as they treat them as just drafts written under tight time constraints, and that (b) your result will be better if you are allowed to fix your apparent errors on a test. However, we will — until foresight trumps hindsight — have to rely on research for the answers to such questions.

References

Attali, Y., & Powers, D. (2008). *A developmental writing scale*. (ETS-RR-08-19). Princeton, NJ: Educational Testing Service.

Benjamin, L. T., Cavell, T. A., & Shallenberger, W. R. (1987). Staying with initial answers on objective tests: Is it a myth? In M. E. Ware & R. J. Millard (Eds.), *Handbook on student development: Advising, career development, and field placement* (pp. 45–53). Hillsdale, NJ: Lawrence Erlbaum.

Bridgeman, B. (2012). Eighty years of research on answer changing is NOT wrong: Reply to van der Linden, Jeon, and Ferrara. *Journal of Educational Measurement, 49*, 216–217.

Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade-point average from the revised and recentered SAT® I: Reasoning Test.* (College Board® Research Report No. 2000-1). New York: The College Board.

Bridgeman, B., Pollack, J., & Burton, N. (2004*). Understanding what SAT Reasoning Test™ scores add to high school grades: A straightforward approach*. (College Board Research Report No. 2004-4 and ETS RR-04-40). New York: The College Board.

Bridgeman, B., Pollack, J., & Burton, N. (2008). *Predicting grades in different types of college courses*. (College Board Research Report 2008-0; ETS RR-08-06). New York: The College Board.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25,* 27–40.

Bridgeman, B., Trapani, C., & Curley, E. (2004). Impact of fewer questions per section on SAT I scores. *Journal of Educational Measurement*, *41*, 291–310.

Bridgeman, B., Trapani, C., and Williamson, D. (April 2011). *The question of validity of automated essay scores and differentially valued evidence.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Godfrey, K. E., & Matos-Elefonte, H. (2010). *Key indicators of college success: Predicting college enrollment, persistence, and graduation*. Paper presented at the annual meeting of the American Educational Research Association, Denver.

Mattern, K. D., & Patterson, B. F. (2011). *Validity of the SAT® for predicting fourth-year grades: 2006 SAT validity sample*. (College Board Statistical Report No. 2011-4). New York: The College Board.

Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. M. (2008). *Differential validity and prediction of the SAT®*. (College Board Research Report No. 2008-4). New York: The College Board.

McMorris, R. F., DeMers, L. P., & Schwarz, S. P. (1987). Attitudes, behaviors, and reasons for changing responses following answer-changing instruction. *Journal of Educational Measurement*, *24*, 131–143.

Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement, 31*, 220–233.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups*. (College Board Research Report No.93-1). New York: The College Board.

R&D Connections is published by

ETS Research & Development
Educational Testing Service
Rosedale Road, MS 19-T
Princeton, NJ 08541-0001
email: **RDWeb@ets.org**

Editor: Hans Sandberg
Copy Editor: Eileen Kerrigan
Layout Design: Sally Acquaviva

Visit ETS Research & Development on the web at **www.ets.org/research**

Follow ETS Research on Twitter®
(@ETSresearch)