The SAT: A review of one of the most prominent achievement assessments

Lisa Klein Weintraub

Wake Forest University

Abstract

The SAT, originally called the Scholastic Aptitude Test, has gone under several revisions since its debut in 1926. The most recent version "is a 4-hour, primarily multiple-choice test with four major sections: (a) Reading, in which individuals read passages and interpret graphics to show their understand of the passage and its context; (b) Writing and Language, in which individuals review a passage and identify and fix mistakes or weaknesses in the passage; (c) Math, in which individuals solve algebraic functions, problem-solve and analyze data, and manipulate test item information to solve complex equations; and (d) Essay in which individuals read a passage and develop an argument to persuade an external audience with ample evidence of that argument" (Hays, 2017, p. 205). Educational Testing Service (ETS) a private nonprofit organization devoted to educational measurement and research is the developer and administrator of the SAT. The College Board sponsored the SAT and other testing programs and decides how they will be constructed, administered and used. As Hays describes in her book, *Assessment in Counseling Procedures and Practices*, "aptitude is generally though of as an ability to acquire a specific type of skill or knowledge; aptitude tests are typically used for prediction purposes" (Hays, 2017, p. 204). It has been posited that academic or scholastic aptitude is significantly related to achievement in various educational programs in high school, colleges, and professional schools. Many professional educators and statisticians challenge the view that the SAT adds substantial predictive value to the success of students in college that is not demonstrated through individual high school performance. The purpose of this paper is to describe strengths and limitations of the SAT and evaluate the predictive validity of the assessment. The procedure used to obtain

the articles utilized in this review is described first. Followed by a comprehensive review of the benefits and disadvantages of the SAT. Reviews are presented to enhance the current review and advise assessment of the evidence. This paper ends with suggestions of potential opportunities for research that may increase the efficacy of the SAT and similar aptitude tests.

The SAT: A review of one of the most prominent achievement assessments

The development of the SAT began with the birth of the College Entrance Examination Board (a.k.a. College Board) in 1899. The College Board set out to develop a standardized test that could measure general intelligence for purposes college admission, in 1925 Carl Brigham, was appointed to head this task. During World War I Brigham served as an assistant in development of the Army Alpha exam, which provided much of the groundwork for the original SAT. The first version of the SAT was administered in 1926 and is nearly unrecognizable when looking at the most recent version of the assessment that was released in 2016. Over nearly a century of administration the SAT has undergone several revisions and each redesign took a variety of considerations in to account, "including fairness issues, scaling issues, cost, public perception, face validity, changes in test-taking populations, changes in patterns of test preparation, and changes in the college admissions process" (Lawrence et al., 2003, p. 1). Despite several revisions the division of the SAT's supporters and critics has deepened and strengthened over the past two decades.

In 2003, Roy Freedle, a cognitive psychologist who had work at the Educational Testing Services (ETS) for more than thirty years published a piece in the Harvard Educational Review asserting that "many of the more difficult SAT items exhibited

differential item functioning (DIF) benefiting African American students, while easier SAT items showed DIF favoring White students" (Santelices & Wilson, 2010, p. 108). Freedle's publication ignited many researchers to investigate the findings that Freedle had made and the rebuttals issued by researchers of ETS. Joseph Soares, Wake Forest University Professor and author of *SAT Wars: The Case for Test-Optional Admissions*, argues that the SAT adds little to High School grade point average when it comes to being able to predict college grades and "furthers social disparities unfavorable to racial minorities, women and low SES youths" (Soares, 2012, p. 6). On the opposite spectrum Hays' writes that "A national college admission test represents a common task for all students and therefore can operate as a correction factor for the high school GPA; additionally for the student with low grades but substantially higher scores than would be expected from those grades, the scores may suggest unrecognized academic potential" (Hays, 2017, p. 207).

Despite the critics of the SAT it is evident that standardized tests are frequently used in the United States and abroad as a basis for making high-stakes decisions about educational opportunities, placements, and diagnoses. "A recent meta-analysis of the predictive validity of the SAT, encompassing roughly 3000 studies and more than one million students, suggested that the SAT is a valid predictor of early-college academic performance (as measured by first-year grade point average (GPA) with validity coefficients generally in the range of .44 to .62" (Sternberg, 2006, p. 322). Sternberg writes that while the SAT has valuable predictive ability there is room for improvement for not only predictive ability but also establishing better group equity.

Method

For my research, I conducted a search for peer reviewed and evidenced based information related to the SAT through the Z. Smith Reynolds Library website. With in the Wake Forest University library website I performed searches through the databases and the ERIC database. I also utilized research publications on the College Board's website. Phrases and keywords like "Scholastic Aptitude Test", "history of Scholastic Aptitude Test", "Disparities Scholastic Aptitude Tests", and "Scholastic Aptitude Test Measures" returned a vast amount of sources associated with the subject. I did not place any limitations on publishing dates when searching but only selected sources published from 2003 to 2017. The ZSR Library database produced the greatest number or related articles all of which were accessible instantly through the database.

The College Board and Educational Testing Services

In February of 2013, the College Board announced that it would undertake a redesign of the SAT in order to develop an assessment that better reflects the work that students will do in college. With the redesign of the assessment College Board conducted a Pilot Predictive Validity Study "in order to examine the relationship between scores between the redesigned test with college outcomes such as first-year grade point average and college course grades" (Shaw et al., 2016, p.6).

This study included over 2,000 college students who had previously taken the SAT in order to make comparisons between the students original scores/redesigned scores and predictive ability of the redesigned SAT in relation to first year GPA. The Validity Study showed that the "correlations coefficients (corrected for restriction of range) representing relationships between admission and test scores and performance in

college or graduate school tend to be in the .40s and .50" (Shaw et al.,2016, p. 12). Further the study showed that contributors to the study show that "the correlations between high school grade point average (HSGPA) and SAT scores with FYGPA are large with strongest relationship represented by the multiple correlation of both HSGPA and SAT together (.58); while the multiple correlation of the SAT verbal and math sections together with FYGPA is .53, and HSGPA alone and FYGPA is .48" (Shaw et al., 2016, p. 13).

While this study does indicate that including both HSGPA and SAT scores to predict Freshman Year Grade Point Average is a better predictor together than either measurement is separately the study reviewed is merely a pilot study. Questions of differential item functioning and bias cannot be determined until a study can be done with a large, nationally representative sample as the redesigned test launched and relationships can be studied between various groups and long term outcomes including persistence, completion, a cumulative GPA.

A Call for Equality in Testing

When Freedle wrote about the disparities he observed in differential item functioning of the SAT many students, parents, and educators were outraged. Santelices and Wilson conducted a study to evaluate the validity of Freedle's findings stating:

"*Regardless of content modifications in recent years, identifying bias in past SAT tests is still critically important, give the role of the SAT in admission to selective institutions of higher education and the role institutions play in dispensing rewards and benefits to members of our society. Fairness is a critical part of test validity, and test should be valid when individual consequences are attached to them. Admissions*

decisions that rely heavily on a test that is not a valid measure for all subgroups of the population raise critical questions of fairness" (Santelices & Wilson, 2010, p. 109).

Santelices and Wilson explain that the goal of test developers in to create tests that have no differential item functioning between groups "since, by definition, these differences in performance are irrelevant to the construct measured and render test results invalid" (Santelices & Wilson, 2010, p. 107). Freedle's report showed a correlation of .50 between the DIF statistics he used and the difficulty of the items. ETS countered Freedle's findings by stating that phenomenon could be resulting from differential speediness, differential guessing strategies among examinees of different racial or ethnic groups, and or methodological issues. Freedle suggested that the reason for the DIF was due to the role of linguistics and cultural differences.

Santelices and Wilson's findings between item difficulty and DIF estimates for White/African American comparison of verbal items falling in the same correlation range as reported by Educational Testing Services. Although the results found were lower than .50 as Freedle had described they were in the range .20 and .40 still showing a relationship between DIF and item difficulty on the verbal test (Santelices & Wilson, 2010, p. 125-126). Santelices and Wilson conclude "although not generalizable to all ethnic subgroups and to all item types, these findings are strong enough to question the validity scores for African American examinees and consequently admission decision based exclusively or predominantly on these scores" (Santelices & Wilson, 2010, p. 128).

Some individuals suggest that "stereotype threat" the threat of being viewed through the lens of a negative stereotype, or the fear of doing something that would inadvertently confirm that stereotype—produces stress, which ultimately leads students to

perform more poorly. However, Zwick and Sklar state that " if stereotype threat depresses standardized test performance but does not affect subsequent academic work, it would be expected to lead to underprediction (of student success) because affected students would preform better in college than their depressed scores would indicate" (Zwick & Scalar, 2005, p. 443). When research shows that FYGPA is commonly overpredicted among black and Latino groups. Zwick and Sklar also state write that "although high-school GPA is usually more highly correlated with FYGPA than are the SAT scores, overprediction tends to be worse if only high school GPA is included in the prediction equation" (Zwick & Sklar, 2005, p. 442); which would appear despite the flaws of the SAT make a case for the assessment to be continued to be used to make admission decisions.

Soares takes the argument against using the SAT a step further by providing research in his book *SAT Wars: The Case for Test-Optional College Admissions* that in addition to the social disparities calcified by the SAT independent scholars "found that neither the SAT nor the ACT adds more than a few percentage points to what is already known from high school GPA" (Soares, 2012, p. 6). Soares also highlights the fact that family income and parental education correlates with test scores but does not correlate with grades earned in high school suggesting that the higher a college's average scores the more economical advantage the next year's applicant pool with be which keeps top tier schools with very affluent applicants. Soares suggests that schools opt to become "test-optional" schools and today many have. "Test-optional" schools meaning that taking the SAT is not a requirement for admission; however, these schools often still require that students submit SAT scores after acceptance in order to collect data on the

performance of test-optional students vs. traditional applicants who include scores initially. Soares reports that since Wake Forest University's decision to become a test-optional school in May 2008 the applicant pool for the following academic year saw a "minority increase by 70%, prior to the policy change 6% of the Wake Forest senior cohort were minorities of color; those admitted thus far as test optional have increased the Black and Hispanic group to 23%, Asian students to 11%, first generation college students to 11% and Pell Grant youths doubled to 11%" (Soares, 2012, p. 10). Soares also writes that WFU non-test-score undergraduates perform equally as well as test score submitters and that "the percent of students from the top 10% of their high school classes has gone up dramatically" (Soares, 2012, p. 11).

A Middle Ground: The Rainbow Project

The Rainbow Project utilizes Sternberg's theory of Successful Intelligence as a basis to provide additional assessment of methodical skills, as well as tests of applied and creative skills, to supplement the SAT in forecasting college performance. Sternberg defines "successful intelligence in terms of the ability to achieve success in life in terms of one's personal standards, with in one's sociocultural context" (Sternberg, 2006, p. 323). Many scholars are familiar with Binet and Simon's original operationalization of intelligence as the skills one needs for success in school, Sternberg provides an argument for this strict academic operationalization of intelligence by countering that while abilities needed to succeed in school are an important part of intelligence and important to the workforce *it is not all there is*. As such, Sternberg believed that tests of intelligence ought to be broadened to include three aspects of successful intelligence including analytical thinking, practical thinking, creative thinking, or a combination of the above.

Sternberg conducted a study that applied the theory of successful intelligence to the creation of assessments that measure analytical, creative, and practical skills. Findings revealed that "triarchic measures alone approximately double the predicted amount of variance in college GPA when compared with the SAT alone" (Sternberg, 2006, p. 344) making a compelling case for furthering the study of the measurement of logical, creative, and applied skills for predicting success in college. Perhaps of equal importance is the fact that "although the group differences in tests were not reduced to zero, the tests did substantially attenuate group differences relative to other measures such as the SAT" (Sternberg, 2006, p.346); this discovery could be an significant step towards safeguarding fair treatment for members of diverse groups in the academic sphere.

Discussion

From a review of the literature it is evident, that the consistently revised SAT still presents a great challenge to developers, administrators, and post-secondary institutions in regard to fairness and predictive ability. Although programs like the Rainbow Project have been studied and implemented in a pilot fashion the SAT is still largely utilized for making admissions decisions along with other high-stakes testing that does not take other facets of intelligence into account. Intelligence is a complex concept because there is arguably no one true definition and the meaning of success is completely individual. Given that there is still room for improvement in validity of the SAT and the results of the Rainbow Project it is important to continue exploring new assessments for future use in college admissions. The Rainbow Project, study should be replicated with larger sample sizes, diverse representation of the population, to test the generalizability of the initial results.

References

Hays, D. G. (2017). *Assessment in counseling: Procedures and practice s*(6th ed.). Alexandria, VA: American Counseling Association.

Lawrence, I. M., Rigol, G. W., Essen, T. V., & Jackson, C. A. (2003). A Historical Perspective on the Content of the SAT. *College Board Publications,*1-19. doi:10.4324/9780203463932_a_historical_perspective_on_the_content

Santelices, M. V., & Wilson, M. (2010). Unfair Treatment? The Case of Freedle, the SAT, and the Standardization Approach to Differential Item Functioning. *Harvard Educational Review,80*(1), 106-134. doi:10.17763/haer.80.1.j94675w001329270

Shaw, E. J., Marini, J. P., Beard, J., Shmueli, D., Young, L., & Ng, H. (2016). The Redesigned SAT Pilot Predictive Validity Study: A First Look. *College Board Publications,*1-21. Retrieved September 28, 2018.

Soares, J. A. (2012). For Tests that are Predictively Powerful and Without Social Prejudice. *Research & Practice In Assessment,7*, 5-11. Retrieved September 28, 2012.

Sternberg, R. J. (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence,34*(4), 321-350. doi:10.1016/j.intell.2006.01.002

Zwick, R., & Sklar, J. C. (2005). Predicting College Grades and Degree Completion Using High School Grades and SAT Scores: The Role of Student Ethnicity and First Language. *American Educational Research Journal,42*(3), 439-464. doi:10.3102/00028312042003439