The Flagging Test Scores of Individuals with Disabilities Who Are Granted the Accommodation of Extended Time: A Report of the Majority Opinion of the Blue Ribbon Panel on Flagging

Noel Gregg[12]
University of Georgia

Nancy Mather[3]
University of Arizona

Sally Shaywitz[4]
Yale University

Stephen Sireci[5]
University of Massachusetts Amherst

[1] The ordering of the authors is alphabetical

[2] . Noel Gregg, Ph.D., is a Distinguished Research Professor and Director of the Regents Center for Learning Disorders, University of Georgia.

[3]. Nancy Mather, Ph.D., is an Associate Professor, Department of Special Education, and University of Arizona.

[4] Sally Shaywitz, M.D., is a Professor of Pediatrics and Co-Director of the Yale Center for the Study of Learning and Attention, Yale University School of Medicine.

[5] Stephen Sireci, Ph.D., is an Associate Professor and Co-Director of the Center for Educational Assessment, School of Education, and University of Massachusetts Amherst.

Flagging scores on the College Board Scholastic Aptitude Test (SAT I) has been challenged as a discriminatory practice specifically penalizing students with learning disabilities. As part of the settlement of the Breimhorst vs. ETS litigation (N.D. Cal, March 27, 2001), the Educational Testing Service (ETS) and the plaintiffs agreed to convene a Blue Ribbon Panel of experts to consider issues relating to the flagging of scores on College Board standardized tests administered with extended time. A panel was chosen consisting have a non-voting chair (Edelman) and six members (Brennan, Gregg, Mather, Saleh, Shaywitz and Sireci). Two members of this panel were designated as the Psychometric Committee (Brennan and Sireci).

Evidence from both parties (i.e., College Board and Disability Rights Advocates) was presented to the full Panel with written supporting documentation. Wayne Camara, Ph.D. (College Board) and Kurt Geisinger, Ph.D. (Disability Rights Advocates) presented to the Panel and provided copies of their transparencies on November 19, 2001.

After reviewing documents submitted to the Panel from Disability Rights Advocates and the College Board, as well as, the evidence provided on November 19, 2001, the Panel met again on March 3, 2002 to determine its position. The majority position of the Panel was to discontinue the practice of flagging the SAT I based on scientific, psychometric, and social evidence. While concern for the integrity of the SAT I is laudable, the Panel determined that this legitimate concern should not result in a bias against applicants taking the SAT I with extended time when scientific, psychometric, and social evidence challenge the continued practice of flagging. The decision of the Majority was determined by considering the multidimensionality of the issue involved. As evidenced by the selection of the panel, representing a wide range of disciplines,

The decision on whether to flag the SAT I for students with disabilities requires the integration of scientific, psychometric and social/ethical evidence.

Scientific Evidence

Compelling evidence to discontinue the practice of flagging comes from current scientific empirical-based research. The vast majority of students with learning disabilities are those with reading disabilities or dyslexia. There is strong evidence that students do not outgrow a reading disability; it is a persistent and chronic problem. Furthermore, there is growing and converging scientific data indicating that students with reading disabilities become increasingly more accurate in reading as they progress in school, but that they continue to remain slow readers. Accumulating neurobiologic evidence demonstrates a functional disruption in children, university students, and adults with reading disabilities in those specific neural systems responsible for fast, automatic reading. Thus, there is epidemiologic evidence of the persistence of reading disabilities, as well as, behavioral and biological validation of the lack of fluency and the need for extra time. Reading fluency is the single best discriminator of college students with and without disabilities, and lack of fluency (slow, effortful reading) characterizes persons with reading disabilities at all ages. The individual with reading disabilities' fundamental need for extra time is further demonstrated by data indicating that while students with reading disabilities show a significant increase in test scores with extra time, other, nondisabled students do <u>not</u> demonstrate such a significant increase. All of these pieces of evidence come together to indicate that in order to be treated fairly and equally, and to have opportunities to pursue higher education, students with reading disabilities must have the accommodation of extra time. As there is strong scientific evidence that fluency is

the core of the disability for the majority of students with learning disabilities, to require flagging of this needed accommodation and no other accommodation, discriminates against a specific group of individuals. In this context, flagging amplifies stereotypes, discourages students from applying for needed accommodations, and represents a profound and artificial barrier preventing students with disabilities, most often those with learning disabilities, from equal access to colleges and future careers.

Psychometric Evidence

The *Standards for Educational and Psychological Testing* developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (AERA/APA/NCME, 1999) was used by the Panel to explore issues of test reliability and validity when non-standardized administration (i.e., extended time) is provided to one population of test users. Standards 10.1, 10.4, and 10.11 are the key benchmarks used in the development of the Majority's argument to discontinue the practice of flagging.

> Standard 10.1: "*In testing individuals with disabilities, test developers, test administrators, and test users should take steps to ensure that the test score inferences accurately reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement*" (p. 106).

If the College Board, as it states, does not consider the SAT to be a measure of speed, then speededness is a source of construct-irrelevant variance that threatens valid test score interpretation. Furthermore, if the SAT were not speeded, then inferences drawn from both timed

and untimed administrations would be similar. However, if speed were a factor affecting test performance, then higher scores would be expected under the conditions of extended time.

Bridgeman, Curley, and Trapani (2001) found that increasing the time limits on the Verbal SAT would result in a 5-to-10-point score increase on average for "standard" examinees. Increasing the time limits on the Math SAT would result in a 20-point increase for this same group. These results suggest that there appears to be a degree of speededness on the SAT. Camara, Copeland, & Rothchild (1998) investigated SAT score gains for students who took the SAT in both their junior and senior years of high school. They found that when students with disabilities took the test with extended time they had average score increases of 45 and 38 points for the Verbal and Math sections, respectively, compared to their scores when taking the tests under standard time limits. For non-disabled students who took the test twice under standard time limits, the score gains were smaller (13 and 12 points for the verbal and math sections, respectively). The results indicate that students with disabilities make substantially larger gains when taking the test with extended time, relative to non-disabled students. This finding is consistent with the argument that students with disabilities need more time to demonstrate their knowledge, skills, and abilities, than the non-disabled student, and suggests that the scores of these students taken under the condition of extended time are more representative of their true performance than are the scores they would obtain from a standard administration.

Therefore, scores for untimed administration of the SAT I do make a significant difference regarding inferences drawn about the aptitude for the population with learning disabilities as compared to the non-disabled (Camara, Copeland & Rothschild, 1998). Evidence provided by the College Board documented that extended time benefits students with learning disabilities even

more so on the verbal section. This is extraordinarily important since many admissions offices weight the verbal section of the SAT I twice as much as the math sections.

> Standard 10.4: *If modifications are made or recommended by test developers. . .(unless) evidence of validity for a given inference has been established for individuals with the specific disabilities, test developers should issue cautionary statements in manuals or supplementary materials regarding confidence in interpretations based on such test scores."*(p. 106).

Review of evidence provided by the College Board for SAT taken with and without extended time shows that the tests have comparable reliability and validity. Specifically, tests administered under extended and standard time demonstrate comparable standard errors of measurement (SEMs), an accepted measure of reliability (Rock, Bennett, Kaplan & Jirele, 1988, pp. 89-90, 95). Tests administered under standard and extended time conditions both show similar factor structures as determined by factor analysis, supporting construct validity. Together, these two measures indicate that the construct being measured is comparable. The *1999 Standards* states, " analysis of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (p. 13). While much of research available to the Panel pertained to the SAT rather than the SAT I, Brennan (2001) did not feel there would be substantial differences for the SAT I related to reliability and validity.

> Standard 10.11: *When there is credible evidence of score comparability across regular and modified administrations, no flag should be attached to a score. When such*

evidence is lacking, specific information about the nature of the modification should be provided, if permitted by law, to assist test users properly to interpret and act on test scores" (p. 108).

The issue of predictive validity, the degree to which the SAT I predicts college grade point averages (CGPA), is key to dismantling the practice of flagging. Cahalan, Mandinach, and Camara (2001) provided the panel with the most recent studies of predictive validity for the SAT I when used with students with learning disabilities under extended time accommodations. These data showed that the standardized residual from predicting first year college grades from the current version of the SAT I for female students with learning disabilities was .02 (overpredicted) and for males .21 (overpredicted). Bridgeman, McCamley-Jenkins, and Ervin (2000) reported that the standardized residuals for African-American students on the SAT were .22 for males (overpredicted) and .03 for females (underpredicted). From these data it is clear that the degree of overprediction noted for the students with learning disabilities who take the SAT with extended time is within the range of overprediction noted for other groups, however, only the group with learning disabilities is flagged, and therefore, the group not the accommodation is flagged by the College Board. In addition, approximately 40% of the 1997-98 College Board seniors with learning disabilities who had extended time on the SAT were females with little prediction error. Thus, for significant group of those with learning disabilities (females), the SAT I comes close to predicting CGPA.

Given that the differences in predictive validity interact with sex of the examinee, we find the DPV results to be equivocal regarding the comparability of scores from standard and extended

time administrations of the SAT. Furthermore, to conclude that the scores from these different administrations are not comparable would require a substantive difference in the predictive relationship between test scores and college grades. There is no commonly accepted criterion for determining at what point differences in predictive validity signify non-comparable scores. Given that the differences in predictive validity across standard and extended time administrations are smaller than those noted for subgroups of examinees who take the SAT (e.g., Caucasian/African American comparisons, Bridgeman, McCamley-Jenkins, & Ervin, 2000), we find no evidence to suggest that the magnitude of the overall difference in predictive validity between standard and extended time administration warrants a cautionary flag to be attached to the scores of students who took the test under the condition of extended time.

The evidence provided by the College Board was insufficient regarding predictive validity. The only means of reliably determining the predictive validity of extended time scores is to compare properly matched conditions: a study in which students who receive extra time on SATs also receive extra time in their first year college courses; such a study has not been carried out. Specifically, we note several limitations of the DPV research that threaten the validity of the interpretation that SAT scores from extended time administrations overpredict college grades. First, it is unknown whether the students who took the test with extended time were also granted testing accommodations at their schools. Thus, the criterion of freshman GPA could be biased against students with disabilities. Second, the study was unable to control for differential courses taken by the two student groups. Differences in courses taken could affect overall GPAs, since some subjects have more stringent grading standards than others. Third, the only college

performance criterion used was freshman GPA. Studies on differential graduation rates or exit GPAs have not been conducted. Therefore, although we believe the studies conducted by the College Board are laudable, they fall short of providing strong evidence those scores from standard and extended time administrations of the SAT are not comparable. Furthermore, there is agreement from both parties that extended time appears to enhance validity, that is, compared to standard test times, tests taken with extended time are more valid measures for students with learning disability.

From a psychometric argument only, the following points must be concluded.:

- The comparability of scores across standard and extended time administrations has not been fully studied and the results from studies that were conducted contain evidence of both comparability and noncomparability. Specifically, we note that (a) no factor analytic studies of differential test structure have been conducted on the SAT I, (b) no formal studies of differential test score reliability have been conducted on the current version of the SAT I, (c) prior research on the earlier version of the SAT I found comparable test structure and test score reliability, and (d) the differential predictive validity results are equivocal.
- There are several reasons why the differential predictive validity results cannot be used to conclude that scores from extended time administration are not comparable to scores from standard administration. These reasons include (a) inconsistency in the magnitude of freshman GPA overprediction across females and males who took the test with extended time, (b) ambiguity in determining the degree of over or underprediction that signifies non-comparability, (c) not controlling for whether students with disabilities received accommodations on their exams

in college, and (d) not controlling for differential course taking patterns across the standard and accommodated groups.

Social and Ethical Evidence

In addition to the strong scientific reasons for stopping flagging are societal and ethical reasons. Many students are reluctant to request extended time on the SAT I because the presence of the flag forces them to reveal a disability. Since the overwhelming majority of students who request extended time demonstrate learning disabilities, the presence of a flag denotes a specific personal characteristic of the examinee – a learning disability. The detrimental effect of such a designation is further supported by findings that students with learning disabilities with flagged scores are under admitted to colleges. Thus, flagging appears to single out and treat the group with learning disabilities unequally, to diminish fair chances for college admission, and to discourage the use of a mandated ADA accommodation; together, these scientific and ethical factors speak to the necessity of removing the flag. The Majority concluded that there are situations when it is necessary to treat people differently in order to treat them equally, and that this is one of them.

The members were also concerned that the policies of granting accommodations must reflect the newest scientific evidence about reading disabilities and not outmoded and inaccurate views. For example, a diagnosis of reading disabilities in young adults is strongly informed by the history of reading difficulties, by current reports of slow and effortful reading, and by measures indicating lack of fluency (slow reading). Sole reliance on measures of reading accuracy is inappropriate and misleading for these students.

The majority also expressed concern about the College Board's seeming propagation of the myth that middle class white students overuse accommodations in contrast to minority students. While it is true that more white students compared to minority students request accommodations, the Majority expresses a concern that this most likely represents an under-representation of minority students requesting accommodations and strongly urges the College Board to make energetic, proactive efforts to inform all students of their rights under the ADA to request and to have accommodations on the SAT, perhaps even targeting minority and disadvantaged students.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for education and psychological testing (2nd ed). Washington, DC: American Psychological Association.

Breimhorst vs. ETS (N.D. Cal, March 27, 2001).

Brennan, R.L. (2001, December). On the camparability of extended-time vs. standard-time scores for college board standardized tests. Paper presented to the Blue Ribbon Panel on Flagging the SAT I.

Bridgeman, B., Curley, E., & Trapani, C. (Draft 2001). To what extend (sic) is SAT I speeded?: Is SAT differentially speeded for ethnic/gender groups? Princeton, NJ: Educational Testing Service.

Camara, W., Copeland, T., & Rothschild, B. (1998). Effects of extended time on the SAT I: Reasoning test score growth for students with learning disabilities. (College Board Research Report 98-7). New York: The College Board.

Cahalan, C., Mandinach, E., & Camara, W. (Draft 2001). Predictive validity of SAT I: Reasoning test for test takers with learning disabilities and extended time accommodations. Princeton, NJ and New York: Education Testing Service and The College Board.

Rock, D.A., Bennett, R.E., Kaplan, B.A., and Jirele, T. (1988). Construct validity. In W.W. Willingham, M. Ragosta, R.E. Bennett, H. Braun, D.A. Rock, & D.E. Powers (Eds.), Testing handicapped people (pp. 99-107). Needham Heights, MA: Allyn and Bacon.