Journal of the National College Testing Association
2016/Volume 2/Issue 1

Retaking the SAT May Boost Scores but This Doesn't Hurt Validity

MICHAEL J. ROSZKOWSKI
La Salle University[1]

SCOTT SPREAT
Woods Services

In 2009, the College Board instituted a policy known as Score Choice, which gives college applicants the option to only release SAT scores from selected administrations. This has raised the concern that test validity may be compromised because only the highest scores will be known to the college. An examination of the historical records of applicants at one institution (n = 96, 295) showed that SAT scores do tend to increase on each retake. However, although averaging scores from all SAT tests taken by an applicant did produce the highest correlation with college grades (n = 20, 915), the advantage of the average SAT score over the highest SAT score was negligible. Moreover, the actual college grades of students who repeated the test were better than predicted by any regression equation developed on the basis of non-repeaters, but the smallest degree of underprediction occurred when the highest SAT score was the predictor. In sum, the likely impact of Score Choice on validity should be minor.

Keywords: SAT, retaking, repeating, scores change, validity, Score Choice policy

In many countries, admission to post-secondary education is based to some extent on performance on a standardized entrance examination (Rotberg, 2006). In the United States, the SAT (which originally was an acronym for the Scholastic Aptitude Test) is a requirement in the application process to many colleges, particularly prestigious institutions (Hawkins & Lautz, 2005). A globally recognized test (Hewitt, 2013), the SAT has undergone a number of revisions (most recently in 2005) in both name and content since its introduction in 1926, but research shows that scores from the various versions are comparable (Burton & Ramist, 2001; Kobrin & Melican, 2007).

Currently, the SAT consists of three sections: Critical Reading (SAT–CR), which was formerly called Verbal (SAT–V), Mathematics (SAT–M), and Writing (SAT–W). Each section results in a standardized score ranging from 200 to 800, where the mean is 500 (SD = 100). Prior to 2005, only the first two parts comprised the SAT, and many colleges still

[1] This article was written while Michael Roszkowski was affiliated with La Salle University. Please direct correspondence to:

Michael J. Roszkowski, PhD
101 Cherry Tree Lane
Cherry Hill, NJ 08002-1006
michaeljroszkowskiphd@gmail.com

*At the time of publication, all urls in this article were valid. If unable to connect by clicking on the url, try copying and pasting it into your browser.

continue to rely on just the SAT–CR and SAT–M in their admissions decisions (Kornblum & Toppo, 2008). Changes to be introduced in 2016 include making the Writing section optional (Resmovits & Klein, 2014).

There is no limit on the number of times that the SAT can be taken. Until a few years ago, the College Board reported a student's scores from all SAT administrations, and left it to the discretion of the college or university to decide how the multiple scores are to be used. In 2009, the College Board introduced the Score Choice policy, a reporting option that allows the student to release to the college or university only the scores from a certain test date, which most likely would be the test session resulting in the highest scores.

Stress is quite frequently associated with college admissions testing (Kruger, Wandle, & Struzziero, 2007; Lee, 2003; Yildirim, 2007), and Score Choice was introduced to reduce stress levels (College Board, n.d., a). However, because this policy permits test takers to report only their best SAT performance to colleges, some schools have opted not to participate in Score Choice due to concerns about compromising the validity of the assessment (Birnbaum, 2009; Rimer, 2008).

In view of the controversial nature of Score Choice, two primary issues are worthy of further investigation and were addressed in our study: (a) the degree of improvement that results from retaking the SAT, and (b) the validity of different methods of treating scores from multiple administrations of the SAT (i.e., first, last, lowest, highest, average of all tests) in terms of the criteria proposed by Boldt, Centra, and Courtney (1986).

REVIEW OF THE LITERATURE

Issues in Test Repetition in General

Fairness of Retaking. Both the Uniform Guidelines on Employee Selection Procedures (Equal Employment Opportunity Commission, 1978) and the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) recommend that retesting be permitted in situations involving admission, graduation, licensing, certification, and promotion. The rationale is that it is possible for the initial assessment to have been erroneous due to such extenuating circumstances as illness, anxiety, or because the person may have improved on the construct being assessed (Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007; Lievens, Buyse, & Sackett, 2005).

Typical Outcomes of Repeating Tests. A surprisingly small number of studies have addressed the outcomes resulting from repeating tests of any sort (Hausknecht, Trevor, & Farr, 2002; Hausknecht et al., 2007; Millman, 1989), but nonetheless certain patterns are evident. Generally, it has been observed that test scores tend to improve on retakes of nonstandardized tests (Cates, 2001; Friedman, 1987; Juhler, Rech, From, & Brogan, 1998) as well as standardized instruments (Andrews & Ziomek, 1998; Blair & Ford, 2006; Geving, Webb, & Davis, 2005; Hausknecht et al, 2002, 2007; Heil et al., 2002; Henriksson & Bränberg, 1994; Kulik, Kulik, & Bangert, 1984; Lane & Feitz, 1976; LeGagnoux, Michael, Hocevar, & Maxwell, 1990; Raymond, Swygert, & Kahraman, 2012; Wing, 1980).

Sometimes, however, the improvements are very minor (e.g., Griffin, Harding, Wilson, & Yeomans, 2008). Not surprisingly, gains are greater on identical test versions than on parallel forms (Hausknecht et al., 2007; Kulik et al., 1984; LeGagnoux et al., 1990). Typically, the lower the initial score, the greater the improvement is upon retesting (e.g., LeGagnoux, et al., 1990). Moreover, while each subsequent retake is likely to result in a higher score (Kulik et al., 1984), the gains tend to become progressively smaller on each successive retest (e.g., Hausknecht et al., 2002; Tornkvist & Henriksson, 2004). Improvements on retests are attributable to some combination of greater familiarity with the instrument, a self-selection effect, and an actual increase in the skill being measured (Hausknecht et al., 2007; Steiner, 2001).

Interpreting Multiple Administrations. Controversy exists about how to best interpret the outcomes from repeated tests, especially for examinations considered to be high stakes. That is, should one consider the most recent score, the best score, or the average score across the multiple administrations? An added complication exists on tests with multiple parts or sections, where the best score can be either the highest sum score from different sections of a test from the same test session, or it can be the sum of the highest section scores from administrations taken on different dates.

The worry is that scores resulting from retakes may be less valid than the initial scores. Whether this concern materializes depends on the nature of the test and the circumstances leading to a retake (Boulet, McKinley, Whelan, & Hambleton, 2003; Griffin et al., 2008; Lievens, Reeve, & Heggestad, 2007; Raymond, Swygert, & Kahraman, 2012). To appreciate this concern, it necessary to realize that the score from any given test session is just an estimate of the person's true score (which we can't measure, only estimate). Due to chance factors, for some individuals the score from a particular test session will be lower than their true score, whereas for other test takers it will be higher than their true score. If it is reasonable to assume that the construct being measured by the test will not change (as might be expected if retests are given in very short succession or when the construct isn't coachable), then the retest score is expected to fall (with 95% confidence) within +/– two standard errors of the original score.

It can be argued that one retest might be justified, just in case the person was unlucky and received a score that reflected a large negative error (i.e., underestimate). However, offering numerous retests when one shouldn't expect a change in the construct and then relying on the highest score as a measure of the construct is viewed as bad practice by some, as it greatly increases the likelihood that someone will get a spuriously high score (i.e., a high positive error) by taking the test numerous times and capitalizing on chance. That is, selective retesting (i.e., only retesting by people who scored low) is likely to lead to a higher score by chance alone. Under such circumstances, the highest score is likely to have lower validity. On the basis of classic psychometric theory, one would expect the average of the multiple scores to be the best predictor of a relevant criterion, other factors being equal, because it balances out errors of overestimation and errors of underestimation (Nunnally & Bernstein, 1994). However, when one can expect a person's level on the construct to really change, retesting may increase the validity of the scores.

Repetition of the SAT

Who Retakes the SAT? Women, higher income students, and nonminorities are more likely to repeat the SAT (Boldt et al., 1986; Nathan & Camara, 1998). The lower one's initial score, the more likely the person is to repeat (Boldt et al., 1986; Nathan & Camara, 1998; Vigdor & Clotfelter, 2003), but self-assessment of ability also influences the decision significantly (Vigdor & Clotfelter, 2003). Students who score low are generally more likely than students who score high to perceive their scores to be inaccurate (Shepperd, 1993), and in many instances they are correct (Alderman, 1981). Personality may also play a role in retaking. According to Zyphur, Islam, and Landis (2007), who examined the relationship of the Big Five personality factors to repeating the SAT, retaking the SAT was a function of greater neuroticism (a tendency to experience unpleasant emotions, such as worry, sadness, and irritability).

How SAT Retakes Are Treated by Colleges. The College Board lists the practices used by individual colleges and scholarship programs for treating scores from retakes (College Board, 2011). Our tabulation of these data indicates that 45% of these organizations use the highest section score (irrespective of test date), 33% consider all scores, and 10% use the scores from the highest single session. In the remaining cases (12%) the student is instructed to contact the institution for information. Obviously, the highest score presents the school in the most favorable light, and this fact may account for its popularity (Black & Anestis, 2009; Jaschik, 2008).

Gains versus Losses on SAT Retakes. Nathan and Camara (1998) studied the entire population of high school juniors and seniors who took the SAT between 1995 and 1997. The most salient finding was that only 10% of juniors retaking the test as seniors experienced no change in performance, whereas 55% improved and 35% scored lower. Generally, the lower the initial score, the more likely it was to subsequently go up. Conversely, the higher the first score, the more likely it was to drop on the retake. More recent data, based on the 2013 cohort of test takers with the newest version of the SAT, indicate a similar pattern (media.collegeboard.com, 2013). A study by Vigdor and Clotfelter (2003) of applicants to three selective institutions (*n* = 22,678) found that even students with above-average initial SAT scores who retook the test typically increased their scores on both the SAT–V and SAT–M.

Validity of SAT Scores from Retakes. A common approach to assessing the value of a predictor is to determine its validity coefficient, which is simply the correlation between the predictor and the criterion. The magnitude of the correlation between a test score and the criterion reflects "how good" the test is. The criterion of success in college is most typically represented in terms of the grade point average (GPA). Over the years, both first-year GPA and cumulative GPA at graduation have served as the criteria in validation studies of the SAT. In Table 1 we present an adaptation of a table from Burton and Ramist (2001), which reports the average validity coefficients from several meta-analyses examining the SAT's ability to predict both types of grades. Burton and Ramist (2001) concur with the conclusion drawn in an earlier analysis by Wilson (1983) that the SAT predicts both types of GPA quite similarly. In a more recent large-scale study, Korbin et al. (2008) investigated the validity of the 2005 revision of the SAT for predicting the first year of college GPA and concluded that despite the changes in the SAT's content and scoring over the years, the validity coefficients have remained remarkably similar.

Apparently, the analysis by Burton and Ramist (2001) did not fully consider how retests were treated in the validity studies they examined. The issue of how various approaches to using multiple SAT scores impact validity was addressed in an earlier investigation conducted by Boldt et al. (1986), who compared the validity coefficients for predicting first-year grades based on (a) the highest SAT score, (b) the most recent SAT score, (c) the score from the administration with the highest SAT–V plus SAT–M total, and (d) the simple arithmetic average of the multiple scores.

Table 2 reports the average validity coefficients Boldt et al. (1986) derived when examining the data from 24 schools where 3,561 students took the SAT three times. This table shows that although the highest validity coefficient resulted when correlating the simple arithmetic average of the multiple SAT scores to the GPA, the differences among the various indices are extremely minor. The authors summarized their findings in the following terms: "…(1) V+M is better than V or M alone, and (2) averaging is slightly the best of the four treatments of multiple scores regardless of the combination of V and M used; the increase for the average over other treatments is .01 to .02 (p. 5)." Boldt et al. (1986) also examined several weighting schemes and found that they were not superior to the simple average.

Notably, the validity coefficient is not the sole basis for judging the value of each type of SAT score. Another criterion advocated by Boldt et al. (1986) to evaluate the merits of different methods of treating multiple SAT scores is the degree to which the various techniques underpredict or overpredict actual GPA based on a regression equation developed on students who only take the SAT once. They found that all methods tended to underpredict the GPAs of multiple-times test takers, in the following order from the least to the most underprediction: highest SAT score, latest SAT score, average SAT score. In other words, the largest degree of underprediction occurred on the basis of the test takers' average SAT, whereas the lowest degree of underprediction resulted on the basis of the student's highest SAT scores. Boldt et al. (1986) therefore concluded: "…. the decision as to which is the most preferable treatment of multiple scores seems to depend on how one evaluates the discrepancy differences as compared to the validity differences" (p. 7).

TABLE 1
Average Predictive Validity of SAT Scores Using as Criteria Two Types of College GPA in Earlier and Later Studies

Authors	Sample Time Frame	SAT–Verbal	SAT–Math	Verbal and Math Combined
First-Year GPA				
Ramist (1984)	Started college between 1960 and 1980	.36	.35	.42
Bridgeman, Jenkins, & Ervin (2000)	Started college in 1995	.30	.30	.35
Cumulative GPA				
Wilson (1983)	Graduated college between 1930 and1980	.43	.31	.42
Burton & Ramist (2001)	Graduated college between 1980 and mid 1990s	.40	.41	.36

Note: Average validity coefficient computed by weighting the institutional correlations by the size of its sample.

TABLE 2
Mean Validity of Different SAT Scores for Predicting First-Year GPA at 24 Schools Studied by Boldt, Centra, and Courtney (1986)

	Highest Combined SAT–V Plus SAT–M	Highest Score	Latest Score	Average Score
SAT–V	.20	.21	.20	.22
SAT–M	.22	.23	.22	.24
Combined SAT –V Plus SAT–M	.26	.25	.25	.27

Method

Setting. The data came from a private university in the mid-Atlantic region of the United States with the following Carnegie classifications: Undergraduate Instructional Program: Bal/SGC; Graduate Instructional Program: S-Doc/Other; Enrollment Profile: HU; Undergraduate Profile: FT4/S/HTI; Size and Setting: M4/HR, Basic: Master's L. (For an explanation of these codes, please consult http://carnegieclassifications.iu.edu). Even though SAT–W scores are available on students testing since 2005, only SAT–CR and SAT–M are used in the admission decision at this institution.

Participants for Analysis of SAT Changes with Retakes. The sample consisted of 96,295 applicants who had submitted scores from 165,932 SATs taken between 1966 and 2009 (May). The number of takes of the SAT per person ranged from 1 to 11, with 1 to 5 accounting for 99.9% of the cases (*M* = 1.72, *SD* = .87). Approximately half had only one administration. The administration dates (years) of these scores were as follows: 0.10% pre-1980, 2.91% between 1980 and 1989, 27.26% between 1990 and 1999, and 69.72% between 2000 and 2009. Only the Verbal (now known as Critical Reading) and the Mathematics sections were used for analysis; the following abbreviated names will be used for each of the two sections: SAT–V/CR and SAT–M. Scores from tests prior to recentering (January 1995 and earlier) were converted to the new scores. The breakdown of the applicants by gender was 45.87% male, 52.97% female, and 1.16% unknown. The ethnic background of these applicants was 57.53% White, 10.59% Black, 4.39% Hispanic, 3.88% Asian, 0.24% American Indian, 1.83% other, and 21.53% unknown.

Participants for Analysis of SAT Validity. A subset consisting of the 20,915 applicants who became students at the university was the basis for examining the validity coefficients using the students' most recent (last available) college cumulative GPA as the criterion. This subsample was 47.78% male and 52.22% female, with the following ethnic distribution: 80.20% White, 6.87% Black, 3.37% Hispanic, 3.06% Asian, 0.14% American Indian, 0.70% other, and 5.67% unknown. It included students who were classified as freshman through senior. At this institution, class level is determined by the number of credit hours earned by the students: zero to 23 = freshman, 24 to 53 = sophomore, 54 to 83 = junior, and 84 plus = senior. The distribution of participants based on this system was 11.75% freshman, 15.20% sophomore, 11.97% junior, and 61.07% senior. Students who left the institution without graduating were included in the analysis.

On SAT–V/CR, the mean scores were: lowest = 505.47 (*SD* = 90.25), highest = 532.44 (*SD* = 89.60), average = 513.98 (*SD* = 88.59). The mean SAT–M scores were: lowest = 503.25 (*SD* = 87.70), highest = 520.96 (*SD* = 86.15), average = 512.13 (*SD* = 85.43). According to institutional policy, only courses taken at the university are incorporated into the GPA; credits transferred from other schools are therefore not part of the GPA. The mean GPA was 2.81 (*SD* =.78) on a system where F = 0, D = 1, C = 2, B = 3, and A = 4. The distribution of the number of credits on which this GPA was based had a mean of 87.10 (*SD* = 41.19) credits. The mean cumulative GPAs as a function of class were: freshman = 1.84 (*SD* = 1.27), sophomore = 2.63 (*SD* = .72), junior = 2.84 (*SD* = .60), senior = 3.04 (*SD* = .50).

RESULTS

Changes in SAT Scores with Retaking

Absolute versus Relative Stability. In discussing changes in SAT scores with retakes, it is necessary to distinguish between absolute stability and relative stability (Caspi, Roberts, & Shiner, 2005). Relative stability is the extent to which individuals maintain their position in a group. Absolute stability, in turn, is the degree to which the person's actual scores remain unchanged over repetitions of the test.

Relative Stability. Table 3 deals with the SAT's relative stability over successive administrations of the test. Only administrations one through seven are considered due to the small number of students who took it eight or more times. As may be observed from inspection of Table 3, the Pearson test-retest correlation coefficients are mainly in the .8–.9 range, indicating impressive long-term relative stability. It must be understood, however, that a high correlation does not necessarily imply that the scores remained the same over the retakes. Changes in absolute stability are possible in the face of high relative stability (Caspi et al., 2005).

Absolute Stability. An analysis of the absolute stability of the SAT may be found in Table 4, where we report the mean scores on each retake of the SAT as a function of the number of overall administrations (takes). Also shown in this table is the average score computed on the basis of all SAT administrations for a given examinee. The last column of the table reports the maximum score that the examinee obtained when all test scores are considered, irrespective of the session from which it resulted. Inspection of this table reveals several interesting relationships, which we discuss next.

TABLE 3
Pearson Correlations between SAT Scores from Successive Test Administrations

Section	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7
Verbal/Critical Reading						
Test 1	.87	.84	.82	.76	.82	.80
Test 2		.85	.84	.80	.85	.86
Test 3			.86	.82	.84	.85
Test 4				.83	.85	.84
Test 5					.89	.88
Test 6						.93
Math						
Test 1	.87	.84	.83	.82	.85	.88
Test 2		.86	.86	.84	.87	.83
Test 3			.87	.86	.88	.88
Test 4				.89	.88	.92
Test 5					.92	.94
Test 6						.95

TABLE 4
Mean SAT Scores on Each Administration for Cohorts Defined by the Maximum Number of SAT Tests Taken

Cohort	*n*	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	All Scores	Maximum Score
Verbal/Critical Reading										
1 Test	48,544	496.70							496.70	496.70
2 Tests	29,879	494.18	506.80						500.49	519.42
3 Tests	14,607	490.96	502.40	512.25					501.87	530.07
4 Tests	2,674	478.46	491.91	501.88	511.18				495.86	530.77
5 Tests	465	467.94	480.92	489.48	501.68	508.28			489.66	531.96
6 Tests	103	457.38	467.48	474.95	488.45	495.53	502.91		481.12	526.80
7 Tests	17	457.06	477.65	478.82	491.76	499.41	510.59	529.41	492.10	544.12
Math										
1 Test	48,544	491.66							491.66	491.66
2 Tests	29,879	491.78	505.34						498.56	517.92
3 Tests	14,607	492.75	508.74	519.17					506.89	535.91
4 Tests	2,674	484.94	504.50	516.40	525.99				507.96	544.58
5 Tests	465	473.38	495.63	507.63	519.37	527.68			504.54	546.82
6 Tests	103	468.35	491.26	503.20	512.82	523.40	532.91		505.32	550.10
7 Tests	17	475.29	475.88	509.41	500.59	528.24	523.53	529.41	506.05	551.76

Decision to Retake Is a Weak Function of the Initial Score. The single-time test takers and the multiple-times test takers (repeaters) were compared on their initial SAT scores. Namely, we aggregated examinees into two groups, single-time test takers and repeaters, and compared their initial scores on (a) SAT–V/CR, (b) SAT–M, and (c) SAT–V/CR and SAT–M combined. For repeaters, the respective mean scores were 491.97 SAT–V/CR, 491.46 SAT–M, and 983.43 combined SAT–V/CR plus SAT–M. For the single-time test takers, the respective scores were 496.70, 491.66, and 988.36.

These three scores were submitted to a MANOVA analysis using the SPSS general linear model routine in which we compared the initial scores of non-repeaters and repeaters. Even with these minor descriptive differences, the multivariate tests picked up a statistically significant difference, $F(2, 96292) = 48.19$, $p = .000$, Wilks Lambda = .999. The between-group comparisons were significant for SAT–V/CR, F (1,96293) = 56.73, $p = .000$ and the combined score of SAT–V/CR and SAT–M, F (1,96293) = 18.57, $p = .000$, but not for SAT–M, F (1,96293) = 0.10, $p = .747$. However, even for the two significant relationships, the effect size was extremely small (SAT–V/CR partial $\eta^2 = .001$, combined SAT–V/CR plus SAT–M partial $\eta^2 = .000$). The sample sizes were very large, hence statistically significant results are possible even with tiny effect sizes.

The previous analysis identified differences between SAT repeaters and SAT non-repeaters. It is also instructive to study the relationship between the quality of the initial SAT scores and the number of retakes. The means reported in Table 4 suggest that on both SAT–V/CR and SAT–M, the number of times the SAT was taken is negatively related to the initial score. Analyzing this relationship in terms of a Pearson correlation, we found that based on all 92,295 cases (including the single-time test takers with zero retakes), the correlation coefficient between the initial score and number of takes is –.04 ($p = .000$) for SAT–V/CR and –.01 ($p = .024$) for SAT–M. If one excludes the single-time takers from the computation ($n = 47{,}751$), then the correlation between number of retakes and the initial score is –.05 ($p = .000$) and –.02 ($p = .000$) for SAT–V/CR and SAT–M, respectively. In other words, despite the apparent differences in means, the effect size for the association between size of initial score and number of takes of the SAT is extremely small.

Increases Occur on Retakes. A comparison of scores from the initial to later administrations of the SAT, as reported in Table 4, shows that on average retake scores are better than the initial score. For each and every cohort formed on the basis of the number of retakes, paired t-tests between the first test score and last test score reveal statistically significant differences at the $p <.001$ level on both the SAT–V/CR and the SAT–M sections (see Appendix, p. 16, for details). The mean increase (weighted for varying sample sizes) between the first and final test session was nearly 17 points on the SAT–V/CR section and almost 20 points on the SAT–M section. Moreover, for the most part, each successive retake resulted in a statistically significant higher score than the immediately preceding one.

Although Table 4 demonstrates that average SAT scores on repetitions generally increase, a better perspective on the issue may be to consider the percentage of test takers who actually gain. Table 5 presents an analysis of the outcomes of repeated takes of the SAT in terms of whether the subsequent score was lower (lost), equal (same), or higher (gained). It is evident that on each retake, the proportion of test takers with either an unchanged score or a higher score exceeded the proportion of those losing points by a nearly a 2:1 ratio on both the SAT–V/ CR and the SAT–M sections. However, disregarding the experience of the 23 people who took the SAT seven times, the probability of gaining over the immediate prior score decreases slightly with each successive retake.

The Greater the Number of Retests, the Larger the Overall Gain. The mean differences on the SAT–V/CR section between the initial score and the last score as a function of one through six retakes are, respectively: 13, 21, 33, 40, 46, and 72 points. On the SAT–M section, the corresponding differences between the initial and the final score as a function of number of retakes are: 14, 26, 41, 54, 65, and 54. In terms of a Pearson correlation, the relationship between the number of retakes (one through six) and the number of points gained between the first and final taking of the SAT equals .25 on the SAT–V/CR section and .30 on the SAT–M section. In a number of instances (e.g., four administrations of the SAT–V/CR), the largest gain occurred between the initial score and the first retake, with the subsequent gains being somewhat lower. However, the diminishing returns effect was not very large, and it may be observed in Table 4 that not every successive retake resulted in a lower gain than the preceding one.

Lower Initial Scores Result in More Gains. There is a relationship between the level of the initial score and how much one gains on the retake. Namely, the lower the initial score, the greater the gain on a retake. The correlation between the initial (first) score and number of points gained by taking the SAT a second time is $r = –.24$ for the SAT–V/CR section and $r = –.23$ for the SAT–M section. A similar relationship can be observed on subsequent retakes. That is, the correlation between number of points change between the second and third test sessions of the SAT, and the score on the second administration is $r = –.25$ and $r = –.26$ on SAT–V/CR and SAT–M, respectively. Likewise, the change between the third and the fourth sessions and the scores from the third session correlate at –.23 on SAT–V/CR and –.26 on the SAT–M. The comparable correlation between the difference score from the fifth administration minus the fourth administration and the score from the fourth test session equals –.27 for SAT–V/CR and –.28 for SAT–M. All the reported correlations are statistically significant.

Retakes Lead to Higher Average Scores. It is also evident that even in terms of the impact on a person's average scores, repeating the SAT produces a statistically significant advantage on both SAT–V/CR, *t* (47,750) = 73.05, *p* = .000 and SAT–M, *t* (47,750) = 85.11, *p* = .000. For the examinees who sat for the SAT on multiple occasions, the mean difference between the individual's initial score and that person's average score is 8.54 (*SD* = 25.54) on SAT–V/CR and 10.26 (*SD* = 26.33) on SAT–M. When the number of takes of the SAT (2 through 11) is correlated with the difference between the initial score and the person's average score over all administrations, *r* = .04 (*p* = .000) for SAT–V/CR and *r* = .13 (*p* = .000) for SAT–M. In other words, although the effect size is small (particularly on SAT–V/CR), the more scores that go into the average, the larger the difference between that average and the score obtained on the first occasion the test was taken.

Validity of Retakes

SAT Validity Coefficients as a Function of Type of Score. For those students who enrolled at the university (*n* = 20,915), Pearson correlations were computed between their last cumulative GPA and the different types of SAT scores (i.e., first, last, lowest, highest, average). Considering just the students who took the SAT only once (*n* = 13,340), the correlation coefficients between their SAT scores and their cumulative college GPA were .233 for SAT–V/CR and .276 for SAT–M. Table 6 reports the Pearson correlations between each section of the SAT and the cumulative GPA for the repeaters as a function of the number of takes and whether one employs the test taker's first, last, lowest, highest, or average SAT score as the basis for computing the validity coefficient. To eliminate small cell sizes, students who took the SAT more than three times were aggregated into a "four plus" category.

TABLE 5
Percent of Repeaters Losing Points, Retaining the Same Score, and Gaining Points over SAT Retests

	n	Lost	No Change	Gained	Total
Verbal/Critical Reading					
Test 2 vs. Test 1	47,751	34.47%	9.83%	55.70%	100%
Test 3 vs. Test 2	17,872	36.35%	9.61%	54.04%	100%
Test 4 vs. Test 3	3,265	34.79%	11.52%	53.69%	100%
Test 5 vs. Test 4	591	37.73%	10.15%	52.12%	100%
Test 6 vs. Test 5	126	35.71%	13.49%	50.79%	100%
Test 7 vs. Test 6	23	17.39%	13.04%	69.57%	100%
Math					
Test 2 vs. Test 1	47,751	32.91%	9.41%	57.68%	100%
Test 3 vs. Test 2	17,872	35.00%	10.11%	54.90%	100%
Test 4 vs. Test 3	3,265	35.93%	11.39%	52.68%	100%
Test 5 vs. Test 4	591	34.52%	10.15%	55.33%	100%
Test 6 vs. Test 5	126	34.92%	15.87%	49.21%	100%
Test 7 vs. Test 6	23	30.43%	17.39%	52.17%	100%

TABLE 6
Validity Coefficients for the SAT Test by Test Session as a Function of Number of Administrations and Type of Score

Total Number of Administrations Undertaken	*n*	*Verbal/Critical Reading*					*Math*				
		First	Last	Lowest	Highest	Mean	First	Last	Lowest	Highest	Mean
1	13,340	.233	—	—	—	—	.276	—	—	—	—
2	4,250	.303	.289	.302	.303	.308	.285	.288	.296	.289	.298
3	2,635	.305	.307	.320	.321	.327	.267	.286	.290	.286	.298
4 Plus	690	.328	.354	.335	.343	.356	.327	.303	.297	.300	.310
Any Multiple	7,575	.295	.296	.296	.311	.311	.274	.290	.287	.292	.298

The differences between these various validity coefficients are minute (hence we report them to the third decimal place). In most instances, the examinee's highest (best) SAT score produced stronger validity coefficients than either the first, last, or lowest score. However, the superiority of the average score over the best score proved to be somewhat nuanced. That is, when one examines the pattern within each individual number of retakes, there is a slight advantage to the average score over the maximum score. For example, among students who took the SAT three times, the SAT–V/CR correlation coefficient was .321 for the maximum score and .327 for the average score. Similarly, the SAT–M validity coefficient was .286 based on the maximum score and .298 based on the average score. But when the data for repeaters were aggregated (disregarding the number of repetitions) into a single set of repeaters (2 or more takes), the validities of the average and maximum scores were identical on SAT–V/CR (r = .311), and on SAT–M the advantage of the average (over the maximum score) was barely evident: .298 vs. 292.

SAT Validity Coefficients as a Function of Number of Administrations. There appears to be a progressive increase in the size of the validity coefficients, particularly for the SAT–V/CR score, as a function of increasing test session number. Namely, the greater the number of SAT administrations that test takers undertook, the higher the correlation between their SAT scores and their GPA (i.e., higher validity coefficients). This is evident when using either the first, last, minimum, maximum, or the average score. In other words, individuals who sat for multiple test sessions are more predictable, even on the basis of their performance on their initial SAT. For instance, considering the initial SAT administration, the validity coefficient for the SAT–V/CR section among students who took the SAT only once is .23. It climbs progressively with each number of takes as follows: .30 for students who took the SAT twice, .31 for students who took it three times, and .33 for students who took it four or more times.

Because of decreasing sample sizes with increasing number of test sessions, we decided to also aggregate the repeaters into a single group and compare their performance on the SAT to that of the single test takers. In other words, we compared the 7,575 students who repeated the SAT with the 13,340 students who took the SAT only once on their scores from their *first* administration. For the students who repeated the SAT (not paying attention to the number of retakes), the correlations with college GPA were .295 and .274 for SAT–V/CR and SAT–M, respectively. The corresponding validity coefficients were .233 and .276 for the students who did not repeat the SAT. The difference between .295 and .233 is statistically significant (z = 4.63, p = .000), and indicates that students who later repeated the test exhibited a slightly higher validity coefficient for the SAT–V/CR section even on the first administration (i.e., first time they took the SAT).

Underprediction of SAT Repeaters. Boldt et al. (1986) recommended that in addition to looking at the validity coefficients, another criterion to be used in judging the value of a particular method of treating multiple SAT scores should be whether the procedure over- or underpredicts the actual college GPA when a regression equation developed on the basis of students who take the SAT only once is applied to students who take the SAT multiple times. If the predicted grade is lower than the actual grade, then there is underprediction. Conversely, if the predicted grade is higher than the actual grade, then overprediction has occurred.

The results of this regression procedure appear in Table 7, which reports the actual GPA and predicted GPA as a function of the number of times the student took the SAT and which type of SAT score was used in the regression equation (SAT–V/CR and SAT–M as two predictors). Since the regression equation was developed on the students who took the SAT just once, the mean actual and the mean predicted GPA are the same for the single-time test takers based on the mathematics of the regression procedure.

TABLE 7
Actual GPA, Predicted GPA, and Residuals as a Function of Number of Takes of the SAT

	1 Take (*n* = 13,340)		2 Takes (*n* = 4,250)		3 Takes (*n* = 2,635)		4 Plus Takes (*n* = 690)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Actual GPA	2.74	.80	2.89	.77	2.96	.70	3.04	.68
Predicted								
Initial SAT	2.74	.23	2.73	.22	2.70	.21	2.67	.21
Maximum SAT	—	—	2.80	.21	2.82	.20	2.84	.21
Minimum SAT	—	—	2.69	.22	2.65	.21	2.62	.22
Average SAT	—	—	2.75	.21	2.74	.20	2.73	.21
Residuals								
Initial SAT	—	—	–.16	.73	–.26	.67	–.38	.64
Maximum SAT	—	—	–.09	.73	–.15	.66	–.21	.64
Minimum SAT	—	—	–.20	.73	–.31	.66	–.42	.64
Average SAT	—	—	–.14	.73	–.23	.66	–.31	.64

Inspection of this table allows for several observations. To begin with, actual GPA increases as a function of number of takes, $F(3,20911) = 100.51$, $p = .000$, $\eta^2 = .014$. Second, the regression equation developed on the scores of the single-time test takers underpredicts the GPA of repeaters in all instances, but the least so on the basis of the highest SAT score and the most so on the basis of the lowest SAT. Third, the size of the residual (extent of underprediction) increases as a function of the number of administrations of the SAT.

DISCUSSION

Repeating high stakes tests is generally permitted, but an outstanding issue is how to best make use of the results from repeated tests in a test-based decision so as not to compromise validity. Given the concern raised about Score Choice, introduced by The College Board in 2009, two primary issues were the focus of this study. First, we wanted to better understand the effect that repeating the SAT had on the quality of the scores obtained on retakes. Second, we sought to determine the validity of different approaches to treating scores from multiple administrations of the SAT (i.e., first, last, lowest, highest, average of all tests). Prior research gave us a good inkling as to what the answer would be to the first question, but we were less sure regarding the answer to the second question. This is because the literature on the changes in scores attributable to repeating tests is more extensive than the research on the validity of scores from the repetitions (Hausknecht et al, 2007). For the decision on whether a college should participate in Score Choice, the most critical issue is seeing what happens if SAT scores from all administrations are not averaged and only the highest score is used.

Scores on Retakes Are Likely to Improve

Although the correlations between the initial score and each subsequent score were very high (r =.8 –.9 range) for test takers who repeated, retaking the SAT generally produced beneficial outcomes. On average, the SAT scores on repetitions were better, especially if the initial score was low. On each retake, the proportion of test takers gaining points or retaining the same score exceeded the proportion losing points by nearly a 2:1 ratio. Moreover, the more SAT test sessions the student experienced, the bigger the difference between the first score and the last score. It is especially noteworthy that the amount of gain would be sufficient to have practical significance in an admission decision. Thus, the advice offered by many guidance counselors to “take the test again” seems to be supported by these results. Most likely, these gains are attributable to some combination of self-selection, regression towards the mean, practice effects, and motivation (Hausknecht et al., 2007).

Retakes Do Not Necessarily Impair Validity

Some university administrators fear that Score Choice will compromise the SAT’s validity since it would seemingly enable the student to capitalize on chance. Namely, the concern is raised that the student can pick the score in which her or his performance is maximized by measurement error. One would expect the introduction of increased measurement error to markedly reduce the utility of the SAT as a predictor of grades, but surprisingly this does not seem to be the case based on the evidence reported here and by Boldt et al. (1986).

Furthermore, the advantage of the person's average SAT score over her or his best score, while there, is surprisingly slight. The maximum SAT score, which presumably would be the score that a rational student would choose to submit to college admission committees, was practically as good a predictor of the college GPA as the average SAT score. More significantly, on the basis of how much each type of score under- or overpredicts the actual college GPA, we found that consistent with Boldt et al. (1986), a regression equation developed on single-time test takers, when applied to repeaters, tends to underpredict their actual college GPA regardless of which type of SAT score is used as a predictor, but this systematic error is minimized when the person's highest score is used.

Retaking May Reflect Academic Motivation

In line with prior research, we found that relative to test takers who only took the SAT once, repeaters had lower initial scores on the SAT and that the number of repeats bears an inverse relationship to the initial score. But the difference in initial scores between repeaters and non-repeaters in our data was quite modest, suggesting that factors besides a poor score were involved in the decision to repeat the test. What we find most intriguing in our results is that retaking seems to be related to a number of indicators suggesting better motivation. Firstly, the larger the number of takes of the SAT, the higher was the student's actual college GPA. Secondly, validity coefficients (SAT–GPA correlations) based on SAT repeaters were superior to such coefficients derived on the basis of single-time test takers, which suggests that repeaters are more predictable than single-time test takers (less random error).

Finally, there is evidence that the college grades of repeat SAT takers are underpredicted on the basis of their SAT, especially the initial one, which implies that SAT repeaters tend to overachieve (or maybe are less likely to underachieve). It has been recognized for some time that low motivation can pose a threat to validity (Arvey, Strickland, Drauden, & Martin, 1990; Revelle, 1993) and perhaps other factors being equal, students who only take the SAT once are exhibiting lower academic motivation than comparable students who retake the SAT. Of course, this notion should be tested more directly by comparing the repeaters and single-time test takers on some formal measure of academic motivation (e.g. Vallerand et al., 1992).

It is instructive to consider our results in the context of Simon's (1955) concept of "satisficing" versus "optimizing." Optimizers seek to obtain the very "best" solution whereas satisficers are willing to accept a merely "good enough" solution (Byron, 1998; Parker, de Bruin, & Fischhoff, 2007; Schwartz et al., 2002). Thus, a satisficing orientation to test taking means that if a score is considered to be good enough, the individual will not try to improve on it. In contrast, operating under an optimizing decision-making orientation, the person will strive to obtain the maximum possible score. Hence, satisficers would be less likely to repeat the SAT than optimizers. Retaking the SAT numerous times may also be an example of the "grit" construct, which is a powerful motivation to achieve an objective in the face of seemingly insurmountable obstacles (Duckworth, Peterson, Matthews, & Kelly, 2007). Despite a disappointing initial score, test takers with "grit" would be likely to repeat until they achieved a score that that they had set as their goal.

CONCLUSION

In sum, there is nothing to suggest that the adoption of Score Choice will introduce much risk to the selection process. The magnitude of the differences in correlations between the various SAT scores (lowest, highest, average) and the GPA is so slight that perhaps it renders them meaningless. Consequently, it is markedly clear that allowing a student to select the best score will not significantly jeopardize prediction of GPA, particularly in light of the fact that most colleges base their admission decision on the applicant's maximum SAT score even when scores from all tests that the student took are available.

Of course, the modest size of the correlations between SAT scores and GPA do suggest that there is need for additional predictors of college success for use by admissions departments. Unquestionably, motivation exerts a powerful influence on academic achievement (Credé & Kuncel, 2008; Noftle & Robins, 2007; Robbins et al., 2004; Roszkowski & Ricci, 2009), and retaking the SAT may be indicative of a motivation to succeed in college. However, not everyone who doesn't try to improve their score suffers from low motivation. Retaking the SAT costs money, and there may be economic reasons for not repeating the test. Thus, Score Choice may put students coming from low income households at a disadvantage. Ironically, for people of reduced means, the policy may be inadvertently promoting test unfairness rather than fairness.

Limitations

There are several caveats to the generality of our conclusions. First, the SAT scores that were the basis for this analysis came from only one institution, and they were obtained over the course of a number of years, during which the content and scoring of the SAT had changed. Aggregation of this sort may produce some unknown artifacts, but some comfort is offered by the findings presented by Kobrin and Melican (2007), demonstrating comparability between the different historical forms of the SAT. Second, the contribution of the Writing score was not considered, but in the next version of the SAT, that section will become optional. Third, the college GPA that served as the criterion was not based on the same number of credits, and it is possible that the SAT is better at predicting first-year college grades than grades in succeeding years (Burton & Ramist, 2001). The probable effect is that our validity coefficients are attenuated, but this should not compromise the observed differences between single-time test takers and test repeaters. Lastly, while all SAT scores at the time of application were used in the analysis, it is possible that some students who we considered to be single-time test takers did in fact repeat the test later, but these scores were not available to us. The impact of this misclassification would be to reduce the observed differences between single-time and repeat test takers.

It is possible to check on the likelihood of this last potential problem by comparing the proportion of examinees sitting for one through five administrations of the SAT in our sample with the national population (N = 1,119,984) pattern reported by Nathan and Camara (1998). According to Nathan and Camara (1998), the percentage of examinees who took the SAT one through five times was: 50.67%, 38.09%, 9.63%, 1.40%, and 0.22% (M = 1.62). In our data, the respective figures were: 50.48%, 31.07%, 15.19%, 2.78%, and 0.48% (M = 1.72). The distributions are not very different, especially regarding single test takers. Consequently, misclassification does not appear to be a major threat. Moreover, our data are consistent with those reported by Nathan and Camara (1998) on the percentage of test takers improving, doing worse, and not changing on retakes of the SAT.

REFERENCES

Alderman, D. L. (1981). Student self-selection and test repetition. *Educational and Psychological Measurement, 41*, 1073–1081.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Andrews, K. M., & Ziomek, R. L. (1998). *Score gains on retesting with the ACT assessment* (ACT Report No 98-7). Iowa City, IA: ACT.

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43,* 695–716.

Birnbaum, M. (2009, April 10). New SAT policy can keep low scores from reaching colleges. *Washington Post.* Retrieved from http://www.washingtonpost.com/wp-dyn/content/article/2009/04/09/AR2009040904051_2.html.

Black, C., & Anestis, M. (2009). *McGraw Hill's SAT 2010*, New York, NY; McGraw Hill.

Blair, M. D., & Ford, J. M. (2006, June). *Factors influencing applicant performance when retaking employment exams.* Paper presented at the 30th Annual International Public Management Association for Human Resources Assessment Council (IPMAAC) Conference on Personnel Assessment, Las Vegas NV.

Boldt , R. F., Centra, J. A., & Courtney, R. G. (1986). *The validity of various methods of treating multiple SAT® scores, (College Board Report No. 86-4).* New York, NY: College Entrance Examination Board. Retrieved from http://research.collegeboard.org/publications/content/2012/05/validity-various-methods-treating-multiple-sat-scores.

Boulet, J. R., McKinley, D. W., Whelan, G. P., & Hambleton, R. K. (2003). The effect of task exposure on repeat candidate scores in a high-stakes standardized patient assessment. *Teaching and Learning in Medicine, 15*, 227–232.

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

Burton, N. W. & Ramist, L. (2001). *Predicting success in college: SAT studies of classes graduating since 1980 (Research Report No. 2001-2).* New York, NY: College Entrance Examination Board.

Byron, M. (1998). Satisficing and optimality. *Ethics, 109*, 67–93.

Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and change. *Annual Review of Psychology, 56,* 453–484.

Cates, W. M. (2001). The efficacy of retesting in relation to improved test performance of college undergraduates. *Journal of Educational Research, 75*, 230–236.

College Board (n.d., a). *Score Choice: New SAT® score-reporting policy*. Retrieved from http://www.collegeboard.com/events/repository/399.pdf.

College Board (2011). *SAT® Score-Use practices by participating institution*. Retrieved from https://secure-media.collegeboard.org/digitalServices/pdf/professionals/sat-score-use-practices-participating-institutions.pdf

Credé, M., & Kuncel, N. R. (2008). Study habits, skills, and attitudes: The third pillar supporting collegiate academic performance. *Perspectives on Psychological Science*, *3*, 425–453.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92,* 1087–1101.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43*(166), 38290–39315.

Friedman, H. (1987). Repeat examinations in introductory statistics courses. *Teaching of Psychology, 14* (1), 20–23.

Geving, A.M., Webb, S., & Davis, B. (2005).Opportunities for repeat testing: Practice doesn't always make perfect. *Applied H.R.M. Research, 10* (2), 47–56.

Griffin, B., Harding, D. W., Wilson, I. G., & Yeomans, N. D. (2008). Does practice make perfect? The effect of coaching and retesting on selection tests used for admission to an Australian medical school. *The Medical Journal of Australia, 189*, 270–273. Retrieved from http://www.mja.com.au/public/issues/189_05_010908/gri10324_fm.html.

Hausknecht, J. P., Trevor, C. O., & Farr, J. L. (2002). Retaking ability tests in a selection setting: Implications for practice effects, training performance, and turnover. *Journal of Applied Psychology, 8*, 243–254.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*, 373–385.

Hawkins, D. A., & Lautz, J. (2005). *State of college admission.* Alexandria, VA: National Association for College Admission Counseling. Retrieved from http://www.nacacnet.org.

Heil, M. C., Detwiler, C. A., Agen, R., Williams, C. A., Agnew, B. O., & King, R. E. (2002). The effects of practice and coaching on the Air Traffic Selection and Training test battery. *FAA Office of Aviation Medicine Reports*, December 1, 1–8, A1–A2.

Henriksson, W., & Bränberg, K. (1994). The effects of practice on the Swedish Scholastic Aptitude Test. *Scandinavian Journal of Educational Research, 38,* 129–148.

Hewitt, T. (2013, March 4). *ACT versus SAT: Popular doesn't mean better*. Retrieved from http://www.icon-plus.com/news-and-media/blog/act/act-versus-sat-popular-doesnt-mean-better.

Jaschik, S. (2008, October 15). Baylor pays for SAT gains. *Inside Higher Ed.* Retrieved from http://www.insidehighered.com/news/2008/10/15/baylor#ixzz36tqEZrkq.

Juhler, S. M., Rech, J. F., From, J. F., & Brogan, M. M. (1998). The effect of optional retesting on college students' achievement in an individualized algebra course. *Journal of Experimental Education, 66*, 125–138.

Kobrin, J. L. & Melican, G. J. (2007). *Comparability of scores on the new and prior versions of the SAT Reasoning Test (College Board Research Note RN31).* New York, NY: Office of Research and Analysis of The College Board.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT for predicting first-year college grade point average (College Board Research Report No. 2008-5).* New York, NY: The College Board.

Kornblum, J., & Toppo, G. (2008, April 24). Studies: SAT writing portion good predictor of grades. *USA Today*. Retrieved from http://usatoday30.usatoday.com/news/education/2008-04-24-sat_N.htm.

Kruger. L. J., Wandle, C., & Struzziero, J. (2007). Coping with the stress of high stakes testing. *Journal of Applied School Psychology, 23*, 109–128.

Kulik, J. A., Kulik, C. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21,* 435–447.

Lane, M. S., & Feitz, R. H. (1976). Cross-year comparison of test-retest MCAT performance. *Journal of Medical Education, 51*, 582–586.

Lee, M. (2003). Korean adolescents' "examination hell" and their use of free time. *New Directions for Child & Adolescent Development, 99*, 9–22.

LeGagnoux, G., Michael, W. B., Hocevar, D., & Maxwell, V. (1990). Retest effects on standardized structure-of-intellect ability measures for a sample of elementary school children. *Educational and Psychological Measurement, 50,* 475–492.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology, 58*, 981–2007.

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology, 92*, 1672–1682.

Media.Collegeboard.com (2013). *SAT percentile student senior-year score gain loss 2013*. Retrieved from http://media.collegeboard.com/digitalServices/pdf/research/SAT-Percentile-Student-Senior-Year-Score-Gain-Loss-2013.pdf.

Millman, J. (1989). If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher, 18* (6), 5–9.

Nathan, J. S. & Camara, W. J. (1998). *Score change when retaking the SAT® I: Reasoning Test (RN-05*). New York, NY: Office of Research, The College Board. Retrieved from http://research.collegeboard.org/sites/default/files/publications/2012/7/researchnote-1998-5-score-change-retaking-sat.pdf.

Noftle, E. F., & Robins, R. W. (2007). Personality predictors of academic outcomes: Big five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology, 93*, 116–130.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.

Parker, A. M., de Bruin, W. B., & Fischhoff, B. (2007). Decision-making styles, competence, and outcomes. *Judgment and Decision Making, 2*, 2007, 342–350.

Ramist, L. (1984). Validity of the ATP tests. In T. Donlon (Ed.), *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests* (pp. 141–170). New York, NY: The College Board.

Raymond, M. R., Swygert, K. A., & Kahraman, N. (2012). Psychometric equivalence of ratings for repeat examinees on a performance assessment for physician licensure. *Journal of Educational Measurement, 49,* 339–361.

Resmovits, P., & Klein, R. (2014, April 16). SAT changes include fewer answer choices, shorter mandatory test time. *Huffington Post*. Retrieved from http://www.huffingtonpost.com/2014/04/16/SAT–changes_n_5146219.html.

Revelle, W. (1993). Individual differences in personality and motivation: 'Non-cognitive' determinants of cognitive performance. In A. Baddeley & L. Weiskrantz (Eds.), *Attention: selection, awareness and control: A tribute to Donald Broadbent* (pp. 346–373). Oxford: Oxford University Press.

Rimer, S. (2008, October 16). Baylor faculty members condemn SAT retaking. *New York Times*, p. A 21. Retrieved from www.nytimes.com/2008/10/16/education/16baylor.html.

Robbins, S., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychological and study skill factors predict college outcome? A meta-analysis. *Psychological Bulletin, 130,* 261–288.

Roszkowski, M. J., & Ricci, R. (2009). Is eighty percent of success just showing up? Student compliance and refusal to complete an advisement form as an indicator of first-term GPA. *The Journal of College Orientation and Transition, 7*, 57–65.

Rotberg, I.C. (2006). Assessment around the world. *Educational Leadership*, *64* (3), 58–63.

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology*, *83*, 1178–1197.

Shepperd, J. A. (1993). Student derogation of the Scholastic Aptitude Test: Biases in perceptions and presentations of College Board scores. *Basic and Applied Social Psychology, 14,* 455473.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 59,* 99–118.

Streiner, D. L. (2001). Regression toward the mean: Its etiology, diagnosis, and treatment. *Canadian Journal of Psychiatry, 46,* 72–76.

Tornkvist, B., & Henriksson, W. (2004). *SweSAT repeat (EM No. 46).* Umeå, Sweden: Umeå, University. Retrieved from http://www.edusci.umu.se/digitalAssets/59/59525_em-no-46.pdf.

Vallerand, R. J., Pelletier, L. G., Blais, M. R., Briere, N. M., Senecal, C., & Vallieres, E. F. (1992). The Academic Motivation Scale: A measure of intrinsic, extrinsic, and amotivation in education. *Educational and Psychological Measurement*, *52*, 1003–1017.

Vigdor, J. L., & Clotfelter, C. T. (2003). Retaking the SAT. *Journal of Human Resources*, *38,* 1–33.

Wilson, K. M. (1983). *A review of research on the prediction of academic performance after the freshman year (College Board Report No. 83-2).* New York: College Board.

Wing, H. (1980). Practice effects with traditional mental test items. *Applied Psychological Measurement*, *4*, 141–155.

Yildirim, I. (2007). Depression, test anxiety and social support among Turkish students preparing for the university entrance examination. *Eurasian Journal of Educational Research, 29,* 171–184.

Zyphur, M. J., Islam, G., & Landis, R. S. (2007). Testing 1, 2, 3, ...4? The personality of repeat SAT test takers and their testing outcomes. *Journal of Research in Personality, 41*, 715–722.

APPENDIX

Paired t-tests between Sores from Repeated Administrations of SATV/CR and SAT–M Scores as a Function of the Maximum Number of Tests Taken

Maximum Number of SAT Administrations	Comparison	SAT–V/CR			SAT–M		
		t	*df*	*p*	*t*	*df*	*p*
2	1 vs. 2	–46.02	29878	.000	–48.68	29878	.000
3	1 vs. 2	–30.21	14606	.000	–41.68	14606	.000
	1 vs. 3	–53.89	14606	.000	–65.85	14606	.000
	2 vs. 3	–25.61	14606	.000	–27.54	14606	.000
4	1 vs. 2	–15.21	2673	.000	–21.42	2673	.000
	1 vs. 3	–25.23	2673	.000	–32.60	2673	.000
	1 vs. 4	–33.06	2673	.000	–42.09	2673	.000
	2 vs. 3	–11.62	2673	.000	–13.67	2673	.000
	3 vs. 4	–10.57	2673	.000	–11.02	2673	.000
5	1 vs. 2	–5.86	464	.000	–10.30	464	.000
	1 vs. 3	–9.53	464	.000	–14.97	464	.000
	1 vs. 4	–13.74	464	.000	–18.30	464	.000
	1 vs. 5	–14.71	464	.000	–21.61	464	.000
	2 vs. 3	–3.94	464	.000	–5.88	464	.000
	3 vs. 4	–5.73	464	.000	–4.91	464	.000
	4 vs. 5	–2.83	464	.005	–4.61	464	.000
6	1 vs. 2	–2.31	102	.023	–5.41	102	.000
	1 vs. 3	–3.70	102	.000	–7.45	102	.000
	1 vs. 4	–7.10	102	.000	–9.61	102	.000
	1 vs. 5	–7.50	102	.000	–10.80	102	.000
	1 vs. 6	–8.69	102	.000	–13.10	102	.000
	2 vs. 3	–1.47	102	.145	–2.61	102	.010
	3 vs. 4	–3.01	102	.003	–2.44	102	.016
	4 vs. 5	–1.54	102	.126	–2.71	102	.008
	5 vs. 6	–1.82	102	.072	–2.69	102	.008
7	1 vs. 2	–2.66	16	.017	–0.04	16	.969
	1 vs. 3	–2.02	16	.061	–3.52	16	.003
	1 vs. 4	–4.17	16	.001	–2.22	16	.041
	1 vs. 5	–3.15	16	.006	–5.42	16	.000
	1 vs. 6	–4.86	16	.000	–4.26	16	.001
	1 vs. 7	–7.36	16	.000	–4.18	16	.001
	2 vs. 3	–.12	16	.905	–2.60	16	.019
	3 vs. 4	–1.67	16	.115	1.14	16	.269
	4 vs. 5	–.66	16	.516	–3.05	16	.008
	5 vs. 6	–1.06	16	.306	.66	16	.522
	6 vs. 7	–2.41	16	.028	–.79	16	.440