SAT Suite of Assessments Administration Report

Delaware
SAT School Day Administration
Test Takers with Accommodations
Spring 2018

Revised September 10, 2018

Executive Summary

This report summarizes the performance of 395 Delaware test takers who took the Spring 2018 SAT School Day administration with accommodations. There were three master forms administered with accommodations in Delaware (form 1 had 395 test takers; form 2 had 1 test takers; form 3 had 18 test takers). This report provides an analysis of the quality of the test forms administered to at least 100 test takers in the state of Delaware. Psychometric and statistical summaries related to the moments, intercorrelations, reliability and standard error of measurement (SEM), item completion rates, form speededness, differential item functioning, and classification accuracy and consistency are also included. Depending on psychometric recommendations for minimum sample sizes for these analyses, results are reported only for forms for which the subgroup sample size was 5 or more, 100 or more, or 200 or more.

This report also summarizes the performance of 367 students who took the SAT Essay with accommodations in the Spring 2018 School Day administration and received non-zero scores. This report includes a summary of descriptive statistics, frequency distributions, correlations of essay dimension scores, and interrater consistency.

Quality of the form(s):

Most of the takers included in this sample were 11th graders. About 73% spoke English or English and another language as their first language. About 62% of the sample was male and 38% was female.

The mean Evidence-Based Reading and Writing (ERW) score was 424 and the standard deviation was 89. The mean Math Section score (MSS) was 404, with a standard deviation of 92. The mean total score was 829 and the standard deviation was 169.

The observed score correlation between ERW and MSS was 0.74. The true score correlation between ERW and MSS was 0.85.

The scale score reliability of ERW was 0.91. The average conditional standard error of measurement (CSEM) for ERW across forms was 27. The scale score reliability of MSS was 0.85. The CSEM for MSS was 35. The scale score reliability of the Total score was 0.93. The CSEM for the Total score was 45.

Over 92% of the sample completed at least 75% of each of the Reading, Writing and Language, Math – No Calculator, and Math – Calculator timed sections of the test across all forms.

None of the items were classified as C+ or C- by differential item functioning analysis.

The percentage of test takers who met Level 3 and Level 4 for ERW was 23%. The percentage of test takers who met Level 3 and 4 for MSS was 9%. The probability of correct classification for the total group was 0.84 for ERW and 0.85 for MSS. The proportion of consistent decisions for the total group was 0.78 for ERW and MSS.

About 414 test takers took the SAT essay test. Out of these test takers, 367 received non-zero essay scores. The average dimension scores were 3.16 for essay reading, 2.49 for essay analysis, and 3.68 for essay writing across all forms.

The observed score correlations of the three essay dimension scores was 0.54 between essay reading and essay analysis, 0.78 between essay reading and essay writing, and 0.61 between essay analysis and essay writing. The range of the correlations between essay dimension scores and Reading Test scores, Writing and Language Test scores and ERW scores was 0.45 to 0.60.

The percentage of exact agreement between the two raters was 69.75 for essay reading, 85.29 for essay analysis, and 70.03 for essay writing. The correlations between the essay dimension scores given by two raters for essay reading was 0.65 with an SEM of 0.43, 0.67 with an SEM of 0.30 for essay analysis, and 0.72 with an SEM of 0.40 for essay writing. The simple Kappa was 0.48 for essay reading, 0.57 for essay analysis, and 0.53 for essay writing. The weighted Kappa was 0.56 for essay reading, 0.61 for essay analysis, and 0.62 for essay writing.

Contents

SAT Suite of Assessments

The SAT Suite of Assessments (i.e., SAT, PSAT/NMSQT®, PSAT™ 10, and PSAT™ 8/9) is designed to measure student readiness for college and postsecondary education. Each assessment contains two sections (Evidence-Based Reading and Writing section [ERW] and the Math section [MSS]), three tests (Reading Test, Writing and Language Test, and Math Test), two cross-tests (Analysis in History/Social Studies and Analysis in Science) and seven subscores (Command of Evidence, Words in Context, Expression of Ideas, Standard English Conventions, Heart of Algebra, Problem Solving and Data Analysis, and Passport to Advanced Math). For the SAT, test takers are given three hours to complete 154 items. Test takers who choose to also take the optional Essay are given an additional 50 minutes.

This report contains summary information about the score tiers; specifically, the total, section, and test scores, as well as the cross-test scores, and the subscores from the Spring 2018 School Day administration of the SAT forms for the state of Delaware. Raw scores were generated from the number of items the student answered correctly within the score tier. Scale scores were generated by applying the appropriate raw-to-scale score conversions. Table 1 describes the number of items and score scale ranges for the SAT.

The Reading Test and Writing and Language Test are administered in separately-timed sections and only contain multiple-choice (MC) items. The Math Test is administered over two separately-timed sections, Math – No Calculator and Math – Calculator. In addition, the Math Test includes two types of items in each timed section, multiple-choice (MC) items and student-produced response (SPR) items. The SAT also includes an optional essay with one prompt. See Table 2 for the number and type of items per timed section for the included forms. The content specifications for the SAT provide additional details for each test within the SAT (College Board, 2014).

The content specifications are deeply informed by evidence about essential requirements for college and career readiness and success. In constructing each test form of the SAT, the content specifications are of primary importance. As such, the SAT forms in the Delaware Spring 2018 School Day administration meets 100% of the target content specifications. The same form was also administered to a national equating sample. More information about the national equating samples used for equating is in Chapter 6 of the SAT Suite of Assessments Technical Manual (College Board, 2017). The target statistical specifications for the SAT Suite are in Appendix A. The target values for difficulty, discrimination, and reliability are summarized in Tables A1 to A5.

SAT Essay

Test takers opting to take the SAT Essay receive an additional 50 minutes at the end of the SAT testing session to compose a clear and cogent analysis of a high-quality source text. The same prompt appears with every essay text:

"As you read the passage below, consider how [the author] uses

- evidence, such as facts or examples, to support claims.
- reasoning to develop ideas and to connect claims and evidence.
- stylistic or persuasive elements, such as word choice or appeals to emotion, to add

> power to the ideas expressed.
>
> Write an essay in which you explain how [the author] builds an argument to persuade [his/her] audience that [author's claim]. In your essay, analyze how [the author] uses one or more of the features listed above (or features of your own choice) to strengthen the logic and persuasiveness of [his/her] argument. Be sure that your analysis focuses on the most relevant features of the passage. Your essay should not explain whether you agree with [the author's] claims, but rather explain how the author builds an argument to persuade [his/her] audience." (College Board, N.D.)

Two readers score each essay, assigning a score from 1 to 4 to each of the Reading, Analysis, and Writing dimensions. Unscorable essays, such as those that are off-topic or written in a language other than English, receive a score of 0. The Reading score assesses the evidence in the essay that the test taker understood the passage, including the interplay of the main themes and the important details. The Analysis score reflects evidence in the essay that the test taker understands how the author builds an argument, including the author's use of evidence, reasoning, and persuasion. A high Writing score is given to essays that are focused, organized, and precise; that show a command of language, including the conventions of standard written English; and that have a variety of sentence structures and consistent, precise word choice.

For each dimension, the two rater scores are added to form the reported score. If one rater gives an essay a score of 0 or the two raters' scores differ by more than one point, a third rater scores the essay. The third rater's score is doubled to yield the reported score. If an essay receives a score of 0 on one dimension, then it is scored 0 on all three dimensions.

Characteristics of the Spring 2018 School Day Administration of the SAT in Delaware

Test Forms and Demographic Information

This report summarizes the data at the master form level for SAT form 1. The master form was built with four timed sections (Reading, Writing and Language, Math - No Calculator, and Math - Calculator). More forms were also administered, but fewer than 100 test takers completed those forms, so the results for those forms are not included in this report.

Along with the test questions, each test taker completed several survey and demographic questions, including gender, current grade level (Not yet in 8th grade; 8th grade; 9th grade; 10th grade; 11th grade; 12th grade or higher; No longer in high school; 1st year of college; 2nd year of college), ethnicity (Hispanic or Latino; Cuban; Mexican; Puerto Rican; Other Hispanic or Latino; or Not Hispanic or Latino) or race (American Indian or Alaska Native; Asian; Black or African American; Native Hawaiian or Other Pacific Islander; or White) and first language spoken (English only; English and another language; Another language). The race/ethnicity question was a two-part question worded in the following way:

What is your ethnicity? (You may mark more than one.)

Hispanic or Latino (including Spanish origin)
  Cuban
  Mexican
  Puerto Rican
Other Hispanic or Latino
Not Hispanic or Latino

What is your race? (You may mark more than one.)
American Indian or Alaska Native
Asian (including Indian subcontinent and Philippines origin)
Black or African American (including African and Afro-Caribbean origin)
Native Hawaiian or Other Pacific Islander
White (including Middle Eastern origin)

If a test taker selected more than one race, they were included in the Two or More Races category only.

Description of the Item Analysis Sample

Before completing the analyses contained in this report, the data used in these analyses were cleaned to exclude any test takers who were not included in the accountability file. See Table 3 for the frequency of test takers in the sample for this administration by grade level, first language, and gender. See Table 4 for the frequency of test takers in the target item analysis sample that responded to the race/ethnicity question.

Description of the Test Analyses

Moments and Score Distributions

Test taker performance is described using the first four moments for all score tiers. The mean, standard deviation, skewness, and kurtosis provide a description of the distribution of scores. Subgroup results are only reported for forms for which the subgroup sample size was 5 or more.

Intercorrelations

The Pearson product moment correlation coefficient provides an evaluation of the pairwise linear relationship between the total, section, test, and cross-test scores, and the subscores. The disattenuated, or true score correlations, are the correlations after correcting for attenuation between the two scores. Subgroup results are only reported for forms for which the subgroup sample size was 100 or more. The formulas for calculating the Pearson correlations and disattenuated, or true score, correlations are in Appendix B1 and B2, respectively.

Reliability and Standard Error of Measurement

Reliability is a measure of consistency in test takers' observed scores. Test takers' observed scores may vary for many reasons. This variance can occur, for example, if the test is administered at two different points in time, across different forms of a test, or due to changes in test administration or scoring conditions. There are many different methods to estimate reliability coefficients, including those based on Generalizability Theory, Classical Test Theory, and Structural Equation Modeling. For the SAT Suite, the compound binomial model is used to calculate reliability for scale scores (See Appendix B3). Reliability estimates range from 0-1, with values near 1 indicating more consistency and values near 0 indicating little to no consistency.

Standard error of measurement (SEM) can be considered a measure of inconsistency in test takers' observed scores. An SEM estimate measures the dispersion of measurement errors over repeated measures of a person on the same instrument. SEM estimates are inversely related to reliability estimates. An SEM value is an average across all observed scores while a conditional standard error of measurement (CSEM) is the estimated SEM for a particular (conditioned on) observed score.

Scale score reliability estimates were derived from averaging the CSEM values obtained from the Delaware Spring 2018 School Day administration. See Section 6.1 of the SAT Suite of Assessments Technical Manual for more details on the scale score reliability estimates. The formulas for calculating the scale score reliability and average CSEM estimates are in Appendix B3 of this document. For the scores that were mathematically derived including Math Test, ERW, and Total scores, the root mean squared CSEM (RMS(CSEM)) was calculated.

Standard error of difference (SED) is calculated to assess how much scores must differ in order to reflect the differences in student ability when comparing scores between students for the same measure. If two scores differ by at least SED times 1.65, it is unlikely that the two scores indicate that the two candidates are equal in ability, since this level difference would occur 10 percent of the time or less. The formula for SED is in Appendix B4.

See the Table 5 series for scale score observed and true score correlations, moments, reliability, and average CSEM values for the total group and gender, race/ethnicity, and grade level subgroups for this administration. In the correlation tables, the values above the diagonal represent the true score correlations. The correlations below the diagonal represent the observed score correlations. Subgroup results are only reported for forms for which the subgroup sample size was 100 or more.

Item Completion Rates and Form Speededness

Item completion rates reflect the percentage of test takers reaching an item within each timed section. A reached item is one that has at least one subsequent item within a timed section with a response. Conversely, a not reached item is one that has no subsequent items within a timed section with a response. Test form speededness is evaluated by examining the following:

- the number of items reached by at least 80% of the test takers
- the percentage of test takers completing at least 75% and 90% of each timed section
- the mean and standard deviation of the number of items not reached

Seventy-five (ninety) percent of a timed section is determined by the ceiling of 75% (90%) of the section length. For example, if a section has 47 items, the statistic is calculated as the percentage of test takers completing 36 or more items in the section. The degree of speededness of a test is negligible when 80% of the students reach the last item and all students reach at least 75% of the questions (van der Linden, 2011). However, judgments of appropriateness of timing should be made using all relevant data. See Tables 6 and 7 for the speededness statistics for this administration. Subgroup results are only reported for forms for which the subgroup sample size was 5 or more.

Differential Item Functioning

Differential item functioning (DIF) is a statistical method that examines the performance of reference and focal subgroups for possible statistical bias. Based on the formulas from Dorans and Holland (1993), found in Appendix B5, the Mantel-Haenszel D-DIF (MH D-DIF) statistic is calculated. MH D-DIF values that are not statistically different from zero are classified as *A* items. Items with a p-value that exceeds 1.96 in absolute value and are significantly larger than 1.5 or less than -1.5 are classified as *C* items. The remaining values are classified as *B* items.

For analysis of DIF for gender, the performance of males is compared to the performance of females, with males serving as the reference group and females as the focal group. For analysis of DIF for race/ethnicity group, the performance of White test takers as the reference group is compared to other race/ethnicity focal subgroups. Ethnicity is defined as Hispanic or non-Hispanic and race is defined as American Indian or Alaska Native (AIAN), Asian, Black or African American, Two or More Races, and White. All non-Hispanic respondents are identified as one of the previously listed race categories with Native Hawaiian or Other Pacific Islander classified as Asian. If a test taker selected more than one race, they were included in the Two or More Races category. DIF analysis for a specific group for an item is only completed if the sample sizes for the item are 200 for the focal group and 500 total. The final DIF category for the item was determined by the worst DIF category compared across gender and race/ethnicity DIF categories. Due to the small sample sizes int this report, DIF results are not reported for test takers with accommodations.

Standardized Differences Between Groups

The test taker performance for each subgroup is described using the mean and standard deviation for all score tiers and the standardized mean differences between the focal and reference groups. See Appendix B6 for the formula for the standardized mean difference. Cohen (1988) suggests standardized mean differences equal to 0.20 are small, 0.50 are medium, and 0.80 are large. See the Table 9 series for the standardized mean differences between subgroups with sample sizes of 100 or more for this administration.

Classification Levels

Classification levels are based on ERW and Math Section cut scores that were determined by state leadership based on recommendations from panelists during a multi-state standard setting held in June 2016 (Morgan, Sweeney, Reshetar, Patel, & McCullough, 2016). The cut scores from the standard setting suggest test takers can be classified into four performance levels with level one being the lowest and level four being the highest. Test takers with an ERW score of at least 480 are considered proficient. Test takers with an MSS of at least 530 are considered proficient.

Upon the establishment of classification levels, one may also examine classification statistics (e.g., classification accuracy and classification consistency). Classification accuracy is the agreement between classifications based on the estimated true scores and observed scores. Classification consistency is the agreement between the classification of expected scores and actual observed scores. The classification accuracy and classification consistency decisions are from the BB-CLASS software (Brennan, 2004). The classification statistics are based on the Livingston & Lewis (1995) method which uses a four-parameter beta-binomial model with effective test length. This method is particularly useful for calculating classification accuracy of composite scores, like ERW. See Appendixes B7 – B14 for the formulas related to classification accuracy and classification consistency. Subgroup results are only reported for forms for which the subgroup sample size was 100 or more. See Tables 10-12 for the classification statistics results.

Description of the SAT Essay Analyses

Description of the Sample

This report summarizes the essay results associated with the SAT master forms administered in Spring 2018. Three prompts were administered in the Spring 2018 SAT Essay test, this report summarizes data at the overall level (i.e., aggregating across all forms and all prompts) and select results are also summarized at the prompt level for prompts with 5 or more test takers.

A score of 0 is assigned to unscorable essays, so a score of 0 is excluded in all of the analyses in this report (e.g., Moments, correlation, and interrater reliability analyses), except for the frequency distributions of scores (including all three dimensions).

Moments and Score Distributions

Test taker performance is described using descriptive statistics (i.e., mean, standard deviation, skewness, and kurtosis) and frequency distributions of scores for all three essay dimension scores. All possible combinations of the three essay dimension scores (512 possible combinations for three dimension scores), along with the frequency and percentage of occurrence provide full information on the joint distribution of the three essay dimension scores. See the Table 13 series for the essay score moments and the Table 14-17 series for the frequency distributions, aggregated across prompts and by prompt.

Intercorrelations

The Pearson product moment correlation coefficient provides an evaluation of the pairwise linear relationship between two essay scores or between essay scores and ERW section, Reading Test, and Writing Test scores. The formula for calculating the Pearson correlations is in Appendix B1. See Table 18 for the correlations between essay dimension scores. See Table 19 for the correlations between essay dimension and relevant ERW section, Reading Test, and Writing Test scores.

Reliability and Standard Error of Measurement

As described previously, reliability refers to the consistency with which an instrument measures some attribute of a person or object. In the context of these analyses, reliability refers to the consistency of test takers' observed scores on the essay dimension scores, given no change in actual ability. There are many reasons a person may score higher or lower on the essay test on any given day. These include situational variables, the particular passage associated with the essay, rater fluctuations, and a number of other factors. If we consider these fluctuations in scores to be errors, then reliability is an index of the proportion of the measurement that is not an error. Reliability estimates range from 0 to 1, with reliability estimates near 1 indicating consistent measurement with very little error. Reliability estimates near zero, on the other hand, would indicate fairly random estimates of the attribute. See Appendixes B15-B20 for formulas related to essay reliability, variance, and SEM.

Percentages of Agreement

Percentage of agreement is an index of interrater agreement. It can be expressed as the number of agreements divided by the total observations (see Appendix B18 for the formula). For ordinal and interval data, percentages of close-but-not-exact agreement (e.g., percentage of adjacent scores – where raters are off by 1) can also be computed and, along with percentage of exact agreement, used as measures of interrater agreement. The percentage of agreement does not take into account agreements due to chance. Therefore, it overestimates the level of agreement (Hallgren, 2012). Percentage of agreement results are presented in the Table 20 series and in Table 21.

Correlation Coefficient and Standard Error of Measurement

The correlation coefficient between the scores given by two raters on the same essay dimension scores is another measure of interrater consistency. Interrater reliability is the reliability of a single rater scoring the essay. This reliability estimate focuses on the stability of the essay scores across raters: How much would the results differ if two different raters were to score the same essay for a test taker? Although the reliability coefficient corresponds to a single rater, the estimation of interrater reliability requires that at least two raters score the same essay for the same test taker, so the reliability of the raters can be estimated. The formulas for computing the Pearson correlation coefficient and related statistics are in Appendixes B1 and B16-18. See Table 22 for the correlation and SEM values for two raters for the essay dimension scores.

Simple Kappa Statistic

Cohen's kappa coefficient (simple kappa statistic; Cohen, 1960) is a statistic that measures the interrater agreement between two raters. It computes the observed level of agreement between two raters, while taking into account the possibility of agreement occurring by chance. The observed agreement is defined by cross-tabulating the scores of the two raters, and the agreement expected by chance is defined by the marginal frequencies of each rater's score. The formula for calculating Cohen's kappa coefficient is given in Appendix B19. Possible values for Cohen's kappa coefficient range from -1 to 1, with 1 indicating complete agreement, 0 indicating complete random agreement, and -1 indicating complete disagreement.

Weighted Kappa Statistic

Weighted kappa coefficient (Cohen, 1968) is an alternative statistic that measures the interrater agreement between two raters, while correcting for the possibility of agreement by chance and penalizing the disagreements. This statistic can be applied to ordinal ratings. The weights used to penalize the disagreement are computed based on the magnitude of disagreement. The formula for calculating weighted kappa coefficient is given in Appendix B20. Possible values for weighted kappa coefficient range from -1 to 1, with 1 indicating complete agreement, 0 indicating complete random agreement, and -1 indicating complete disagreement.

See Table 23 for simple and weighted kappa coefficients for the essay dimension scores.

Standardized Differences Between Groups

See the Table 24 series for the standardized mean essay dimension score differences between the reference and focal subgroups for this administration.

Bibliography/References

Brennan, R. L. (2004). BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy. Available from: https://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs/

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Education and Psychological Measurement. 20, 37-46.

Cohen, J. (1968). Weighted Kappa: Nominal Scales Agreement Provision for Scaled Disagreement or Partial Credit. Psychological Bulletin. 70, 213-220.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

College Board. (N.D.). SAT essay. New York, NY: College Board. Retrieved from https://collegereadiness.collegeboard.org/sat/inside-the-test/essay

College Board. (2014). Test specification for the redesigned SAT. New York, NY: College Board. Retrieved from https://collegereadiness.collegeboard.org/pdf/test-specifications-redesigned-sat-1.pdf.

College Board. (2017). SAT Suite of Assessments Technical Manual: Characteristics of the SAT New York, NY: College Board.

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. Belmont, CA: Wadsworth Group/Thomson Learning.

Dorans, N.J. & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland and H. Wainer (Eds.). *Differential Item functioning* (p 35 – 66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. Tutorials in quantitative methods for psychology, 8(1), 23.

Hanson, B. A. & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. Journal of Educational Measurement, 27(4), 345 – 359.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. Journal of Educational Measurement, 32(2), 179–197.

Morgan, D. L., Sweeney, K., Reshetar, R., Patel, P., & McCullough, J. (2016). Final report on the 2016 SAT multi-state standard setting. (Unpublished Technical Report). New York, NY: The College Board.

Schumacker R.E., & Muchinsky P. M. (1996). Disattenuating correlation coefficients. *Rasch* Measurement Transactions, 10(1), 479. Retrieved from the web on January 20, 2016 from http://www.rasch.org/rmt/rmt101g.htm.

van der Linden, W. J. (2011). Test design and speededness. *Journal of Educational Measurement*, 48(1), 44-60.

Tables

Table 1. Score Scales and Number of Items Contributing to Each Score

Scores	SAT Items	 Scale
Test Scores		
Reading	52	10-40
Writing and Language (WL)	44	10-40
Math (MTS)	58	10-40
No Calculator	20	
Calculator	38	
Cross-Test Scores		
Analysis in History/Social Studies (HSS)	35	10-40
Analysis in Science (SCI)	35	10-40
Subscores		
Command of Evidence (COE)	18	1-15
Words in Context (WIC)	18	1-15
Expression of Ideas (EOI)	24	1-15
Standard English Conventions (SEC)	20	1-15
Heart of Algebra (HOA)	19	1-15
Problem Solving and Data Analysis (PSD)	17	1-15
Passport to Advanced Mathematics (PAM)	16	1-15
Section Scores		
Evidence-Based Reading and Writing (ERW)	96	200-800
Math (MSS)	58	200-800
Total	154	400-1600

Table 2. Number and Type of Items per Timed Section

	SAT	
Timed Section	Items	Timing
Reading	52 MC	65
Writing and Language (WL)	44 MC	35
Math Test - No Calculator	15 MC; 5 SPR	25
Math Test - Calculator	30 MC; 8 SPR	55

Table 3. Frequency and Percentage of Test Takers in Item Analysis Sample by Grade Level, First Language, and Gender

Subgroup	n	%
Grade Level		
11th graders	394	99.49
First Language		
English	253	63.89
English and another language	35	8.84
Another language	14	3.54
No response	33	8.33
Missing	61	15.40
Gender		
Male	246	62.12
Female	149	37.63

Only subgroups with sample size >=5 have statistics reported.

Table 4. Frequency and Percentage of Racial/Ethnic Subgroups in Item Analysis Sample

Subgroup	n	%
White	135	34.09
Black or African American	82	20.71
Hispanic	53	13.38
Asian	5	1.26
Two or more races	15	3.79
Other/Missing	103	26.01

Note. If a test taker selected more than one race then they were included in the Two or More Races category. Only subgroups with sample size <= 5 have statistics reported.

Table 5.a. Scale Score Moments, Intercorrelations, and Reliability

	R	WL	MTS	HSS	SCI	COE	WIC	EOI	SEC	HOA	PSD	PAM	ERW	MSS	Total
N = 395															
R	1	0.93	0.80	1.00	1.00	1.00	1.00	0.96	0.90	0.80	0.84	0.68	1.00	0.80	0.98
WL	0.78	1	0.86	1.00	0.97	1.00	1.00	1.00	1.00	0.84	0.88	0.82	1.00	0.86	1.00
MTS	0.68	0.73	1	0.91	0.91	0.87	0.85	0.88	0.83	1.00	1.00	1.00	0.85	1.00	1.00
HSS	0.87	0.82	0.76	1	0.95	1.00	1.00	1.00	0.91	0.84	1.00	0.77	1.00	0.91	1.00
SCI	0.88	0.79	0.75	0.76	1	1.00	1.00	1.00	0.90	0.91	0.95	0.77	1.00	0.91	1.00
COE	0.82	0.76	0.65	0.78	0.79	1	0.96	1.00	0.90	0.87	0.88	0.77	1.00	0.87	1.00
WIC	0.79	0.78	0.63	0.78	0.74	0.62	1	1.00	0.95	0.82	0.89	0.75	1.00	0.85	1.00
EOI	0.75	0.93	0.70	0.82	0.78	0.79	0.79	1	0.97	0.83	0.93	0.80	1.00	0.88	1.00
SEC	0.68	0.89	0.63	0.68	0.66	0.60	0.63	0.69	1	0.84	0.80	0.81	1.00	0.83	0.98
HOA	0.61	0.64	0.88	0.63	0.68	0.59	0.55	0.60	0.58	1	0.89	1.00	0.83	1.00	1.00
PSD	0.66	0.69	0.87	0.78	0.72	0.61	0.61	0.68	0.57	0.64	1	0.90	0.87	1.00	1.00
PAM	0.45	0.54	0.81	0.50	0.50	0.45	0.44	0.50	0.48	0.63	0.56	1	0.76	1.00	1.00
ERW	0.94	0.94	0.74	0.90	0.89	0.84	0.83	0.89	0.84	0.66	0.71	0.53	1	0.85	1.00
MSS	0.68	0.73	1.00	0.76	0.75	0.65	0.63	0.70	0.63	0.88	0.87	0.81	0.74	1	1.00
Total	0.86	0.89	0.94	0.88	0.87	0.79	0.78	0.85	0.78	0.83	0.85	0.72	0.93	0.94	1
Mean	21.72	20.72	20.21	21.02	21.70	6.73	6.01	6.11	5.34	5.82	5.13	6.25	424.35	404.23	828.58
S.D.	4.71	4.71	4.60	5.09	4.99	2.19	2.94	2.56	2.53	2.48	3.02	2.36	88.85	92.01	168.88
Skewness	0.63	0.79	0.92	0.48	0.46	0.91	0.27	0.78	0.91	1.05	0.47	0.51	0.89	0.92	1.01
Kurtosis	0.57	0.59	1.36	0.11	0.28	1.77	-0.33	0.31	0.51	1.62	-0.55	0.71	0.78	1.36	1.20
Reliability	0.84	0.83	0.85	0.81	0.79	0.66	0.64	0.74	0.68	0.70	0.73	0.53	0.91	0.85	0.93
RMS(CSEM)	1.90	1.95	1.77	2.21	2.29	1.28	1.77	1.30	1.42	1.36	1.56	1.62	27.27	35.32	44.62
SED	2.69	2.76	2.50	3.12	3.24	1.80	2.50	1.84	2.01	1.92	2.21	2.29	38.57	49.95	63.11
SED x 1.65	4.44	4.56	4.12	5.15	5.35	2.98	4.12	3.04	3.32	3.17	3.64	3.78	63.63	82.42	104.13

Note. The values above the diagonal represent the true score correlations. The correlations below the diagonal represent the observed score correlations. SED=Standard Error of Difference. Only subgroups with sample size >=100 have statistics reported.

Table 5.b.1. Scale Score Moments, Intercorrelations, and Reliability for Male Test Takers

	R	WL	MTS	HSS	SCI	COE	WIC	EOI	SEC	HOA	PSD	PAM	ERW	MSS	Total
N = 245															
R	1	0.95	0.85	1.00	1.00	1.00	1.00	0.98	0.91	0.84	0.88	0.75	1.00	0.85	1.00
WL	0.79	1	0.87	1.00	0.99	1.00	1.00	1.00	1.00	0.84	0.87	0.86	1.00	0.87	1.00
MTS	0.72	0.74	1	0.95	0.92	0.90	0.90	0.89	0.83	1.00	1.00	1.00	0.87	1.00	1.00
HSS	0.88	0.82	0.80	1	0.97	1.00	1.00	1.00	0.91	0.86	1.00	0.82	1.00	0.95	1.00
SCI	0.89	0.80	0.77	0.78	1	1.00	1.00	1.00	0.89	0.92	0.94	0.83	1.00	0.92	1.00
COE	0.83	0.79	0.68	0.80	0.81	1	0.99	1.00	0.95	0.89	0.90	0.86	1.00	0.90	1.00
WIC	0.81	0.78	0.68	0.79	0.77	0.65	1	1.00	0.95	0.87	0.93	0.81	1.00	0.90	1.00
EOI	0.78	0.93	0.72	0.81	0.81	0.82	0.79	1	0.95	0.84	0.92	0.84	1.00	0.89	1.00
SEC	0.68	0.89	0.63	0.68	0.65	0.63	0.63	0.68	1	0.83	0.78	0.85	1.00	0.83	0.98
HOA	0.65	0.65	0.88	0.66	0.69	0.61	0.60	0.62	0.57	1	0.88	1.00	0.85	1.00	1.00
PSD	0.70	0.69	0.88	0.83	0.73	0.63	0.65	0.70	0.56	0.65	1	0.93	0.89	1.00	1.00
PAM	0.50	0.57	0.82	0.54	0.54	0.51	0.48	0.53	0.51	0.65	0.59	1	0.82	1.00	1.00
ERW	0.95	0.95	0.77	0.90	0.89	0.86	0.84	0.90	0.83	0.68	0.73	0.57	1	0.87	1.00
MSS	0.72	0.74	1.00	0.80	0.77	0.68	0.68	0.72	0.63	0.88	0.88	0.82	0.77	1	1.00
Total	0.88	0.89	0.95	0.90	0.88	0.81	0.80	0.86	0.77	0.83	0.86	0.74	0.94	0.95	1
Mean	21.36	20.53	20.24	20.88	21.49	6.65	5.76	6.02	5.24	5.80	5.22	6.19	418.98	404.90	823.88
S.D.	4.70	4.74	4.81	5.23	5.07	2.17	3.03	2.62	2.47	2.54	3.11	2.38	89.47	96.13	174.68
Skewness	0.72	0.77	0.89	0.48	0.49	0.96	0.36	0.78	0.90	1.10	0.38	0.65	0.91	0.89	0.99
Kurtosis	0.72	0.51	1.17	-0.01	0.50	2.13	-0.42	0.22	0.56	1.74	-0.77	1.00	0.84	1.17	1.17
Reliability	0.83	0.83	0.87	0.82	0.80	0.66	0.66	0.76	0.67	0.71	0.75	0.53	0.91	0.87	0.93
RMS(CSEM)	1.92	1.95	1.76	2.21	2.29	1.26	1.77	1.29	1.42	1.36	1.55	1.62	27.40	35.18	44.59
SED	2.72	2.76	2.49	3.13	3.23	1.79	2.51	1.82	2.01	1.92	2.19	2.30	38.75	49.75	63.06
SED x 1.65	4.48	4.56	4.10	5.16	5.34	2.95	4.14	3.01	3.32	3.17	3.61	3.79	63.94	82.08	104.05

Note. The values above the diagonal represent the true score correlations. The correlations below the diagonal represent the observed score correlations. SED=Standard Error of Difference. Only subgroups with sample size >=100 have statistics reported.

Table 5.b.2. Scale Score Moments, Intercorrelations, and Reliability for Female Test Takers

	R	WL	MTS	HSS	SCI	COE	WIC	EOI	SEC	HOA	PSD	PAM	ERW	MSS	Total
N = 149															
R	1	0.90	0.73	1.00	1.00	1.00	1.00	0.92	0.89	0.74	0.78	0.55	1.00	0.73	0.95
WL	0.75	1	0.85	1.00	0.95	0.94	1.00	1.00	1.00	0.83	0.90	0.73	1.00	0.85	1.00
MTS	0.60	0.70	1	0.84	0.90	0.82	0.77	0.86	0.84	1.00	1.00	1.00	0.81	1.00	1.00
HSS	0.86	0.82	0.68	1	0.93	1.00	1.00	1.00	0.92	0.79	0.94	0.66	1.00	0.84	1.00
SCI	0.87	0.76	0.72	0.73	1	1.00	1.00	0.98	0.92	0.89	0.99	0.67	1.00	0.90	1.00
COE	0.80	0.70	0.60	0.74	0.76	1	0.92	1.00	0.82	0.83	0.87	0.61	1.00	0.82	0.98
WIC	0.77	0.77	0.54	0.78	0.69	0.58	1	1.00	0.97	0.72	0.86	0.64	1.00	0.77	1.00
EOI	0.71	0.93	0.66	0.82	0.73	0.73	0.79	1	0.99	0.80	0.95	0.71	1.00	0.86	1.00
SEC	0.68	0.90	0.64	0.69	0.68	0.56	0.63	0.70	1	0.86	0.85	0.73	1.00	0.84	1.00
HOA	0.56	0.62	0.88	0.58	0.65	0.55	0.46	0.56	0.60	1	0.91	0.99	0.80	1.00	1.00
PSD	0.60	0.68	0.85	0.70	0.73	0.59	0.55	0.67	0.60	0.62	1	0.85	0.86	1.00	1.00
PAM	0.36	0.48	0.79	0.43	0.42	0.36	0.36	0.43	0.44	0.58	0.51	1	0.66	1.00	0.97
ERW	0.94	0.94	0.70	0.90	0.87	0.80	0.82	0.88	0.85	0.63	0.68	0.45	1	0.81	1.00
MSS	0.60	0.70	1.00	0.68	0.72	0.60	0.54	0.66	0.64	0.88	0.85	0.79	0.70	1	1.00
Total	0.84	0.89	0.92	0.86	0.86	0.76	0.74	0.84	0.81	0.81	0.83	0.67	0.92	0.92	1
Mean	22.32	21.05	20.19	21.26	22.11	6.86	6.43	6.25	5.52	5.87	5.01	6.37	433.69	403.89	837.58
S.D.	4.68	4.67	4.25	4.89	4.83	2.23	2.75	2.47	2.63	2.39	2.87	2.33	87.45	84.95	158.92
Skewness	0.49	0.85	0.98	0.50	0.43	0.83	0.18	0.82	0.89	0.96	0.63	0.27	0.89	0.98	1.11
Kurtosis	0.52	0.78	1.76	0.37	-0.04	1.31	0.01	0.50	0.42	1.46	-0.06	0.34	0.81	1.76	1.32
Reliability	0.84	0.82	0.82	0.80	0.77	0.66	0.60	0.71	0.71	0.68	0.69	0.52	0.90	0.82	0.92
RMS(CSEM)	1.87	1.95	1.78	2.19	2.30	1.30	1.74	1.32	1.42	1.36	1.59	1.62	27.05	35.54	44.66
SED	2.65	2.76	2.51	3.10	3.25	1.83	2.47	1.87	2.01	1.92	2.24	2.28	38.26	50.26	63.16
SED x 1.65	4.37	4.56	4.15	5.11	5.36	3.03	4.07	3.08	3.32	3.17	3.70	3.77	63.12	82.93	104.22

Note. The values above the diagonal represent the true score correlations. The correlations below the diagonal represent the observed score correlations. SED=Standard Error of Difference. Only subgroups with sample size >=100 have statistics reported.

Table 5.c.1. Scale Score Moments, Intercorrelations, and Reliability for White Test Takers

	R	WL	MTS	HSS	SCI	COE	WIC	EOI	SEC	HOA	PSD	PAM	ERW	MSS	Total
								N = 134							
R	1	0.92	0.75	1.00	1.00	1.00	1.00	0.96	0.89	0.71	0.81	0.66	1.00	0.75	0.95
WL	0.78	1	0.85	0.99	0.98	0.94	1.00	1.00	1.00	0.78	0.87	0.84	1.00	0.85	1.00
MTS	0.65	0.74	1	0.88	0.93	0.85	0.71	0.86	0.83	1.00	1.00	1.00	0.82	1.00	1.00
HSS	0.90	0.84	0.75	1	1.00	1.00	1.00	1.00	0.94	0.80	0.98	0.75	1.00	0.88	1.00
SCI	0.89	0.81	0.77	0.83	1	1.00	1.00	1.00	0.93	0.88	0.99	0.80	1.00	0.93	1.00
COE	0.82	0.72	0.67	0.76	0.81	1	0.84	1.00	0.84	0.81	0.86	0.83	1.00	0.85	0.98
WIC	0.81	0.79	0.56	0.77	0.75	0.58	1	1.00	0.97	0.66	0.79	0.58	1.00	0.71	0.92
EOI	0.77	0.94	0.71	0.83	0.80	0.75	0.78	1	1.00	0.80	0.90	0.82	1.00	0.86	1.00
SEC	0.70	0.92	0.68	0.73	0.71	0.60	0.70	0.75	1	0.77	0.82	0.85	1.00	0.83	0.98
HOA	0.58	0.64	0.91	0.65	0.69	0.60	0.49	0.61	0.59	1	0.89	1.00	0.76	1.00	0.98
PSD	0.65	0.70	0.87	0.77	0.76	0.62	0.57	0.67	0.61	0.68	1	0.87	0.86	1.00	1.00
PAM	0.48	0.61	0.84	0.53	0.56	0.54	0.38	0.56	0.58	0.71	0.59	1	0.77	1.00	1.00
ERW	0.94	0.95	0.74	0.92	0.90	0.82	0.84	0.91	0.87	0.65	0.71	0.58	1	0.82	1.00
MSS	0.65	0.74	1.00	0.75	0.77	0.67	0.56	0.71	0.68	0.91	0.87	0.84	0.74	1	1.00
Total	0.85	0.90	0.94	0.89	0.89	0.79	0.75	0.86	0.82	0.84	0.85	0.77	0.93	0.94	1
Mean	23.90	22.75	22.42	23.31	23.90	7.66	7.25	7.20	6.28	6.84	6.44	7.11	466.49	448.43	914.93
S.D.	4.74	5.06	5.04	5.13	4.85	2.36	2.99	2.66	2.85	2.90	3.12	2.59	92.42	100.84	180.32
Skewness	0.51	0.41	0.77	0.34	0.39	0.63	-0.06	0.36	0.41	0.73	0.07	0.64	0.60	0.77	0.73
Kurtosis	-0.19	-0.41	0.67	-0.49	-0.39	0.89	-0.16	-0.43	-0.55	0.37	-0.63	0.45	-0.38	0.67	0.05
Reliability	0.85	0.85	0.88	0.83	0.79	0.69	0.70	0.75	0.74	0.78	0.75	0.62	0.92	0.88	0.94
RMS(CSEM)	1.83	1.93	1.72	2.11	2.24	1.31	1.64	1.32	1.45	1.35	1.58	1.60	26.64	34.37	43.49
SED	2.59	2.73	2.43	2.99	3.17	1.86	2.33	1.87	2.06	1.91	2.23	2.26	37.68	48.61	61.50
SED x 1.65	4.28	4.51	4.01	4.93	5.23	3.07	3.84	3.08	3.39	3.15	3.68	3.73	62.17	80.20	101.47

Note. The values above the diagonal represent the true score correlations. The correlations below the diagonal represent the observed score correlations. SED=Standard Error of Difference. Only subgroups with sample size >=100 have statistics reported.

Table 6. Item Level Completion Rates

Item Number	Reading	Writing and Language	Math-No Calculator	Math-Calculator
1	100.00	99.24	98.99	97.72
2	100.00	99.24	98.99	97.47
3	100.00	99.24	98.99	97.47
4	100.00	99.24	98.73	97.47
5	100.00	99.24	98.48	97.47
6	99.75	99.24	98.48	97.47
7	99.75	99.24	98.48	97.47
8	99.75	99.24	98.23	97.22
9	99.75	98.99	98.23	96.96
10	99.75	98.99	98.23	96.96
11	99.49	98.99	98.23	96.96
12	99.49	98.99	97.47	96.96
13	99.49	98.73	97.47	96.71
14	99.49	98.73	97.47	96.46
15	99.49	98.73	97.22	96.46
16	98.99	98.48	86.33	96.46
17	98.99	98.48	81.27	96.20
18	98.73	98.48	76.96	96.20
19	98.23	98.23	75.95	95.95
20	98.23	97.72	71.90	95.70
21	97.97	97.72	-	94.94
22	97.97	97.47	-	94.43
23	97.97	97.22	-	94.18
24	97.47	97.22	-	93.92
25	97.47	96.71	-	93.92
26	96.96	96.71	-	93.67
27	96.71	96.71	-	93.67
28	96.20	96.71	-	93.67
29	95.95	96.20	-	93.67
30	95.70	95.95	-	93.16
31	94.43	94.94	-	84.56
32	94.18	94.43	-	83.04
33	94.18	93.92	-	82.03
34	93.92	93.42	-	82.03
35	93.42	93.16	-	78.99
36	93.16	92.41	-	76.71
37	92.66	92.41	-	75.19
38	92.41	91.90	-	69.11
39	91.90	90.89	-	-
40	91.90	90.63	-	-
41	91.90	89.87	-	-
42	90.38	88.35	-	-
43	89.87	88.10	-	-
44	89.87	88.10	-	-
45	89.37	-	-	-
46	89.11	-	-	-
47	88.86	-	-	-
48	88.61	-	-	-
49	88.61	-	-	-

Table 6. Item Level Completion Rates

Item Number	Reading	Writing and Language	Math-No Calculator	Math-Calculator
50	88.61	-	-	-
51	88.35	-	-	-
52	88.10	-	-	-

Table 7.a. Section Completion Rates by Timed Section

Test	Category	N=395
Reading	# Items Reached by 80%	52
	# Items in Section	52
	% Completing 75%	92.15
	% Completing 90%	89.11
	% Completing Section	88.35
	Mean Not Reached	2.36
	S.D. Not Reached	7.41
Writing and Language	# Items Reached by 80%	44
	# Items in Section	44
	% Completing 75%	94.18
	% Completing 90%	90.89
	% Completing Section	88.35
	Mean Not Reached	1.71
	S.D. Not Reached	6.09
Math-No Calculator	# Items Reached by 80%	17
	# Items in Section	20
	% Completing 75%	97.47
	% Completing 90%	77.22
	% Completing Section	72.15
	Mean Not Reached	1.34
	S.D. Not Reached	2.96
Math-Calculator	# Items Reached by 80%	34
	# Items in Section	38
	% Completing 75%	93.92
	% Completing 90%	79.24
	% Completing Section	69.37
	Mean Not Reached	2.91
	S.D. Not Reached	7.24

Table 7.b. Section Completion Rates by Gender

Test	Category	Male (N=245)	Female (N=149)
Reading	# Items Reached by 80%	52	52
	# Items in Section	52	52
	% Completing 75%	92.65	91.95
	% Completing 90%	90.20	87.92
	% Completing Section	89.80	86.58
	Mean Not Reached	2.34	2.26
	S.D. Not Reached	7.60	6.95
Writing and Language	# Items Reached by 80%	44	44
	# Items in Section	44	44
	% Completing 75%	94.29	94.63
	% Completing 90%	91.84	89.93
	% Completing Section	89.39	87.25
	Mean Not Reached	1.76	1.54
	S.D. Not Reached	6.49	5.31
Math-No Calculator	# Items Reached by 80%	17	17
	# Items in Section	20	20
	% Completing 75%	97.55	97.32
	% Completing 90%	77.96	76.51
	% Completing Section	72.24	72.48
	Mean Not Reached	1.38	1.23
	S.D. Not Reached	3.15	2.63
Math-Calculator	# Items Reached by 80%	34	35
	# Items in Section	38	38
	% Completing 75%	93.88	94.63
	% Completing 90%	76.73	83.89
	% Completing Section	68.16	71.81
	Mean Not Reached	3.28	2.19
	S.D. Not Reached	7.96	5.74

Only subgroups with sample size >=5 have statistics reported.

Table 7.c. Section Completion Rates by Race/Ethnicity

Test	Category	White (N=135)	Black (N=82)	Hispanic (N=53)	Asian (N=5)	NHPI (N=1)	AIAN (N=2)	Two or More Races (N=15)
Reading	# Items Reached by 80%	52	52	52	52	–	–	52
	# Items in Section	52	52	52	52	–	–	52
	% Completing 75%	99.25	85.37	86.79	100.00	–	–	93.33
	% Completing 90%	97.01	82.93	86.79	80.00	–	–	86.67
	% Completing Section	96.27	81.71	86.79	80.00	–	–	86.67
	Mean Not Reached	0.59	4.24	3.25	2.20	–	–	2.80
	S.D. Not Reached	3.29	10.13	8.78	4.92	–	–	8.87
Writing and Language	# Items Reached by 80%	44	44	44	44	–	–	44
	# Items in Section	44	44	44	44	–	–	44
	% Completing 75%	100.00	89.02	90.57	100.00	–	–	86.67
	% Completing 90%	97.76	82.93	86.79	80.00	–	–	86.67
	% Completing Section	94.78	81.71	83.02	80.00	–	–	86.67
	Mean Not Reached	0.58	3.01	2.47	1.20	–	–	3.80
	S.D. Not Reached	3.95	7.60	6.74	2.68	–	–	11.61
Math-No Calculator	# Items Reached by 80%	20	15	19	20	–	–	15
	# Items in Section	20	20	20	20	–	–	20
	% Completing 75%	99.25	96.34	96.23	100.00	–	–	93.33
	% Completing 90%	88.81	62.20	83.02	100.00	–	–	73.33
	% Completing Section	84.33	54.88	77.36	80.00	–	–	66.67
	Mean Not Reached	0.75	2.22	1.09	0.40	–	–	2.40
	S.D. Not Reached	2.54	3.72	2.73	0.89	–	–	5.28
Math-Calculator	# Items Reached by 80%	38	30	34	37	–	–	30
	# Items in Section	38	38	38	38	–	–	38
	% Completing 75%	100.00	87.80	86.79	100.00	–	–	93.33
	% Completing 90%	92.54	68.29	79.25	80.00	–	–	73.33
	% Completing Section	84.33	58.54	69.81	60.00	–	–	73.33
	Mean Not Reached	0.93	4.85	4.36	1.80	–	–	4.13
	S.D. Not Reached	3.70	9.77	9.91	3.49	–	–	9.93

Note. AIAN stands for American Indian/Alaska Native, NHPI stands for Native Hawaiian or other Pacific Islander. Only subgroups with sample size >=5 have statistics reported.

Table 9.a. Scale Score Mean, Standard Deviation, and Standardized Difference between Gender Groups

		Male			Female		
Score	N	Mean	S.D.	N	Mean	S.D.	Std. Diff.
R	245	21.36	4.70	149	22.32	4.68	0.20
WL		20.53	4.74		21.05	4.67	0.11
MTS		20.24	4.81		20.19	4.25	-0.01
HSS		20.88	5.23		21.26	4.89	0.07
SCI		21.49	5.07		22.11	4.83	0.12
COE		6.65	2.17		6.86	2.23	0.10
WIC		5.76	3.03		6.43	2.75	0.23
EOI		6.02	2.62		6.25	2.47	0.09
SEC		5.24	2.47		5.52	2.63	0.11
HOA		5.80	2.54		5.87	2.39	0.03
PSD		5.22	3.11		5.01	2.87	-0.07
PAM		6.19	2.38		6.37	2.33	0.08
ERW		418.98	89.47		433.69	87.45	0.17
MSS		404.90	96.13		403.89	84.95	-0.01
Total		823.88	174.68		837.58	158.92	0.08

Note. Std. Diff.=Standardized Difference for female mean - male mean. Only subgroups with sample size >=100 have statistics reported.

Table 10. Percentage of Test Takers in Each Classification Level for SAT by Subgroup

		Evidence-Based Reading and Writing				Math			
Level		**Level 1**	**Level 2**	**Level 3**	**Level 4**	**Level 1**	**Level 2**	**Level 3**	**Level 4**
Score Range	**N**	**200-410**	**420-470**	**480-620**	**630-800**	**200-410**	**420-520**	**530-640**	**650-800**
Grade Level									
All	395	56.20	20.76	19.24	3.80	65.06	25.57	6.58	2.78
Gender									
Male	245	57.55	20.41	17.96	4.08	63.67	25.71	7.35	3.27
Female	149	53.69	21.48	21.48	3.36	67.11	25.50	5.37	2.01
Race/Ethnicity									
White	134	34.33	28.36	29.85	7.46	44.78	36.57	11.94	6.72
Black or African American	82	80.49	10.98	8.54	0.00	79.27	18.29	2.44	0.00
Hispanic	53	58.49	22.64	16.98	1.89	73.58	20.75	5.66	0.00
Other/Missing	99	67.68	19.19	12.12	1.01	73.74	22.22	4.04	0.00

Note. Classification levels are not reported for groups with less than 30 test takers.

Table 11. Classification Accuracy

	Evidence-Based Reading and Writing			Math		
	Probability of correct classification	False positive	False negative	Probability of correct classification	False positive	False negative
Grade Level						
All	0.84	0.09	0.07	0.85	0.09	0.06
Gender						
Male	0.84	0.09	0.07	0.85	0.09	0.06
Female	0.82	0.10	0.08	0.85	0.09	0.06
Race/Ethnicity						
White	0.80	0.11	0.09	0.80	0.12	0.08
Individual cut points						
Level 1 vs. Level 2 - 4	0.91	0.05	0.04	0.90	0.06	0.05
Level 1 - 2 vs. Level 3 - 4	0.94	0.03	0.03	0.96	0.02	0.02
Level 1 - 3 vs. Level 4	0.99	0.01	0.00	0.99	0.01	0.00

Note. Classification accuracy is reported for groups with more than 100 test takers.

Table 12. Classification Consistency

	Evidence-Based Reading and Writing				Math			
	Proportion of consistent decisions	Chance proportion of consistent decision	Kappa Statistic	Probability of misclassification	Proportion of consistent decisions	Chance proportion of consistent decision	Kappa Statistic	Probability of misclassification
Grade Level								
All	0.78	0.40	0.63	0.22	0.78	0.49	0.58	0.22
Gender								
Male	0.78	0.41	0.63	0.22	0.79	0.48	0.60	0.21
Female	0.76	0.38	0.61	0.24	0.78	0.52	0.55	0.22
Race/Ethnicity								
White	0.72	0.30	0.61	0.28	0.72	0.35	0.57	0.28
Individual cut points								
Level 1 vs. Level 2 - 4	0.87	0.51	0.75	0.13	0.85	0.54	0.68	0.15
Level 1 - 2 vs. Level 3 - 4	0.92	0.64	0.77	0.08	0.94	0.83	0.67	0.06
Level 1 - 3 vs. Level 4	0.98	0.93	0.73	0.02	0.98	0.95	0.64	0.02

Note. Classification consistency is reported for groups with more than 100 test takers.

Table 13.a. Descriptive Statistics for Essay Dimension Scores

Score	Mean	SD	Skewness	Kurtosis
Reading				
Rating 1	1.59	0.75	1.00	0.05
Rating 2	1.58	0.71	0.86	-0.37
Dimension Score	3.16	1.34	0.96	-0.10
Analysis				
Rating 1	1.27	0.54	2.02	3.71
Rating 2	1.23	0.50	2.15	3.82
Dimension Score	2.49	0.95	2.00	3.52
Writing				
Rating 1	1.83	0.76	0.44	-0.71
Rating 2	1.86	0.76	0.47	-0.49
Dimension Score	3.68	1.42	0.46	-0.63
N	367	367	367	367

Note: Dimension scores of zero were excluded from the computation of all four moments. For each dimension, the two rater scores are added to form the dimension score. If the two raters' scores differ by more than one point, then a third rater scores the Essay. The third rater's score is doubled to yield the dimension score.

Table 13.b.1. Descriptive Statistics for Essay Dimension Scores for Prompt 1

Score	Mean	SD	Skewness	Kurtosis
Reading				
Rating 1	1.51	0.66	1.15	1.12
Rating 2	1.50	0.64	0.93	-0.20
Dimension Score	3.03	1.18	1.13	0.94
Analysis				
Rating 1	1.30	0.52	1.52	1.44
Rating 2	1.27	0.51	1.73	2.20
Dimension Score	2.57	0.96	1.53	1.38
Writing				
Rating 1	1.74	0.71	0.42	-0.91
Rating 2	1.80	0.72	0.65	0.27
Dimension Score	3.54	1.31	0.58	-0.28
N	100	100	100	100

Note: Dimension scores of zero were excluded from the computation of all four moments. For each dimension, the two rater scores are added to form the dimension score. If the two raters' scores differ by more than one point, then a third rater scores the Essay. The third rater’s score is doubled to yield the dimension score.

Table 13.b.2. Descriptive Statistics for Essay Dimension Scores for Prompt 2

Score	Mean	SD	Skewness	Kurtosis
Reading				
Rating 1	1.60	0.77	0.92	-0.31
Rating 2	1.59	0.72	0.85	-0.37
Dimension Score	3.18	1.37	0.90	-0.37
Analysis				
Rating 1	1.25	0.54	2.22	4.74
Rating 2	1.22	0.51	2.25	4.21
Dimension Score	2.48	0.97	2.15	4.22
Writing				
Rating 1	1.88	0.79	0.42	-0.70
Rating 2	1.90	0.77	0.34	-0.78
Dimension Score	3.76	1.45	0.36	-0.76
N	249	249	249	249

Note: Dimension scores of zero were excluded from the computation of all four moments. For each dimension, the two rater scores are added to form the dimension score. If the two raters' scores differ by more than one point, then a third rater scores the Essay. The third rater’s score is doubled to yield the dimension score.

Table 13.b.3. Descriptive Statistics for Essay Dimension Scores for Prompt 3

Score	Mean	SD	Skewness	Kurtosis
Reading				
Rating 1	1.75	0.93	1.13	0.68
Rating 2	1.75	0.77	0.49	-1.06
Dimension Score	3.50	1.59	0.85	-0.14
Analysis				
Rating 1	1.25	0.58	2.38	5.31
Rating 2	1.06	0.25	4.00	16.00
Dimension Score	2.19	0.54	3.03	9.09
Writing				
Rating 1	1.75	0.77	0.49	-1.06
Rating 2	1.63	0.89	1.55	2.28
Dimension Score	3.38	1.54	1.14	0.66
N	16	16	16	16

Note: Dimension scores of zero were excluded from the computation of all four moments. For each dimension, the two rater scores are added to form the dimension score. If the two raters' scores differ by more than one point, then a third rater scores the Essay. The third rater’s score is doubled to yield the dimension score.

Table 14.a. Frequency Distributions of the Three Essay Dimension Scores

	Essay Reading		Essay Analysis		Essay Writing	
Score	**Freq**	**Percent**	**Freq**	**Percent**	**Freq**	**Percent**
0	47	11.35	47	11.35	47	11.35
2	167	40.34	272	65.70	105	25.36
3	72	17.39	36	8.70	59	14.25
4	68	16.43	40	9.66	112	27.05
5	27	6.52	13	3.14	40	9.66
6	29	7.00	5	1.21	42	10.14
7	4	0.97	1	0.24	8	1.93
8	0	0.00	0	0.00	1	0.24
Total	*414*	*100.00*	*414*	*100.00*	*414*	*100.00*

Table 14.b.1. Frequency Distributions of the Three Essay Dimension Scores for Prompt 1

	Essay Reading		Essay Analysis		Essay Writing	
Score	Freq	Percent	Freq	Percent	Freq	Percent
0	22	18.03	22	18.03	22	18.03
2	45	36.89	69	56.56	28	22.95
3	22	18.03	11	9.02	21	17.21
4	25	20.49	15	12.30	32	26.23
5	2	1.64	4	3.28	9	7.38
6	5	4.10	1	0.82	8	6.56
7	1	0.82	0	0.00	2	1.64
8	0	0.00	0	0.00	0	0.00
Total	*122*	*100.00*	*122*	*100.00*	*122*	*100.00*

Table 14.b.2. Frequency Distributions of the Three Essay Dimension Scores for Prompt 2

	Essay Reading		Essay Analysis		Essay Writing	
Score	**Freq**	**Percent**	**Freq**	**Percent**	**Freq**	**Percent**
0	23	8.46	23	8.46	23	8.46
2	115	42.28	188	69.12	70	25.74
3	47	17.28	23	8.46	34	12.50
4	40	14.71	24	8.82	76	27.94
5	23	8.46	9	3.31	30	11.03
6	22	8.09	4	1.47	33	12.13
7	2	0.74	1	0.37	5	1.84
8	0	0.00	0	0.00	1	0.37
Total	*272*	*100.00*	*272*	*100.00*	*272*	*100.00*

Table 14.b.3. Frequency Distributions of the Three Essay Dimension Scores for Prompt 3

	Essay Reading		Essay Analysis		Essay Writing	
Score	Freq	Percent	Freq	Percent	Freq	Percent
0	2	11.11	2	11.11	2	11.11
2	6	33.33	14	77.78	6	33.33
3	3	16.67	1	5.56	4	22.22
4	3	16.67	1	5.56	3	16.67
5	2	11.11	0	0.00	1	5.56
6	1	5.56	0	0.00	1	5.56
7	1	5.56	0	0.00	1	5.56
8	0	0.00	0	0.00	0	0.00
Total	*18*	*100.00*	*18*	*100.00*	*18*	*100.00*

Table 15.a. Frequency Distributions of the Three Essay Dimension Scores by Rater

	Essay Reading				Essay Analysis				Essay Writing			
	Rater Set 1		Rater Set 2		Rater Set 1		Rater Set 2		Rater Set 1		Rater Set 2	
Score	Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent
0	47	11.35	47	11.35	47	11.35	47	11.35	47	11.35	47	11.35
1	206	49.76	201	48.55	285	68.84	297	71.74	139	33.57	131	31.64
2	111	26.81	121	29.23	67	16.18	56	13.53	155	37.44	164	39.61
3	46	11.11	44	10.63	14	3.38	14	3.38	69	16.67	66	15.94
4	4	0.97	1	0.24	1	0.24	0	0.00	4	0.97	6	1.45
Total	*414*	*100.00*	*414*	*100.00*	*414*	*100.00*	*414*	*100.00*	*414*	*100.00*	*414*	*100.00*

Table 15.b.1. Frequency Distributions of the Three Essay Dimension Scores by Rater for Prompt 1

	Essay Reading				Essay Analysis				Essay Writing			
	Rater Set 1		Rater Set 2		Rater Set 1		Rater Set 2		Rater Set 1		Rater Set 2	
Score	Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent
0	22	18.03	22	18.03	22	18.03	22	18.03	22	18.03	22	18.03
1	57	46.72	58	47.54	73	59.84	76	62.30	41	33.61	36	29.51
2	36	29.51	34	27.87	24	19.67	21	17.21	44	36.07	50	40.98
3	6	4.92	8	6.56	3	2.46	3	2.46	15	12.30	12	9.84
4	1	0.82	0	0.00	0	0.00	0	0.00	0	0.00	2	1.64
Total	*122*	*100.00*	*122*	*100.00*	*122*	*100.00*	*122*	*100.00*	*122*	*100.00*	*122*	*100.00*

Table 15.b.2. Frequency Distributions of the Three Essay Dimension Scores by Rater for Prompt 2

	Essay Reading				Essay Analysis				Essay Writing			
	Rater Set 1		Rater Set 2		Rater Set 1		Rater Set 2		Rater Set 1		Rater Set 2	
Score	Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent
0	23	8.46	23	8.46	23	8.46	23	8.46	23	8.46	23	8.46
1	140	51.47	135	49.63	198	72.79	204	75.00	90	33.09	85	31.25
2	70	25.74	81	29.78	40	14.71	34	12.50	104	38.24	108	39.71
3	37	13.60	32	11.76	10	3.68	11	4.04	51	18.75	53	19.49
4	2	0.74	1	0.37	1	0.37	0	0.00	4	1.47	3	1.10
Total	*272*	*100.00*	*272*	*100.00*	*272*	*100.00*	*272*	*100.00*	*272*	*100.00*	*272*	*100.00*

Table 15.b.3. Frequency Distributions of the Three Essay Dimension Scores by Rater for Prompt 3

	Essay Reading				Essay Analysis				Essay Writing			
	Rater Set 1		Rater Set 2		Rater Set 1		Rater Set 2		Rater Set 1		Rater Set 2	
Score	Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent
0	2	11.11	2	11.11	2	11.11	2	11.11	2	11.11	2	11.11
1	8	44.44	7	38.89	13	72.22	15	83.33	7	38.89	9	50.00
2	5	27.78	6	33.33	2	11.11	1	5.56	6	33.33	5	27.78
3	2	11.11	3	16.67	1	5.56	0	0.00	3	16.67	1	5.56
4	1	5.56	0	0.00	0	0.00	0	0.00	0	0.00	1	5.56
Total	*18*	*100.00*	*18*	*100.00*	*18*	*100.00*	*18*	*100.00*	*18*	*100.00*	*18*	*100.00*

Table 16.a. Frequency Distributions of Observed Combinations of the Three Essay Dimension Scores

Essay Reading	Essay Analysis	Essay Writing	Freq	Percent
0	0	0	47	11.35
2	2	2	93	22.46
2	2	3	28	6.76
2	2	4	31	7.49
2	3	3	1	0.24
2	3	4	6	1.45
2	3	5	1	0.24
2	3	6	2	0.48
2	4	3	2	0.48
2	4	4	3	0.72
3	2	2	10	2.42
3	2	3	20	4.83
3	2	4	23	5.56
3	2	5	3	0.72
3	2	6	1	0.24
3	3	2	1	0.24
3	3	4	6	1.45
3	3	6	1	0.24
3	4	4	3	0.72
3	4	5	3	0.72
3	5	5	1	0.24
4	2	2	1	0.24
4	2	3	8	1.93
4	2	4	28	6.76
4	2	5	7	1.69
4	2	6	4	0.97
4	3	4	1	0.24
4	3	5	2	0.48
4	3	6	1	0.24
4	4	4	7	1.69
4	4	5	3	0.72
4	4	6	2	0.48
4	5	5	4	0.97
5	2	4	1	0.24
5	2	5	3	0.72
5	3	5	4	0.97
5	3	6	5	1.21
5	4	4	1	0.24
5	4	5	6	1.45
5	4	6	4	0.97
5	5	5	1	0.24
5	5	6	2	0.48
6	2	4	1	0.24
6	2	5	2	0.48
6	2	6	8	1.93
6	3	4	1	0.24
6	3	6	2	0.48
6	4	6	5	1.21

Table 16.a. Frequency Distributions of Observed Combinations of the Three Essay Dimension Scores

Essay Reading	Essay Analysis	Essay Writing	Freq	Percent
6	4	7	1	0.24
6	5	6	3	0.72
6	6	6	1	0.24
6	6	7	3	0.72
6	6	8	1	0.24
6	7	7	1	0.24
7	3	6	1	0.24
7	3	7	1	0.24
7	5	7	2	0.48
Total			*414*	*100.00*

Table 16.b.1. Frequency Distributions of Observed Combinations of the Three Essay Dimension Scores for Prompt 1

Essay Reading	Essay Analysis	Essay Writing	Freq	Percent
0	0	0	22	18.03
2	2	2	24	19.67
2	2	3	7	5.74
2	2	4	6	4.92
2	3	3	1	0.82
2	3	4	2	1.64
2	3	6	1	0.82
2	4	3	2	1.64
2	4	4	2	1.64
3	2	2	4	3.28
3	2	3	8	6.56
3	2	4	3	2.46
3	2	5	1	0.82
3	3	4	4	3.28
3	4	4	1	0.82
3	4	5	1	0.82
4	2	3	3	2.46
4	2	4	7	5.74
4	2	5	2	1.64
4	2	6	2	1.64
4	3	4	1	0.82
4	3	6	1	0.82
4	4	4	5	4.10
4	4	5	1	0.82
4	4	6	1	0.82
4	5	5	2	1.64
5	3	6	1	0.82
5	4	4	1	0.82
6	2	5	2	1.64
6	4	6	1	0.82
6	5	6	1	0.82
6	6	7	1	0.82
7	5	7	1	0.82

Table 16.b.2. Frequency Distributions of Observed Combinations of the Three Essay Dimension Scores for Prompt 2

Essay Reading	Essay Analysis	Essay Writing	Freq	Percent
0	0	0	23	8.46
2	2	2	63	23.16
2	2	3	20	7.35
2	2	4	25	9.19
2	3	4	4	1.47
2	3	5	1	0.37
2	3	6	1	0.37
2	4	4	1	0.37
3	2	2	5	1.84
3	2	3	11	4.04
3	2	4	19	6.99
3	2	5	2	0.74
3	2	6	1	0.37
3	3	2	1	0.37
3	3	4	2	0.74
3	3	6	1	0.37
3	4	4	2	0.74
3	4	5	2	0.74
3	5	5	1	0.37
4	2	2	1	0.37
4	2	3	3	1.10
4	2	4	20	7.35
4	2	5	5	1.84
4	2	6	2	0.74
4	3	5	2	0.74
4	4	4	2	0.74
4	4	5	2	0.74
4	4	6	1	0.37
4	5	5	2	0.74
5	2	5	3	1.10
5	3	5	4	1.47
5	3	6	4	1.47
5	4	5	5	1.84
5	4	6	4	1.47
5	5	5	1	0.37
5	5	6	2	0.74
6	2	4	1	0.37
6	2	6	7	2.57
6	3	6	2	0.74
6	4	6	4	1.47
6	4	7	1	0.37
6	5	6	2	0.74
6	6	6	1	0.37
6	6	7	2	0.74
6	6	8	1	0.37
6	7	7	1	0.37
7	3	6	1	0.37
7	5	7	1	0.37

Table 16.b.3. Frequency Distributions of Observed Combinations of the Three Essay Dimension Scores for Prompt 3

Essay Reading	Essay Analysis	Essay Writing	Freq	Percent
0	0	0	2	11.11
2	2	2	5	27.78
2	2	3	1	5.56
3	2	2	1	5.56
3	2	3	1	5.56
3	2	4	1	5.56
4	2	3	2	11.11
4	2	4	1	5.56
5	2	4	1	5.56
5	4	5	1	5.56
6	2	6	1	5.56
7	3	7	1	5.56

Table 17.a. Frequency Distributions of Observed Combinations of the Three Essay Dimension Scores by Rater

			Rater Set 1		Rater Set 2	
Essay Reading	Essay Analysis	Essay Reading	Freq	Percent	Freq	Percent
0	0	0	47	11.35	47	11.35
1	1	1	125	30.19	116	28.02
1	1	2	61	14.73	65	15.70
1	1	3	3	0.72	5	1.21
1	2	1	1	0.24	2	0.48
1	2	2	15	3.62	10	2.42
1	2	3	0	0.00	3	0.72
1	3	3	1	0.24	0	0.00
2	1	1	13	3.14	13	3.14
2	1	2	58	14.01	66	15.94
2	1	3	9	2.17	13	3.14
2	2	2	18	4.35	18	4.35
2	2	3	11	2.66	8	1.93
2	3	3	2	0.48	3	0.72
3	1	2	1	0.24	5	1.21
3	1	3	15	3.62	13	3.14
3	1	4	0	0.00	1	0.24
3	2	2	2	0.48	0	0.00
3	2	3	16	3.86	15	3.62
3	2	4	1	0.24	0	0.00
3	3	3	9	2.17	6	1.45
3	3	4	2	0.48	4	0.97
4	2	3	3	0.72	0	0.00
4	3	4	0	0.00	1	0.24
4	4	4	1	0.24	0	0.00
Total			*414*	*100.00*	*414*	*100.00*

Table 17.b.1. Frequency Distributions of Observed Combinations of the Three Essay Dimension Scores by Rater for Prompt 1

			Rater Set 1		Rater Set 2	
Essay Reading	Essay Analysis	Essay Reading	Freq	Percent	Freq	Percent
0	0	0	22	18.03	22	18.03
1	1	1	34	27.87	31	25.41
1	1	2	13	10.66	16	13.11
1	1	3	1	0.82	1	0.82
1	2	1	1	0.82	1	0.82
1	2	2	7	5.74	8	6.56
1	2	3	0	0.00	1	0.82
1	3	3	1	0.82	0	0.00
2	1	1	6	4.92	4	3.28
2	1	2	14	11.48	17	13.93
2	1	3	3	2.46	4	3.28
2	2	2	9	7.38	7	5.74
2	2	3	3	2.46	2	1.64
2	3	3	1	0.82	0	0.00
3	1	2	0	0.00	2	1.64
3	1	3	2	1.64	1	0.82
3	2	2	1	0.82	0	0.00
3	2	3	2	1.64	2	1.64
3	3	3	1	0.82	1	0.82
3	3	4	0	0.00	2	1.64
4	2	3	1	0.82	0	0.00

Table 17.b.2. Frequency Distributions of Observed Combinations of the Three Essay Dimension Scores by Rater for Prompt 2

			Rater Set 1		Rater Set 2	
Essay Reading	Essay Analysis	Essay Reading	Freq	Percent	Freq	Percent
0	0	0	23	8.46	23	8.46
1	1	1	84	30.88	77	28.31
1	1	2	46	16.91	49	18.01
1	1	3	2	0.74	4	1.47
1	2	1	0	0.00	1	0.37
1	2	2	8	2.94	2	0.74
1	2	3	0	0.00	2	0.74
2	1	1	6	2.21	7	2.57
2	1	2	40	14.71	46	16.91
2	1	3	6	2.21	9	3.31
2	2	2	9	3.31	10	3.68
2	2	3	8	2.94	6	2.21
2	3	3	1	0.37	3	1.10
3	1	2	1	0.37	1	0.37
3	1	3	13	4.78	11	4.04
3	2	3	13	4.78	13	4.78
3	2	4	1	0.37	0	0.00
3	3	3	7	2.57	5	1.84
3	3	4	2	0.74	2	0.74
4	2	3	1	0.37	0	0.00
4	3	4	0	0.00	1	0.37
4	4	4	1	0.37	0	0.00

Table 17.b.3. Frequency Distributions of Observed Combinations of the Three Essay Dimension Scores by Rater for Prompt 3

			Rater Set 1		Rater Set 2	
Essay Reading	Essay Analysis	Essay Reading	Freq	Percent	Freq	Percent
0	0	0	2	11.11	2	11.11
1	1	1	6	33.33	7	38.89
1	1	2	2	11.11	0	0.00
2	1	1	1	5.56	2	11.11
2	1	2	4	22.22	3	16.67
2	2	2	0	0.00	1	5.56
3	1	2	0	0.00	1	5.56
3	1	3	0	0.00	1	5.56
3	1	4	0	0.00	1	5.56
3	2	3	1	5.56	0	0.00
3	3	3	1	5.56	0	0.00
4	2	3	1	5.56	0	0.00

Table 18. Correlations of the Three Essay Dimension Scores

Score	Essay Reading	Essay Analysis	Essay Writing
N	367	367	367
Dimension Score			
Essay Reading	1		
Essay Analysis	0.54	1	
Essay Writing	0.78	0.61	1
Rater Set 1			
Essay Reading	1		
Essay Analysis	0.51	1	
Essay Writing	0.73	0.57	1
Rater Set 2			
Essay Reading	1		
Essay Analysis	0.46	1	
Essay Writing	0.69	0.53	1

Note: Scores of zero were excluded from the computation of correlations.

Table 19. Correlations between the Reading Test Score, Writing & Language Test Score, the ERW Section Score, and the Dimension Scores on Essay

Score	Essay Reading	Essay Analysis	Essay Writing
Reading Test Score	0.47	0.55	0.58
Writing Test Score	0.45	0.54	0.54
ERW Section Score	0.49	0.58	0.60
N	367	367	367

Note: Scores of zero were excluded from the computation of correlations.

Table 20.a. Cross-tabulated Score Distributions between the Two Raters for Essay Reading Score

	Rater Set 2					
Rater Set 1	0	1	2	3	4	Total
1	0 0.00	164 44.69	39 10.63	3 0.82	0 0.00	206 56.13
2	0 0.00	33 8.99	66 17.98	12 3.27	0 0.00	111 30.25
3	0 0.00	4 1.09	15 4.09	26 7.08	1 0.27	46 12.53
4	0 0.00	0 0.00	1 0.27	3 0.82	0 0.00	4 1.09
Total	0 0.00	201 54.77	121 32.97	44 11.99	1 0.27	367 100.00

Table 20.b. Cross-tabulated Score Distributions between the Two Raters for Essay Analysis Score

			Rater Set 2			
Rater Set 1	**0**	**1**	**2**	**3**	**4**	**Total**
1	0	271	13	1	0	285
	0.00	73.84	3.54	0.27	0.00	77.66
2	0	23	37	7	0	67
	0.00	6.27	10.08	1.91	0.00	18.26
3	0	3	6	5	0	14
	0.00	0.82	1.63	1.36	0.00	3.81
4	0	0	0	1	0	1
	0.00	0.00	0.00	0.27	0.00	0.27
Total	0	297	56	14	0	367
	0.00	80.93	15.26	3.81	0.00	100.00

Table 20.c. Cross-tabulated Score Distributions between the Two Raters for Essay Writing Score

			Rater Set 2			
Rater Set 1	0	1	2	3	4	Total
1	0	104	33	2	0	139
	0.00	28.34	8.99	0.54	0.00	37.87
2	0	26	110	19	0	155
	0.00	7.08	29.97	5.18	0.00	42.23
3	0	1	21	42	5	69
	0.00	0.27	5.72	11.44	1.36	18.80
4	0	0	0	3	1	4
	0.00	0.00	0.00	0.82	0.27	1.09
Total	0	131	164	66	6	367
	0.00	35.69	44.69	17.98	1.63	100.00

Table 21. Interrater Agreement between the Two Raters for Each Dimension

Agreement	Essay Reading	Essay Analysis	Essay Writing
Percent Agreement	69.75	85.29	70.03
Percent Adjacent	28.07	13.62	29.16
Percent More than Adjacent	2.18	1.09	0.82
N	367	367	367

Note: Scores of zero were excluded from the computation of interrater agreement.

Table 22. Interrater Reliability (Pearson Correlations) between the Two Rater Scores for Each Dimension

Score	Pearson Correlation	Standard Error of Measurement
Essay Reading	0.65	0.43
Essay Analysis	0.67	0.30
Essay Writing	0.72	0.40
N	367	367

Note: Scores of zero were excluded from the computation of interrater agreement.

Table 23. Interrater Consistency (Kappa) between the Two Rater Scores for Each Dimension

Score	Kappa Statistic	Value	ASE[1]	95% Confidence Limits	
Essay Reading	Simple	0.477	0.040	0.398	0.555
	Weighted	0.558	0.036	0.486	0.629
Essay Analysis	Simple	0.570	0.047	0.477	0.663
	Weighted	0.612	0.044	0.525	0.699
Essay Writing	Simple	0.533	0.037	0.461	0.606
	Weighted	0.620	0.032	0.558	0.681

[1] ASE represents asymptotic standard error.

Note: Scores of zero were excluded from the computation of correlations.

Table 24.a. Essay Dimension Score Mean, Standard Deviation, and Standardized Difference Between Gender Groups

		Male			Female		
Score	N	Mean	S.D.	N	Mean	S.D.	Std. Diff.
Essay Reading	223	2.93	1.26	143	3.51	1.40	0.44
Essay Analysis		2.40	0.85		2.64	1.07	0.25
Essay Writing		3.44	1.35		4.06	1.45	0.44

Note: Scores of zero were excluded from the analysis.

Appendix A: Target Specifications for the SAT Suite of Assessments

The target statistical specifications for the SAT Suite of Assessments describe the desired distribution or range of values on the assessment in terms of item difficulty, item discrimination, and overall reliability. Tables A1 - A3 outline exactly how many items are included at each difficulty level (i.e., easy, medium, hard).   The bounds for item difficulty levels are based on historical data. The current difficulty classifications based on p-values are used in combination with the target statistical specifications to identify the number of items per difficulty classification for each score tier.

Table A1. Target Number of Items per Difficulty Classification by Reading and Writing and Language Test Scores and Subscores

Score and difficulty level	Number of Items
Reading	
Hard (.03 ≤ p ≤ .45)	19
Medium (.46 ≤ p ≤ .81)	18
Easy (p ≥ .82)	15
Writing and Language	
Hard (.03 ≤ p ≤ .45)	9
Medium (.46 ≤ p ≤ .81)	16
Easy (p ≥ .82)	19
Expression of Ideas	
Hard (.03 ≤ p ≤ .45)	5
Medium (.46 ≤ p ≤ .81)	9
Easy (p ≥ .82)	10
Standard English Conventions	
Hard (.03 ≤ p ≤ .45)	4
Medium (.46 ≤ p ≤ .81)	7
Easy (p ≥ .82)	9
Words in Context	
Hard (.03 ≤ p ≤ .45)	3 R; 3 W/L
Medium (.46 ≤ p ≤ .81)	4 R; 2 W/L
Easy (p ≥ .82)	3 R; 3 W/L
Command of Evidence	
Hard (.03 ≤ p ≤ .45)	3 R; 3 W/L
Medium (.46 ≤ p ≤ .81)	4 R; 2 W/L
Easy (p ≥ .82)	3 R; 3 W/L

Table A2. Target Number of Items per Difficulty Classification by Math Test Score, Cross-Test Scores, and Subscores

Score and difficulty level	MC	SPR
Math		
Hard (.03 ≤ p ≤ .45)	19	6
Medium (.46 ≤ p ≤ .81)	15	4
Easy (p ≥ .82)	11	1
Any	0	2
Analysis in History/Social Studies		
Hard (.03 ≤ p ≤ .45)	8 R; 2 W/L; 2 M	2
Medium (.46 ≤ p ≤ .81)	7 R; 2 W/L; 2 M	1
Easy (p ≥ .82)	6 R; 2 W/L; 1 M	0
Analysis in Science		
Hard (.03 ≤ p ≤ .45)	8 R; 2 W/L; 2 M	2
Medium (.46 ≤ p ≤ .81)	7 R; 2 W/L; 2 M	1
Easy (p ≥ .82)	6 R; 2 W/L; 1 M	0
Heart of Algebra		
Hard (.03 ≤ p ≤ .45)	5	2
Medium (.46 ≤ p ≤ .81)	6	2
Easy (p ≥ .82)	4	0
Problem Solving and Data Analysis		
Hard (.03 ≤ p ≤ .45)	6	1
Medium (.46 ≤ p ≤ .81)	2	1
Easy (p ≥ .82)	5	0
Any	0	2
Passport to Advanced Mathematics		
Hard (.03 ≤ p ≤ .45)	7	1
Medium (.46 ≤ p ≤ .81)	6	1
Easy (p ≥ .82)	1	0

Table A3. Target Average Item Difficulty Estimates and Standard Deviations

Score	n	Mean	S.D.
Reading	52	0.579	0.285
Writing and Language	44	0.684	0.263
Math	58	0.520	0.279
Analysis in History/Social studies	35	0.564	0.273
Analysis in Science	35	0.564	0.273
Command of Evidence	18	0.592	0.303
Words in Context	18	0.592	0.303
Expression of Ideas	24	0.678	0.265
Standard English Conventions	20	0.691	0.261
Heart of Algebra	19	0.557	0.270
Problem Solving and Data Analysis	17	0.555	0.308
Passport to Advanced Mathematics	16	0.439	0.252

Table A4. Target Average Item Discrimination Bounds

Score	Lower	Upper
Reading	0.340	0.403
Writing and Language	0.475	0.538
Math	0.410	0.473
Analysis in History/Social studies	0.407	0.470
Analysis in Science	0.407	0.470
Command of Evidence	0.398	0.461
Words in Context	0.398	0.461
Expression of Ideas	0.490	0.551
Standard English Conventions	0.497	0.556
Heart of Algebra	0.444	0.501
Problem Solving and Data Analysis	0.458	0.512
Passport to Advanced Mathematics	0.454	0.509

Table A5. Target Reliability Bounds

Score	Lower	Upper
Reading	0.850	0.899
Writing and Language	0.920	0.943
Math	0.910	0.937
Analysis in History/Social studies	0.844	0.891
Analysis in Science	0.844	0.891
Command of Evidence	0.708	0.797
Words in Context	0.708	0.797
Expression of Ideas	0.863	0.900
Standard English Conventions	0.839	0.882
Heart of Algebra	0.774	0.835
Problem Solving and Data Analysis	0.730	0.800
Passport to Advanced Mathematics	0.743	0.809

Appendix B: Test Analysis Formulas

B1. Pearson Product Moment Correlation Coefficient

$$\rho_{XY} = \frac{\sum Z_X Z_Y}{N}$$

where Z_X and Z_Y represent z-scores of observed scores X and Y , respectively and N represents the number of test takers (Crocker & Algina, 1986)

B2. Disattenuated Correlations/True Score Correlations

$$\rho_{\mathrm{T}} = \frac{\rho_{XY}}{\sqrt{SA_X SA_Y}}$$

where ρ_{XY} is the correlation between observed scores X and Y, and SA_X and SA_Y represent the stratified alpha reliability of score X and Y, respectively (Schumacker & Muchinsky, 1996).

B3. Scale-score CSEM and Reliability Estimates

The reliabilities for scale scores were estimated from the average CSEM using the following equation:

$$Reliability_{SC} = 1 - \frac{MS(CSEM)_{SC}}{SD^2_{SC}},$$ where

SD^2_{SC} is the variance of scale score. The mean squared *CSEM, MS(CSEM)* was obtained as the weighted average of the squared *CSEMs* for the scales directly established. Thus the *MS(CSEM)* can be written as

$$\mathrm{MS(CSEM)}_{\mathrm{SC}} = \int \mathrm{CSEM}^2_{\mathrm{SC}(\tau)} \mathrm{Prob}(\tau)\,\mathrm{d}\tau$$, where

$CSEM^2_{SC(\tau)}$ is the squared scale score *CSEM* at τ, and the average of these is obtained over the probability distribution of τ, *Prob*(τ).

For the scores that were mathematically derived including Math Test, ERW, and Total scores, the following equations were used to compute the root mean squared *CSEM, RMS(CSEM)*:

$$\mathrm{RMS(CSEM)}_{\mathrm{MTS}} = \sqrt{\frac{MS(CSEM)_{MSS}}{20^2}}$$

$$\mathrm{RMS(CSEM)}_{\mathrm{ERW}} = \sqrt{MS(CSEM)_R \cdot 10^2 + MS(CSEM)_{WL} \cdot 10^2}$$

$$\mathrm{RMS(CSEM)}_{\mathrm{Total}} = \sqrt{MS(CSEM)_{ERW} + MS(CSEM)_{MSS}}.$$

B4. Standard Error of the Difference

The formula for computing the Standard Error of the Difference (SED) is:

$$\mathrm{SED} = \sqrt{2 * SEM^2}$$

where it is assumed that scores of two students would be independent with equal SEMs across testing times, so that the variance of the score difference could be estimated by doubling the squared SEM.

When comparing scores between students for the same measure (Reading, Writing, Math), the standard error of the difference (SED) can be used to assess how much scores must differ in order to reflect true differences in ability. If two scores differ by at least SED times 1.65, it is unlikely that the two scores indicate that the two candidates are equal in ability since this level difference would occur 10 percent of the time or less. For example, when the SED is 40 points, you can be reasonably confident that if the score difference between two test-takers is greater than 66 points (40 x 1.65), the two test-takers are not likely to be equal in true ability.

B5. Mantel-Haenszel D-DIF Statistic

Based on the formulas from Dorans and Holland (1993), the Mantel-Haenszel D-DIF (MH D-DIF) statistics is calculated for subgroups of gender and ethnicity/race with the following formula:

$$MH\ \ D-DIF=-2.35\ln[\alpha_{MH}],$$

where α_{MH} is an estimate of the odds ratio. “Positive values of MH D-DIF favor the focal group, whereas, negative values favor the reference group” (Dorans & Holland, 1993, p 41). The odds ratio is calculated as

$$\alpha_{MH}=\frac{\sum_m R_{rm}\frac{W_{fm}}{N_{tm}}}{\sum_m R_{fm}\frac{W_{rm}}{N_{tm}}}$$

where R_{rm} is the number correct in reference group at ability level m, W_{fm} is the number incorrect in the focal group at ability level m, N_{tm} is the number in total group at ability level m, R_{fm} is the number correct in the focal group at ability level m, and W_{rm} is the number incorrect in the reference group at ability level m. At the test development stage, the minimum sample size requirement for the focal group is 100 when calculating the statistics.

B6. Standardized Mean Difference

The formula for computing a standardized mean difference is:

$$d=\frac{\overline{X}_f-\overline{X}_r}{SD_T}$$

where $\bar{X}_f$ and $\bar{X}_r$ represent mean scores for the focal group and reference group (white or male), respectively, and SD_T represents the total group (pooled) standard deviation (Cohen, 1988):

$$SD_T=\sqrt{\frac{(n_f-1)SD_f^2+(n_r-1)SD_r^2}{n_f+n_r-2}}$$

where n_f and n_r represent sample sizes for the focal group and reference group, respectively, and SD_f^2 and SD_r^2 represent standard deviations for the focal group and reference group, respectively (Cohen, 1988).

B7. False Positive Rate

The formula for computing the false positive rate is:

$$R_{fp}=\int_0^{\tau_0}\Pr(\mathrm{X}\geq \mathrm{x}_0\,|\,\tau)g(\tau)d\tau$$

where τ_0 is the true score, x_0 is the raw score cut point, X is the raw score obtained by a randomly selected test-taker, $g(\tau)$ is the true score density, which is obtained using the four-parameter beta-binomial model with effective test length (Brennan, 2004; Livingston & Lewis, 1995 ; Hanson & Brennan, 1990).

B8. False Negative Rate

The formula for computing the false negative rate is:

$$R_{fn} = \int_{\tau_0}^{1} \Pr(X \le x_0 - 1 \mid \tau) g(\tau) d\tau$$

where τ_0 is the true score, x_0 is the raw score cut point, X is the raw score obtained by a randomly selected test-taker, $g(\tau)$ is the true score density, which is obtained using the four-parameter beta-binomial model with effective test length (Brennan, 2004; Livingston & Lewis, 1995 ; Hanson & Brennan, 1990).

B9. Probability of Correct Classification

The formula for computing the probability of correct classification is:

$$P = 1 - R_{fp} - R_{fn}$$

where R_{fp} is the false positive rate and R_{fn} is the false negative rate.

B10. Effective Test Length

The formula for effective test length is:

$$\tilde{n} = \frac{(\mu_x - X_{min})(X_{max} - \mu_x) - r\sigma_x^2}{\sigma_x^2(1 - r)}$$

where X_{min} is the lowest score for raw score X, X_{max} is the highest score, μ_x is the mean, σ_x^2 is the variance, and r is the reliability (Brennan, 2004; Livingston & Lewis, 1995).

B11. Proportion of Consistent Decisions

The formula for computing the proportion of consistent decisions is:

$$p = \Pr(X_1 \le x_0 - 1, X_2 \le x_0 - 1) + \Pr(X_1 \ge x_0, \mathrm{X}_2 \ge x_0)$$

where X_1 and X_2 are raw score random variables for two independent administrations and x_0 is the raw score cut point (Brennan, 2004; Livingston & Lewis, 1995; Hanson & Brennan, 1990).

B12. Proportion of Consistent Decisions by Chance

The formula for computing the proportion of consistent decisions by chance is:

$$p_c = \Pr(X_1 \le x_0 - 1)\Pr(X_2 \le x_0 - 1) + \Pr(X_1 \ge x_0)\Pr(\mathrm{X}_2 \ge x_0)$$

where X_1 and X_2 are raw score random variables for two independent administrations and x_0 is the raw score cut point (Brennan, 2004; Livingston & Lewis, 1995; Hanson & Brennan, 1990).

B13. Kappa Statistic

The formula for computing the kappa statistic is:

$$k = \frac{p - p_c}{1 - p_c}$$

where p is the proportion of consistent decisions and p_c is the proportion of consistent decisions by chance (Brennan, 2004; Livingston & Lewis, 1995; Hanson & Brennan, 1990).

B14. Probability of Misclassification

The formula for computing the probability of misclassification is:

$$p_m = 1 - p$$

where p is the proportion of consistent decisions.

B15. Single-Rater Reliability Coefficient

The single-rater reliability coefficient $\rho_{RR'}$ for a given dimension is estimated by the Pearson correlation between the first and second rater scores.

B16. Single-Rater Variance

The single-rater variance σ_R^2 for a dimension score or for the sum of dimension scores can be computed on either the first or second rater scores. Because both rater scores are generated from the same pool of raters, the two estimates are equivalent. In these analyses, the single-rater variance is estimated using the arithmetic average of the variances of the first and second rater scores:

$$\sigma_R^2 = \frac{1}{2}\left(\sigma_{R1}^2 + \sigma_{R2}^2\right).$$

B17. Single-Rater Standard Error of Measurement

The variance error of measurement for a single rater SEM_R is given by:

$$SEM_R = \sqrt{\left(1 - \rho_{RR'}\right)\sigma_R^2}.$$

B18. Percentage of Agreement

The percentage of agreement (in percentage) is computed as

$$p = \sum p_{ij} \text{ for all } i = j.$$

B19. Simple Kappa Coefficient

The simple kappa coefficient is given by

$$\hat{\kappa} = (p_0 - p_e)/(1 - p_e)$$

where p_0 is the observed probability of agreement and is computed as $p_0 = \Sigma p_{ij}$ for all $i=j$. p_e is the expected probability of agreement and is computed as $p_e = \Sigma p_{i.} p_{.j}$ for all $i=j$.

The asymptotic variance of the simple kappa coefficient is computed as

$$\mathrm{var}\left(\hat{k}\right) = \frac{\sum_{i=j}\left(p_{ij}\left(1-\left(p_{i.}+p_{.j}\right)\left(1-\hat{k}\right)\right)^2\right) + \left(1-\hat{k}\right)^2 \sum\sum_{i\neq j}\left(p_{ij}\left(p_{i.}+p_{.j}\right)^2\right) + \left(\hat{k}-p_e\left(1-\hat{k}\right)\right)^2}{\left(1-p_e\right)^2 * n}$$

The asymptotic standard error (ASE) is the square root of the asymptotic variance. The confidence limits are computed as

$$\kappa \pm (1.96 * \sqrt{var(\hat{\kappa})})$$

B20. Weighted Kappa Coefficient

The weighted Kappa coefficient is a generalization of the simple Kappa coefficient that uses the weights to quantify the relative difference between categories. It is computed as

$$\hat{k}_w = \left(p_{0(w)} - p_{e(w)}\right) / \left(1 - p_{e(w)}\right)$$

where p_0 is the observed probability agreement and is computed as $p_{0(w)} = \Sigma_i \Sigma_j w_{ij} p_{ij}$ and $p_{e(w)}$ is the expected probability agreement and is computed as $p_{e(w)} = \Sigma_i \Sigma_j w_{ij} p_{i.} p_{.j}$. The weights w_{ij} are constructed so that $w_{ij}=1$ for all $i=j$,$0=w_{ij}<1$ for all $i=j$, and $w_{ij}=w_{ji}$. The asymptotic variance of the weighted kappa coefficient is computed as

$$\mathrm{var}\left(\hat{k}_w\right) = \frac{\sum_i \sum_j p_{ij}\left(w_{ij} - \left(\sum_j p_{.j} w_{ij} + \sum_i p_{i.} w_{ij}\right)\left(1-\hat{k}_w\right)\right)^2 - \left(\hat{k}_w - p_{e(w)}\left(1-\hat{k}_w\right)\right)^2}{\left(1-p_{e(w)}\right)^2 * n}$$

The asymptotic standard error (ASE) is the square root of the asymptotic variance. The confidence limits are computed as

$$\hat{k}_w \pm \left(1.96 * \sqrt{\mathrm{var}\left(\hat{k}_w\right)}\right).$$

About the College Board

The College Board is a mission-driven, not-for-profit organization that connects students to college success and opportunity. Founded in 1900, the College Board was created to expand access to higher education. Today, the membership association is made up of over 6,000 of the world’s leading educational institutions and is dedicated to promoting excellence and equity in education. Each year, the College Board helps more than seven million students prepare for a successful transition to college through programs and services in college readiness and college success — including the SAT® and the Advanced Placement Program®. The organization also serves the education community through research and advocacy on behalf of students, educators, and schools. For further information, visit www.collegeboard.org.