SAT Suite of Assessments Essay Report

Delaware

SAT School Day Administration

April 2017

2017

Revised September 12, 2017

Executive Summary

This report summarizes the performance of 9,516 students who took the SAT Essay in Delaware during the April 2017 administration. This report is a summary of descriptive statistics and frequency distributions, and correlations of essay dimension scores, and inter-rater consistency.

Table of Contents

SAT Suite of Assessments 5
SAT Essay 5
Characteristics of the Administrations of the SAT 6
Test forms included 6
Description of the analysis sample 6
Essay Analyses 6
Moments and Score Distributions 6
Intercorrelations 7
Reliability and standard error of measurement 7
Standardized differences between groups 8
Tables 9
Table 1. Essay Dimension Score Descriptions and Score Ranges 9
Table 2. Descriptive Statistics for Essay Dimension Scores 10
Table 3. Frequency Distributions of the Three Essay Dimension Scores 11
Table 4. Frequency Distributions of the Three Essay Dimension Scores by Rater 12
Table 5. Frequency Distributions of All Possible Combinations of the Three Essay Dimension Scores 13
Table 6. Frequency Distributions of All Possible Combinations of the Three Essay Dimension Scores by Rater 15
Table 7. Correlations of the Three Essay Dimension Scores 16
Table 8. Correlations Between the Reading Test score, the Writing & Language Test score, the Evidence-Based Reading and Writing (ERW) Section Score, and the Dimension Scores on Essay 16
Table 9a. Cross-tabulated Score Distributions Between the Two Raters for Essay Reading Score 17
Table 9b. Cross-tabulated Score Distributions Between the Two Raters for Essay Analysis Score 17
Table 9c. Cross-tabulated Score Distributions Between the Two Raters for Essay Writing Score 18
Table 10. Inter-rater Agreement Between the Two Raters for Each Dimension 18
Table 11. Inter-rater Reliability (Pearson Correlations) Between the Two Rater Scores for Each Dimension 18
Table 12. Inter-rater Consistency (Kappa) Between the Two Raters for Each Dimension 19
Table 13a. Essay Dimension Score Mean, Standard Deviation, and Standardized Difference Between Gender Groups 20

Table 13b. Essay Dimension Score Mean, Standard Deviation, and Standardized Difference Between Racial/Ethnic Groups 20

Bibliography/References 21

About the College Board 22

Appendix A: Correlation Indices and Reliability Estimates 23

A1. Pearson product-moment correlation coefficient 23

A2. Percentage of agreement 23

A3. Single-rater reliability coefficient 23

A4. Single-rater variance 23

A5. Single-rater Standard error of measurement 23

A6. Simple Kappa Coefficient 23

A7. Weighted Kappa Coefficient 24

SAT Suite of Assessments

The SAT Suite of Assessments (SAT, PSAT/NMSQT®, PSAT™ 10, and PSAT™ 8/9) is designed to measure student readiness for college and postsecondary education. Each assessment comprises two sections (the Evidence-Based Reading and Writing section and the Math section), three tests (the Reading Test, the Writing and Language Test, and the 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 primarily concerns the Essay, although analyses make use of the Evidence-Based Reading and Writing (ERW) and Math section scores.

SAT Essay

Test takers opting to take the SAT Essay receive an additional 50 minutes at the end of the SAT testing session to respond to the essay. A prompt appears with every essay text. The following prompt, or a nearly identical one, is used every time:

> 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 Reading, Analysis, and Writing. 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 score. If one rater gives an essay a score of 0, or 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 score. If an essay receives a score of 0 on one dimension, then it is scored 0 on all three dimensions.

This report contains information about the three essay dimension scores, specifically Essay Reading, Essay Analysis, and Essay Writing. Table 1 lists the descriptions of the three Essay scores examined in this report.

Characteristics of the Administrations of the SAT

Test forms included

This report summarizes the Essays associated with the SAT test form administered in April 2017. Test takers were given one SAT form and four prompts were administered as part of the Essay test. This report summarizes the data at the overall level (i.e., aggregating all prompts).

Description of the analysis sample

Before completing the analyses contained in this report, the sample available at 38 days after the test administration was cleaned to exclude any students not in grade 11. See Table 3 in the *SAT Suite of Assessments Administration Report, Delaware, SAT School Day Administration, April 2017,* for the frequency of test takers in the item analysis sample for this administration by first language and gender.

A score of 0 is assigned to unscorable essays, so a score of 0 is not included in all the analyses in this report (eg., moments, correlation, and inter-rater reliability analyses), except the frequency distribution of scores (including all three dimensions).

Essay Analyses

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 Tables 2-6 for these results. Note that in several tables 'Rater Sets' are mentioned. These

rater sets are an aggregation of scores across all raters who provided the first (Rater Set 1), second (Rater Set 2), or third (Rater Set 3) score.

Intercorrelations

The Pearson product moment correlation coefficient provides an evaluation of the pairwise linear relationship between two essay or rater scores. The formula for calculating the Pearson correlations is in Appendix A1. See Tables 7 and 8 for these results.

Reliability and standard error of measurement

Reliability refers to the consistency with which an instrument measures some attribute of a person or object. In the context of this report, 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, and a number of other factors. If we consider these fluctuations in scores to be error, then reliability is an index of the proportion of the measurement that is not error. Reliability estimates range from 0 to 1, with reliability estimates near 1 indicating consistent measurement with very little error. On the other hand, reliability estimates near zero would indicate fairly random estimates of the attribute.

If the reliability estimate indicates consistency of measurement, the standard error of measurement (SEM) can be considered a measure of inconsistency in test takers' observed scores. Unlike reliability estimates, which are standardized to range from 0 to 1, the SEM is presented in the unit of measurement of the observed score. The SEM represents the average margin of error around the observed score. The SEM is inversely related to reliability (see Appendix A5 for the formulas).

Inter-rater (single rater) reliability

Inter-rater reliability is the reliability of a single rater scoring the Essay. This reliability 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? Just as with Essay reliability, although the reliability coefficient corresponds to a single rater, the estimation of inter-rater reliability requires that at least two raters score the same Essay for the same examinee. Then the reliability of the raters can be estimated.

Percentages of agreement

Percentage of agreement is an index of inter-rater agreement. It can be expressed as the number agreeing divided by the total observations (see Appendix A2 for the formula). For ordinal and interval data, percentages of close-but-not-exact agreement (e.g., percentage of adjacent – where raters are off by 1) can also be computed and, along with percentage of exact agreement, used as measures of inter-rater agreement. The percentage of agreement does not

take into account agreements due to chance. Therefore, it overestimates the level of agreement (Hallgren, 2012). See Tables 9a-10 for these results.

Correlation coefficient

The correlation coefficient between the scores given by two raters on the same essay dimension scores is a measure of the inter-rater consistency. The formula for computing the Pearson correlation coefficient is in Appendix A1. See Table 11 for these results.

Simple kappa statistic

Cohen's kappa coefficient (simple kappa statistic; Cohen, 1960) is a statistic that measures the inter-rater 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 ratings of the two raters, and the agreement expected by chance is defined by the marginal frequencies of the ratings given by each rater. The formula for calculating Cohen's kappa coefficient is given in Appendix A6. Possible values for Cohen's kappa coefficients range from -1 to 1, with 1 indicating complete agreement, 0 indicating complete random agreement, and -1 indicating complete disagreement. See Table 12 for these results.

Weighted kappa statistic

Weighted kappa coefficient (Cohen, 1968) is an alternative statistic that measures the inter-rater agreement between two raters while correcting the possibility of agreement by chance and penalizing the disagreements. It can be applied to ordinal ratings and 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 A7. See Table 12 for these results.

Standardized differences between groups

The test taker performance for subgroups, with samples greater than or equal to 200, is described using the mean and standard deviation for all Essay scores and the standardized mean differences between the focal and reference groups. Cohen (1988) suggests standardized mean differences equal to 0.20 are small, 0.50 are medium, and 0.80 are large. See Tables 13a and 13b for Essay score subgroup moments and differences.

Tables

Table 1. Essay Dimension Score Descriptions and Score Ranges

Score	What the score reflects	Scale*
Dimension Scores		
Reading	• Comprehension of source text • Understanding of central ideas • Correctness of interpretation of text • Use of textual evidence	2 - 8
Analysis	• Insightfulness of analysis of source text • Evaluation of the author's use of evidence, reasoning, and persuasion • Relevance and sufficiency of support for claims • Consistency of focus on relevant features of the text	2 - 8
Writing	• Cohesiveness; command of the language • Effectiveness of progression of ideas throughout the essay • Variety of sentence structures; preciseness of word choice; style and tone • Command of conventions of standard written English; freedom from errors	2 - 8

* Each dimension score is the sum of two rater scores, each ranging from 1 to 4.

Table 2. Descriptive Statistics for Essay Dimension Scores

Score	Mean	SD	Skewness	Kurtosis
Reading				
Rating 1	2.38	0.69	-0.24	-0.38
Rating 2	2.38	0.70	-0.31	-0.41
Dimension Score	4.76	1.22	-0.35	-0.33
Analysis				
Rating 1	1.78	0.75	0.50	-0.73
Rating 2	1.78	0.75	0.51	-0.67
Dimension Score	3.54	1.35	0.43	-0.83
Writing				
Rating 1	2.30	0.72	-0.23	-0.61
Rating 2	2.29	0.72	-0.24	-0.60
Dimension Score	4.59	1.30	-0.29	-0.61
N	9,272	9,272	9,272	9,272

Note: Zero dimension scores were excluded from the computation of all four moments. For each dimension, the two rater scores are added to form the 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 3. Frequency Distributions of the Three Essay Dimension Scores

	Essay Reading		Essay Analysis		Essay Writing	
Score	Frequency	Percent	Frequency	Percent	Frequency	Percent
0	244	2.56	244	2.56	244	2.56
2	487	5.12	2953	31.03	760	7.99
3	776	8.15	1696	17.82	989	10.39
4	2583	27.14	2302	24.19	2658	27.93
5	2407	25.29	1442	15.15	2058	21.63
6	2677	28.13	777	8.17	2533	26.62
7	305	3.21	97	1.02	244	2.56
8	37	0.39	5	0.05	30	0.32
Total	9,516	100.00	9,516	100.00	9,516	100.00

Table 4. 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	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent
0	237	2.49	249	2.62	237	2.49	249	2.62	237	2.49	249	2.62
1	879	9.24	938	9.86	3,823	40.17	3,814	40.08	1,284	13.49	1,289	13.55
2	4,202	44.16	4,047	42.53	3,791	39.84	3,789	39.82	4,132	43.42	4,135	43.45
3	3,987	41.9	4,086	42.94	1,596	16.77	1,582	16.62	3,698	38.86	3,679	38.66
4	211	2.22	196	2.06	69	0.73	82	0.86	165	1.73	164	1.72
Total	9,516	100.00	9,516	100.00	9,516	100.00	9,516	100.00	9,516	100.00	9,516	100.00

Note: Rater Set is an aggregate of scores across all raters who provided the first (Rater Set 1) or second (Rater Set 2) score

Table 5. Frequency Distributions of All Possible Combinations of the Three Essay Dimension Scores

Essay Reading	Essay Analysis	Essay Writing	Frequency	Percent
0	0	0	244	2.56
2	2	2	372	3.91
2	2	3	56	0.59
2	2	4	19	0.20
2	2	5	1	0.01
2	3	2	5	0.05
2	3	3	12	0.13
2	3	4	9	0.09
2	3	5	4	0.04
2	4	3	1	0.01
2	4	4	6	0.06
2	5	5	2	0.02
3	2	2	219	2.30
3	2	3	278	2.92
3	2	4	103	1.08
3	2	5	6	0.06
3	2	6	4	0.04
3	3	2	10	0.11
3	3	3	57	0.60
3	3	4	51	0.54
3	3	5	5	0.05
3	3	6	1	0.01
3	4	3	4	0.04
3	4	4	27	0.28
3	4	5	9	0.09
3	4	6	1	0.01
3	5	6	1	0.01
4	2	2	132	1.39
4	2	3	375	3.94
4	2	4	682	7.17
4	2	5	73	0.77
4	2	6	4	0.04
4	3	2	11	0.12
4	3	3	107	1.12
4	3	4	517	5.43
4	3	5	74	0.78
4	3	6	11	0.12
4	4	2	3	0.03
4	4	3	17	0.18
4	4	4	366	3.85
4	4	5	130	1.37
4	4	6	37	0.39
4	5	4	13	0.14
4	5	5	16	0.17
4	5	6	12	0.13
4	6	4	1	0.01
4	6	5	1	0.01
4	6	6	1	0.01
5	2	2	2	0.02
5	2	3	41	0.43
5	2	4	211	2.22
5	2	5	165	1.73
5	2	6	33	0.35
5	3	2	6	0.06
5	3	3	27	0.28
5	3	4	226	2.37
5	3	5	271	2.85
5	3	6	47	0.49
5	4	3	9	0.09
5	4	4	226	2.37
5	4	5	469	4.93
5	4	6	212	2.23
5	5	3	2	0.02
5	5	4	38	0.40
5	5	5	226	2.37
5	5	6	153	1.61
5	5	7	2	0.02
5	6	4	1	0.01
5	6	5	11	0.12
5	6	6	26	0.27
5	6	7	2	0.02
5	7	7	1	0.01
6	2	3	2	0.02
6	2	4	32	0.34
6	2	5	66	0.69
6	2	6	70	0.74
6	3	3	1	0.01
6	3	4	34	0.36
6	3	5	93	0.98
6	3	6	115	1.21
6	4	4	71	0.75
6	4	5	215	2.26
6	4	6	479	5.03
6	4	7	5	0.05
6	5	4	20	0.21
6	5	5	167	1.75
6	5	6	708	7.44
6	5	7	19	0.20

Essay Reading	Essay Analysis	Essay Writing	Frequency	Percent
6	6	4	3	0.03
6	6	5	43	0.45
6	6	6	487	5.12
6	6	7	34	0.36
6	6	8	1	0.01
6	7	6	8	0.08
6	7	7	4	0.04
7	2	5	3	0.03
7	2	6	2	0.02
7	2	7	2	0.02
7	3	5	1	0.01
7	3	7	1	0.01
7	4	4	1	0.01
7	4	5	1	0.01
7	4	6	4	0.04
7	4	7	7	0.07
7	5	4	1	0.01
7	5	5	2	0.02
7	5	6	36	0.38
7	5	7	22	0.23
7	6	5	3	0.03
7	6	6	63	0.66
7	6	7	80	0.84
7	6	8	4	0.04
7	7	6	10	0.11
7	7	7	55	0.58
7	7	8	5	0.05
7	8	8	2	0.02
8	4	6	1	0.01
8	4	7	1	0.01
8	5	6	1	0.01
8	5	7	1	0.01
8	6	5	1	0.01
8	6	6	5	0.05
8	6	7	3	0.03
8	6	8	7	0.07
8	7	6	1	0.01
8	7	7	5	0.05
8	7	8	8	0.08
8	8	8	3	0.03
Total			9,516	100.00

Table 6. Frequency Distributions of All Possible Combinations of the Three Essay Dimension Scores by Rater

Essay Reading	Essay Analysis	Essay Writing	Rater Set 1		Rater Set 2	
			Frequency	Percent	Frequency	Percent
0	0	0	237	2.49	249	2.62
1	1	1	669	7.03	706	7.42
1	1	2	150	1.58	160	1.68
1	1	3	7	0.07	4	0.04
1	2	1	5	0.05	7	0.07
1	2	2	43	0.45	54	0.57
1	2	3	3	0.03	7	0.07
1	3	3	2	0.02	0	0.00
2	1	1	559	5.87	541	5.69
2	1	2	1722	18.10	1673	17.58
2	1	3	97	1.02	130	1.37
2	2	1	43	0.45	31	0.33
2	2	2	1325	13.92	1301	13.67
2	2	3	384	4.04	313	3.29
2	3	1	0	0.00	1	0.01
2	3	2	20	0.21	15	0.16
2	3	3	52	0.55	42	0.44
3	1	1	5	0.05	3	0.03
3	1	2	292	3.07	301	3.16
3	1	3	318	3.34	294	3.09
3	2	1	3	0.03	0	0.00
3	2	2	501	5.26	534	5.61
3	2	3	1473	15.48	1529	16.07
3	2	4	1	0.01	4	0.04
3	3	2	79	0.83	97	1.02
3	3	3	1268	13.32	1273	13.38
3	3	4	36	0.38	36	0.38
3	4	3	7	0.07	7	0.07
3	4	4	4	0.04	8	0.08
4	1	3	2	0.02	2	0.02
4	1	4	2	0.02	0	0.00
4	2	3	8	0.08	8	0.08
4	2	4	2	0.02	1	0.01
4	3	3	69	0.73	59	0.62
4	3	4	70	0.74	59	0.62
4	4	3	8	0.08	11	0.12
4	4	4	50	0.53	56	0.59
Total			9,516	100.00	9,516	100.00

Note: Rater Set is an aggregate of results across all raters providing the first (Rater Set 1) or second (Rater Set 2) score

Table 7. Correlations of the Three Essay Dimension Scores

Score	Essay Reading	Essay Analysis	Essay Writing
N	9,272	9,272	9,272
Reported Score			
Essay Reading	1		
Essay Analysis	0.69	1	
Essay Writing	0.82	0.73	1
Rater Set 1			
Essay Reading	1		
Essay Analysis	0.61	1	
Essay Writing	0.75	0.65	1
Rater Set 2			
Essay Reading	1		
Essay Analysis	0.62	1	
Essay Writing	0.75	0.65	1

Note: Zero scores were excluded from the computation of correlations. Rater Set is an aggregate of results across all raters providing the first (Rater Set 1) or second (Rater Set 2) score

Table 8. Correlations Between the Reading Test score, the Writing & Language Test score, the Evidence-Based Reading and Writing (ERW) Section Score, and the Dimension Scores on Essay

Score	Essay Reading	Essay Analysis	Essay Writing
Reading Scale Score	0.57	0.61	0.62
Writing Scale Score	0.57	0.62	0.62
ERW	0.6	0.65	0.65
N	9,272	9,272	9,272

Note: Zero scores were excluded from the computation of correlations.

Table 9a. Cross-tabulated Score Distributions Between the Two Raters for Essay Reading Score

	Rater Set 2					
Rater Set 1	**0**	**1**	**2**	**3**	**4**	**Total**
0	2 0.02	1 0.01	0 0.00	0 0.00	0 0.00	3 0.03
1	4 0.04	460 4.96	350 3.77	57 0.61	1 0.01	872 9.40
2	0 0.00	426 4.59	2517 27.15	1238 13.35	19 0.20	4200 45.30
3	0 0.00	51 0.55	1168 12.60	2626 28.32	141 1.52	3986 42.99
4	0 0.00	0 0.00	12 0.13	164 1.77	35 0.38	211 2.28
Total	6 0.06	938 10.12	4047 43.65	4085 44.06	196 2.11	9272 100.00

Note: Rater Set is an aggregate of results across all raters providing the first (Rater Set 1) or second (Rater Set 2) score

Table 9b. Cross-tabulated Score Distributions Between the Two Raters for Essay Analysis Score

	Rater Set 2					
Rater Set 1	**0**	**1**	**2**	**3**	**4**	**Total**
0	2 0.02	1 0.01	0 0.00	0 0.00	0 0.00	3 0.03
1	4 0.04	2875 31.01	840 9.06	92 0.99	2 0.02	3813 41.12
2	0 0.00	856 9.23	2200 23.73	714 7.70	21 0.23	3791 40.89
3	0 0.00	81 0.87	728 7.85	733 7.91	54 0.58	1596 17.21
4	0 0.00	0 0.00	21 0.23	43 0.46	5 0.05	69 0.74
Total	6 0.06	3813 41.12	3789 40.86	1582 17.06	82 0.88	9272 100.00

Note: Rater Set is an aggregate of results across all raters providing the first (Rater Set 1) or second (Rater Set 2) score

Table 9c. Cross-tabulated Score Distributions Between the Two Raters for Essay Writing Score

	Rater Set 2					
Rater Set 1	**0**	**1**	**2**	**3**	**4**	**Total**
0	2 0.02	1 0.01	0 0.00	0 0.00	0 0.00	3 0.03
1	4 0.04	743 8.01	482 5.20	47 0.51	0 0.00	1276 13.76
2	0 0.00	507 5.47	2596 28.00	1014 10.94	14 0.15	4131 44.55
3	0 0.00	38 0.41	1043 11.25	2495 26.91	121 1.31	3697 39.87
4	0 0.00	0 0.00	13 0.14	123 1.33	29 0.31	165 1.78
Total	6 0.06	1289 13.90	4134 44.59	3679 39.68	164 1.77	9272 100.00

Note: Rater Set is an aggregate of results across all raters providing the first (Rater Set 1) or second (Rater Set 2) score

Table 10. Inter-rater Agreement Between the Two Raters for Each Dimension

Agreement	Essay Reading	Essay Analysis	Essay Writing
Percent Agreement	60.83	62.72	63.25
Percent Adjacent	37.66	34.94	35.54
Percent More than Adjacent	1.51	2.34	1.21
Total	100.00	100.00	100.00

Note: Zero scores were excluded from the computation of inter-rater agreement.

Table 11. Inter-rater Reliability (Pearson Correlations) Between the Two Rater Scores for Each Dimension

Score	Pearson Correlation	Standard Error of Measurement
Essay Reading	0.54	0.47
Essay Analysis	0.61	0.47
Essay Writing	0.61	0.45
N	9,272	9,272

Note: Zero scores were excluded from the computation of correlations.

Table 12. Inter-rater Consistency (Kappa) Between the Two Raters for Each Dimension

Score	Kappa Statistic	Value	ASE[1]	95% Confidence Limits	
Essay Reading	**Simple**	0.35	0.01	0.33	0.37
	Weighted	0.43	0.01	0.42	0.45
Essay Analysis	**Simple**	0.41	0.01	0.40	0.43
	Weighted	0.50	0.01	0.49	0.52
Essay Writing	**Simple**	0.41	0.01	0.40	0.43
	Weighted	0.50	0.01	0.49	0.52

[1] ASE represents asymptotic standard error.

Note: Zero scores were excluded from the computation of correlations.

Table 13a. Essay Dimension Score Mean, Standard Deviation, and Standardized Difference Between Gender Groups

Score	Male		Female		
	Mean	S. D.	Mean	S. D.	Std. Diff.
Essay Reading	4.56	1.24	4.97	1.16	0.35
Essay Analysis	3.32	1.30	3.76	1.37	0.33
Essay Writing	4.34	1.31	4.84	1.24	0.39

Note: Zero scores were excluded from the analysis.

Table 13b. Essay Dimension Score Mean, Standard Deviation, and Standardized Difference Between Racial/Ethnic Groups

Score	White		Black			Hispanic			Asian		
	Mean	S.D.	Mean	S.D.	Std. Diff.	Mean	S.D.	Std. Diff.	Mean	S.D.	Std. Diff.
Essay Reading	5.05	1.16	-	-	-	4.67	1.15	-0.33	5.53	1.07	0.42
Essay Analysis	3.85	1.37	-	-	-	3.34	1.27	-0.38	4.51	1.39	0.48
Essay Writing	4.91	1.22	-	-	-	4.48	1.25	-0.36	5.45	1.15	0.44
Score	White		NHPI			AIAN			Two or more races		
	Mean	S.D.	Mean	S.D.	Std. Diff.	Mean	S.D.	Std. Diff.	Mean	S.D.	Std. Diff.
Essay Reading	5.05	1.16	-	-	-	-	-	-	4.91	1.17	-0.12
Essay Analysis	3.85	1.37	-	-	-	-	-	-	3.71	1.32	-0.11
Essay Writing	4.91	1.22	-	-	-	-	-	-	4.75	1.26	-0.13

Note: Zero scores were excluded from the analysis. Results are only included if the non-white group sample is equal to or greater than 200.

Bibliography/References

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

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

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement*, *fourth edition* (pp. 65-110). Westport, CT: Praeger.

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.

Appendix A: Correlation Indices and Reliability Estimates

A1. 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)

A2. Percentage of agreement

The percentage of agreement (in percentage) is computed as

$$p = \sum p_{ii}.$$

A3. 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 (result A1).

A4. 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:

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

A5. Single-rater Standard error of measurement

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

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

A6. 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 $p_0 = \sum_i p_{ii}$. p_e is the expected probability of agreement and is computed as $p_e = \sum_i p_{i.}\, p_{.i}$.

The asymptotic variance of the simple kappa coefficient is computed as

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

Where p_e is the expected probability of agreement and is computed as $p_e = \sum_i p_{i.}\, p_{.i}$.

The asymptotic standard error (ASE) is the square root of the asymptotic variance.

The confidence limits are computed as $\kappa \pm (1.96 * \sqrt{var(\hat{\kappa})})$

A7. 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

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

Where p_0 is the observed probability agreement and is computed as $p_{0(w)} = \sum_i \sum_j w_{ij} p_{ij}$. $p_{e(w)}$ is the expected probability agreement and is computed as $p_{e(w)} = \sum_i \sum_j w_{ij} p_{i.} p_{.j}$. The weights w_{ij} are constructed so that $0 \leq w_{ij} \leq I$ for all $i \neq j$ and $w_{ij} = I$ for all $i = j$, and $w_{ij} = w_{ij}$.

The asymptotic variance of the weighted kappa coefficient is computed as

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

Where $p_{e(w)}$ is the expected probability agreement and is computed as $p_{e(w)} = \sum_i \sum_j w_{ij} p_{i.} p_{.j}$.

The asymptotic standard error (ASE) is the square root of the asymptotic variance.

The confidence limits are computed as $\hat{\kappa}_w \pm (1.96 * \sqrt{var(\hat{\kappa}_w)})$