Does Quantity Equal Quality?

The Relationship Between Length of Response and Scores on the SAT Essay

Jennifer L. Kobrin, Hui Deng, and Emily J. Shaw

The College Board

Abstract

This study was designed to address two frequent criticisms of the SAT essay -- that essay length is the best predictor of scores, and that there is an advantage in using more "sophisticated" examples as opposed to personal experience. The study was based on 2,820 essays from the first three administrations of the new SAT. Each essay was coded for number of words, number of paragraphs, whether or not the response included first-person, and whether or not the response went to the second page. Analyses included descriptive statistics and group comparisons on the essay response features, correlations between essay length and scores, and hierarchical multiple regression to examine the contribution of each essay feature variable to the prediction of essay scores. The number of words in the essay explained 39% of the variance of essay scores. Whether or not the essay reached the second page explained an additional 1.5%, and whether or not the essay was written in first person explained an additional 1.1% . An examination of these features potentially affecting SAT essay scores is essential to maintain that the SAT writing section promotes valid interpretations of students' writing skills. The research described in this paper may benefit other testing programs that include essay assessments. The careful analysis of response features and the identification of potential construct-irrelevant features in essay assessments are important for evaluating the content and construct validity of writing assessments.

Introduction

The new SAT Reasoning Test, first administered in March 2005, was designed to incorporate a number of important changes. Perhaps the most noteworthy change to the test was the addition of a writing sectionthat includes a 25-minute student-composed essay in response to a prompt focusing on a particular issue. The SAT essay prompt includes a short paragraph from an authentic text and asks students to consider the issue at hand, develop a point of view, and demonstrate critical thinking by using clear and appropriate examples, reasons, and other evidence to support their positions (College Board, 2004). All essay prompts go through the same rigorous development and pre-testing process to ensure comparability and accessibility to all types of students. The prompts are pretested in a diverse sample of schools. A sample of at least 300 responses to each prompt is read by a group of experienced teachers to determine whether a particular prompt elicits responses that can be scored reliably and that provide differentiation among better and poorer writers (College Board, in press).

Two trained essay raters holistically score each essay on a scale of 0 to 6, where 6 represents an outstanding essay demonstrating clear and consistent mastery, and 1 represents an essay that is fundamentally lacking or demonstrating minimal or no mastery. A score of 0 is reserved for students who do not write an essay, essays written on a topic that was not addressed in the prompt, or severely illegible essays that have been confirmed by a number of people to be impossible to score. To date, there have been no severely illegible essays deemed impossible to score.

When SAT essays are scored holistically, a number of aspects are taken into account, including: the development of a point of view, the logical presentation of ideas, clear reasoning, sustained focus, appropriate choices of evidence, skillful coherence, effective organization,

precise use of language, and engagement with the reader (College Board, 2004). Nevertheless, critics have insisted that other factors such as handwriting (Setoodeh, 2005), essay length (Perelman, 2005), the sophistication of examples (Skutches, 2005), or the essay prompt topic (Baron, 2005) affect students' essay scores.

A popular criticism of the SAT essay is that it rewards essay length over content in the scoring process. Perelman (2005) wrote, "…the test [SAT essay] encourages wordiness. Longer essays consistently score higher." Perelman was quoted in a related article stating, "I have never found a quantifiable predictor in 25 years of grading that was anywhere near as strong as this one. If you just graded them [SAT essays] based on length without ever reading them, you'd be right over 90 percent of the time" (Winerip, 2005). Perelman's comment is based on a review of a limited number of scored essays used to train raters. It is important to note that many studies on the relationship between word count and essay score have found there to be a significant positive relationship between the two, with an average correlation in the .60s (Breland, Camp, Jones, Morris, & Rock, 1987; Breland, Danos, Kahn, Kubota, & Sudlow, 1991; Breland & Jones, 1988; Powers, Fowles, Farnum, & Ramsey, 1994).

Numerous newspaper and magazine articles have criticized the SAT essay for rewarding students for using a formulaic five-paragraph essay format (Baron, 2005; Cioffi, 2005; Hoover, 2005; Thomas, 2004) or essays with more than three paragraphs (Katzman, Lutz, & Olson, 2004). While there has been little scientific research on this topic, English teachers and professors have written about the five-paragraph essay's role in stunting students' critical thinking abilities or prohibiting the flexibility of ideas (Nunnally, 1991; Wesley, 2000). Nunnally (1991), however, does not completely reject the five-paragraph essay format because of its value in developing solid principles of composition, noting that students should not be

asking, "What *three* ideas can I use to support my thesis?" but should be asking, "What ideas can I use to support my thesis?" (Nunnally, 1991, p.71).

An examination of the construct-irrelevant essay features affecting SAT essay scores is essential to maintain that the SAT promotes valid interpretations of students' writing skills. The purpose of this study was to examine some of the features that have been the focus of criticism on the SAT. The first phase of the study, which is the focus of this paper, was designed to respond to pressing questions about essay scores that have been raised at the College Board and in the media. This first phase focused on determining the association between surface features of the essay, most notably, length of the response, with essay scores. A second phase of this research, designed to examine a wider array of essay features, is currently underway and scheduled to be completed by early 2007.

Methodology

The SAT Writing Section.

The SAT writing section includes a 25-minute and a 10-minute multiple-choice section, and a 25-minute essay. The 49 multiple-choice items are combined to produce a scaled sub-score from 20 to 80. The essay is scored by two trained readers on a scale from 0 to 6, and the scores are combined to produce a raw sub-score from 2 to 12. If the two readers disagree by more than one point, the essay is sent to an expert reader to determine the final score on the 0-6 scale and this is doubled to produce a raw sub-score. The multiple-choice and essay sub-scores are combined to form a total score for the writing section on the 200-800 scale.

Data Sources

This study was based on 2,820 essay papers from the first three administrations of the newly revised SAT. The essay papers that were studied were written in response to the prompts

used on the East and West coasts for each test administration because these had the largest number of test-takers. To ensure that the sample used in the study represented the SAT test-taker population in terms of racial/ethnic and best language subgroups, stratified random sampling was employed. The pooled data from the March, May, and June 2005 administrations were used to set percentage targets for the major racial/ethnic (White, African-American, Asian-American, and Hispanic) and best language subgroups (English only, English and another language, another language). Six stratified samples were drawn separately for prompts used on the East or West coast in the March, May and June 2005 administrations. Several subgroups were over-sampled to ensure that the mean of each subpopulation sample was within 0.5 points of the mean of the subpopulation with a 95 percent confidence interval. The minimum sample size required for each subgroup was determined, and additional test-takers were selected for the subgroups for which the random sample produced sample sizes below the target.

The essay responses were coded by temporary staff and entered into an Access database. Temporary staff were trained and given practice coding before beginning the task. The essays were coded on the total number of words, number of paragraphs, and use of first person voice. The coding form is found in the appendix to this paper. Approximately ten percent of the essay papers were coded independently a second time to assess the reliability of the coding. The agreement in the double-coding for number of words was 77%, allowing for a difference of no more than two words. Allowing for a difference of no more than 50 words, the agreement was 94%. The agreement for coding of number of paragraphs was 87%, reaching the second page, 91%, and first-person voice, 79%.[1]

[1] It was discovered that one of the temporary workers coding the essays misinterpreted the directions for coding first-person voice. Therefore, the authors of the study went back and recoded all of these essays, so the agreement of the coding for first-person voice was actually higher than 79%.

Data Analyses

The data analyses included descriptive statistics and group comparisons on the essay response features, with associated standardized mean differences (effect sizes). Correlations were computed to examine the association between essay length and scores. Analysis of variance and t-tests were performed to compare mean essay scores by number of paragraphs, reaching the second page, and use of first-person voice. Chi Square tests of association were performed to compare the frequency of essay features among gender, ethnic, and language sub-groups, and among prompts. Finally, hierarchical multiple regression was performed to examine the contribution of each essay feature variable to the prediction of essay scores beyond its effect in combination with other features.

Results

Association Between Essay Length and Scores

Table 1 shows the descriptive statistics and correlation of number of words and number of paragraphs with essay scores and total SAT writing scores, across the six prompts. The average essay score for the sample in this study was 7.1 (SD = 1.7), which is nearly identical to the average for the SAT college-bound senior population in 2006 (M = 7.2, SD = 1.7). The correlation of number of words and essay score was .62, which is in line with prior research in this area.

Table 2 shows the mean essay scores by number of paragraphs in the essay. Most students wrote three to five paragraphs and the highest mean score was for students writing five paragraphs. An analysis of covariance was performed to ascertain whether there was a significant difference in essay scores by number of paragraphs, holding constant the number of words in the essay. After controlling for number of words, the number of paragraphs was

statistically significant (F (5, 2813) = 16.582, p<.001) but the effect size was negligible. (partial eta squared = .029). A series of post-hoc simple contrasts were performed to compare the mean essay scores for students writing five paragraphs with the mean scores based on the other numbers of paragraphs. All but one of the contrasts were statistically significant. Students writing five paragraphs scored significantly higher than students writing one, two, four, or six or more paragraphs, but did not score significantly different from students writing three paragraphs. This study did not ascertain how many of the students writing five paragraphs used a formulaic structure similar to that criticized in the press. This question will be addressed in the second phase of this research.

Table 1

Means, Standard Deviations, and Correlations of Number of Words and Number of Paragraphs with SAT Essay and Composite Scores

Variable	Mean (Min, Max)	Standard Deviation	Correlation with Essay Score	Correlation with SAT-W Score
N Words	290 (1, 619)	81.5	.62	.42
N Paragraphs	3.7 (1, 9)	1.2	.34	.23
Essay Score	7.1	1.7	---	.73
SAT-W Composite Score	487.5	111.7	.73	---

Table 2

Mean SAT Essay Score by Number of Paragraphs

Number of Paragraphs	N	Mean Essay Score
1	200	5.6
2	199	5.9
3	625	6.7
4	1,081	7.5
5	642	7.7
6 or more	73	7.2

There was some variability in the correlation of number of words and essay score across prompts. The range of correlations across the six prompts was .57 to .68 for number of words and .27 to .38 for number of paragraphs. As shown in Table 3, the magnitude of the correlation of number of words and essay score varied somewhat across gender, ethnic, and language subgroups. The highest correlations (0.65) were found for male and Asian students, and the lowest correlations (0.59) were found for Black and Hispanic students. None of the differences in the correlations were statistically significant across gender, ethnic and language groups.

Reaching the Second Page

Some students believe that reaching the second page of the essay will improve their score by giving the impression of length. Table 3 shows the percentage of students by gender and ethnicity that reached the second page. Overall, a large majority (84%) of essay responses reached the second page. The percentage varied somewhat by prompt but the association of prompt and frequency reaching the second page was not statistically significant ($\chi^2_{(5, N=2820)} = 6.3$, $p = .28$). The mean essay score for those reaching the second page was 7.4 compared to 5.3 for those not reaching the second page. The difference was statistically significant, ($t_{(2,818)} = 25.9$, $p<.001$), with a very large effect size of 1.19. Of those not reaching the second page, 9% received a score of 8 or higher; the highest score among this group was a 10.

When the number of words in the essay was controlled with analysis of covariance, the adjusted mean essay score for essay responses reaching the second page was 7.2 compared to 6.5 for those not reaching the second page. The difference was still statistically significant, but the effect size was reduced to .40, which is considered a medium effect (Cohen, 1988). Adjusting for the number of words in the essay takes into account essays with unusually large handwriting or large spacing. Once this adjustment is made, reaching the second page has a much smaller

relationship with scores. Females, White students, and students speaking English as their first language were much more likely than the other sub-groups to reach the second page. There was a significant association between reaching the second page and gender ($\chi^2_{(1, N=2820)} = 37.0$, p = < .001), ethnicity ($\chi^2_{(3, N=2820)} = 64.9$, $p < .001$), and best language ($\chi^2_{(2, N=2820)} = 28.4$, p < .001).

Table 3

Mean SAT Essay Scores and Essay Features by Gender, Ethnic, and Language Sub-Groups

Sub-Group	N	Mean Essay Score	Correlation of N Words and Essay Score	% Reaching Second Page	% Using First Person
Female	1,535	7.14	.60	88	50
Male	1,285	7.03	.65	79	51
Asian	475	6.85	.65	78	54
Black	594	6.56	.59	77	58
Hispanic	387	6.42	.59	81	56
White	1,364	7.59	.61	90	45
English First Language	1,950	7.34	.62	86	48
English & Another Language	406	6.93	.61	83	54
Another Language	464	6.18	.63	76	58

Use of First-Person Voice

Across all six prompts, half of the essay responses used first-person and the other half did not. There was substantial variability in the use of first-person across prompts, ranging from 38% to 64%. This variation across prompts was statistically significant ($\chi^2_{(5, N=2820)} = 87.6$, p <.001). Some of the essay prompts more readily elicited first-person responses, while others were not as conducive of first-person responses. As shown in Table 3, the percentage using first-person varied across subgroups, ranging from 45% of White students to 58% of Black students. There was a significant association between use of first person and ethnicity ($\chi^2_{(3, N=2820)} = 38.6$,

$p < .001$) and between the use of first person and best language ($\chi^2_{(2, N=2820)} = 17.5$, $p < .001$). White students and students reporting their best language as English used the first-person voice less frequently than the other ethnic and language subgroups. The mean essay score of those using first-person in their essay response was 6.9 compared to 7.3 for those not using first-person. Six percent of the students using first-person voice received a score of 10 or higher.

Prediction of Essay Scores from Features of Responses

Hierarchical multiple regression analysis using stepwise entry was performed to determine the contribution of the number of words, number of paragraphs, reaching the second page, and use of first-person to the prediction of total essay score. The results are shown in Table 4. Across all prompts, the number of words in the essay explained 39% of the variance of essay scores. Whether or not the essay reached the second page explained an additional .015, and whether or not the essay was written in first person explained an additional .011. The number of paragraphs in the essay added very little to the prediction of essay scores (.001), primarily because this variable is highly correlated with number of words (r=.448). The number of words is also highly correlated with reaching the second page (r=.534), yet reaching the second page was still a significant predictor of essay scores, with a semi-partial correlation of .124 after number of words is held constant.

Table 4

Multiple Regression Results of SAT Essay Scores on Essay Features

Variables in the Model	R-square	R-square increment	Standard Error of the Estimate	Sig. F Change
Number of words	.390	.390	1.362	.000
Reached 2nd page	.406	.016	1.345	.000
First person	.417	.011	1.333	.000
Number of paragraphs	.418	.001	1.332	.014

Summary and Discussion

The addition of a writing test to the SAT in March 2005 garnered a lot of attention. Much of this attention focused on the validity of the essay for measuring the type of writing that students do in college. This study was designed to address two frequent criticisms of the SAT essay -- that essay length is the best predictor of scores, and that there is an advantage in using more "sophisticated" examples as opposed to personal experience. As found in other similar studies, essay length is indeed related to scores, but the correlation is not nearly as high as previous critics have claimed (Perelman, 2005). A certain length is needed to effectively develop a point of view on the issue presented in the essay prompt, and this is one of the aspects taken into account in the scoring. In this study, an essay with 220 words was awarded a total score of 11 and an essay with 329 words was awarded a perfect score of 12, demonstrating that very long essays are not necessarily the only ones to receive high scores.

In this study, students who wrote a five-paragraph essay got the highest scores. However, this does not necessarily support the critics' contention that the SAT essay is encouraging writing strategies that are not aligned with high quality instruction. Future research will determine how many of the five-paragraph essays follow the formulaic approach that has been criticized, and compare the scores of the formulaic essays to those not following this format.

The instructions to the SAT essay tell students that they may use a variety of types of evidence to support their viewpoint(s). Examples taken from personal experience or observation that include first-person voice can be very effective, and of those essays receiving a perfect score of 12, 22% were written in first-person. Therefore, despite the differences found in the mean essay scores of students using first-person voice and students who did not, one should not assume that it is the use of first-person voice which leads to lower scores. More importantly,

these results should not be used as evidence to discourage students from using first-person voice, as this may be the most effective way for some students to communicate their essay responses. Without further research using experimental methods, the results from this study should not be used to develop new or modify existing test preparation programs. The second phase of this research will more fully examine essay scores based on the different categories of evidence that students use in their essays.

An examination of the features potentially affecting SAT essay scores is essential to maintain that the SAT writing section is valid for its intended uses. This study is a first step towards gathering validity evidence to support the College Board's intended purpose of the essay and the scoring procedures that have been designed to support it. Additional research will not only examine a wider array of essay features and their association with scores, but will also focus on the nature of the essay prompts and discern whether there are characteristics of the prompts that are associated with higher or lower scores for several different test-taker sub-groups. While this research is non-experimental in nature, future studies utilizing an experimental design would be useful. For example, essay responses can be constructed on the same topic using the same examples, differing only in length. These essays can then be scored using typical operational approaches and the scores can be compared. The research described in this paper may benefit other testing programs that include essay assessments. The careful analysis of response features and the identification of potential construct-irrelevant features in essay assessments are important elements in the establishment of the assessment's content and construct validity.

References

Baron, D. (2005, May 6). The College Board's new essay reverses decades of progress toward literacy. *The Chronicle of Higher Education,* pp. B14-B15.

Breland, H.M., Camp, R., Jones, R.J., Morris, M.M., & Rock, D.R. (1987). *Assessing writing skill* (College Board Research Rep. No. 11). New York: College Entrance Examination Board.

Breland, H.M., Danos, D.O., Kahn, H.D., Kubota, M.Y., & Sudlow, M.W. (1991). *A study of gender and performance on Advanced Placement history examinations* (College Board Research Rep.No. 91-4). New York: College Entrance Examination Board.

Breland, H.M., & Jones, R.J. (1988). *Remote scoring of essays* (College Board Research Rep. 88-3). New York: College Entrance Examination Board.

Cioffi, F.L. (2005, May 20). Argumentation in a culture of discord. *The Chronicle of Higher Education, p. B6.*

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

College Board. (2004). *ScoreWrite: A guide to preparing for the new SAT essay*. New York: College Entrance Examination Board.

College Board (in press). *Technical Manual for the SAT Reasoning Test*. E. Kimmel (Ed.). New York: College Entrance Examination Board.

Hoover, E. (2005, April 22). And on the 11th day, the readers rested. *The Chronicle of Higher Education,* p. A39.

Katzman, J., Lutz, A., & Olson, E. (2004, March). Would Shakespeare get into Swarthmore? *The Atlantic Monthly,* pp. 97-100.

Nunnally, T.E. (1991). Breaking the five-paragraph-theme barrier. *The English Journal, 80,* 67-71.

Perelman, L. (2005, May 29). New SAT: Write longly, badly, and prosper. *Los Angeles Times,* p. M5.

Powers, D.E., Fowles, M.E., Farnum, M., & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement, 31,* 220-233.

Setoodeh, R. (2005, March 14). Penmanship, the newest SAT worry. *Newsweek,* p. 48.

Skutches, G. (2005, July 24). New SAT essay format is "teaching to the test," not teaching writing, critical thinking [Opinion]. *The Morning Call,* p. D1.

Thomas, P. (2004). The negative impact of testing writing skills. *Educational Leadership, 62,* 76-79.

Wesley, K. (2000). The ill effects of the five paragraph theme. *English Journal, 90,* 57-60.

Winerip, M. (2005, May 4). SAT essay test rewards length and ignores errors of fact. *New York Times,* p. B9.

Appendix

Essay Features Study
Coding Protocol for Phase I

Student ID

Prompt ID

Prompt Descriptor

Number of Words

Number of Paragraphs

Essay response reached second page Yes No

Essay response included first-person pronoun ("I") Yes No