Quality vs. Quantity: How to Score Higher on the SAT Essay Component

Milo Beckman

This paper discusses the correlation between the length of the essay written by a student on the SAT and the score the student is given on the essay. As each essay grader is given hundreds, sometimes thousands[1] of essays to score and less than three minutes to score each one[2], graders often resort to length to measure a student's essay-writing capabilities. This paper tests to see if such a correlation exists for SAT essay-writers at Stuyvesant High School.

1 Introduction

In November 2009 I took the SAT without having prepared, in order to gage how much practice I would need before I took it again. I was surprised to find how little time was set aside for the essay component, as I had never before been asked to plan and write an essay in the course of 25 minutes. Although I would have liked to write three body paragraphs, the time constraint forced me to write only two, but I nonetheless felt good about the essay. When I received my scores, I learned that both scorers had given my essay a 4 out of 6, making my total a decent but not excellent score of 8.

When I took the SAT for a second time in March 2010, I was more prepared for the time constraint. I had time to write a significantly longer essay, although one of my body paragraphs seemed to be proving a point far different

from my thesis. Additionally, I discovered later that day that I had made an important factual error regarding a sequence of historical events, making one of my supporting points essentially false. This essay was given a 9.

What did I do in March but not in November that gave me that extra point? I wrote more.

2 Les Perelman and the MIT Study

I was naturally curious as to how these essays were graded, so I did some quick research. The first result was the College Board's own description of the grading process, in which it explains that the SAT essay is "scored in a holistic manner by qualified educators."[3] Without saying what this holistic manner is, it goes on to say that the essay is scored not on individual traits, but on the "total impression the essay creates."

Not satisfied with this explanation, I continued my search. I soon found out from multiple sources about Dr. Les Perelman, Writing Director at MIT, who was also unsatisfied with the College Board's description of the essay grading policy. While attending a panel regarding the newly created SAT essay in 2005, he realized that the rule governing grading was "the longer the essay, the higher the score."[4] When a panel member from the College Board told him otherwise, he decided to take matters into his own hands.

Dr. Perelman gathered every sample essay the College Board made public and counted the number of words in each. These 54 sample essays showed an astounding correlation − from the word count alone, Dr. Perelman discovered that one could guess the score on the 1-6 scale over 90% of the time.[4] Although this result received wide publicity, there were several aspects of the study that I found unconvincing.

First and most importantly, Dr. Perelman used the College Board's sample

essays rather than essays written by actual students. He gathered his data "shortly after the test was first administered in March,"[5] before the first round of actual SAT essays had been graded. As a result, his study was limited to the 54 sample essays distributed by the College Board prior to the exam intended to prepare the students.[4]

It is very plausible that these sample essays are not representative of the true scoring policy. For example, perhaps the length-score correlation in the samples was to encourage students preparing for the exam to write longer essays. Even if this is not the case, the fact that the essays were written by the same organization could easily distort the results and create a false positive.

Second, he used word counts rather than line counts. It is unlikely that, if the grading is in fact length-based, the reader's "general impression" will be proportional to the number of words. A more visible measure of an essay's length is its line count, as a reader will be more likely to formulate an opinion based on the amount of space left empty than based on the number of words the writer used. Furthermore, many sources, including Dr. Perelman himself,[5] say that the essay graders are impressed with students who use big words, which is not taken into account when using word counts.

Several other issues with Dr. Perelman's study include his relatively small sample size of 54 essays and his use of the 1-6 scale rather than 2-12. This "MIT study" was interesting in theory, and the results were certainly astonishing, but I still was not entirely satisfied.

3 The Theory

My own experience and the MIT study seemed to point to a hypothesis: there is a correlation between the length of an SAT essay and the score the essay is given. Why would such a correlation exist?

The first reason lies in the method used by the College Board to select its essay graders. Dan Verner told the story of his training, which "consisted of 17 hours of reading sample essays, doing 10 at a time and then being assessed on his scoring."[1] He got the job because the scores he gave the essays agreed for the most part with the College Board's opinions. As Dr. Perelman's research shows, the College Board's scoring method is heavily influenced by length, so those graders like Verner who pass the training will probably also have length-based scoring methods.

The second reason is the conditions under which graders score essays. The number of essays given to the average grader is said to be around eight hundred[1], and the amount of time spent per essay has been variously estimated around two to three minutes[6], approximately two minutes[7], 2 1/2 minutes[1], and at most 60 seconds[8]. As a result, the graders will be unable to analyze each essay thoroughly, and will more likely base the grade on their original "total impression," as the College Board says. Only having skimmed the essay, this "total impression" could very likely be based on length.

With the essay graders selected to agree with the College Board's length-based grading system, and with the essays graded in under three minutes, it seems likely that a correlation would arise between length and score. However, Stuyvesant students are stereotypically lazy and smart, so perhaps the same rule wouldn't hold. To test for a correlation, I decided to perform a study of my own on Stuyvesant students.

4 Data Collection

The first step was to determine a method of data collection. Fortunately, every student who has taken the SAT has a scanned copy of their essay available to them on the College Board website. I just needed to decide on a measure of

length and a method of finding participants.

Regarding the measure of length, I decided for several reasons to use line count. First, as mentioned above, it seems more likely that if SAT graders do indeed score based on length, their impression of length will probably be based on the amount of space left empty, which is measured in lines. Second, as also mentioned above, the alternative measure of word counts does not factor in the use of long words, which is generally agreed to improve one's score. Finally, I realized that I would receive far more data points if I chose line count, as it is unlikely that students would be willing to count the number of words in their essays.

Regarding the method of finding participants, I had to consider several factors. First, for a participant to be able to give the line count of his essay, he would have to be at a computer with internet access. Thus it would make sense for the medium for gathering to be internet-based too. My first thought was a chain email, but I soon ruled this out for a number of reasons: many students would think it was junk mail, most students would decide not to forward the email to others, and the number of email contacts I had access to was considerably smaller than the sample size I would have liked. For all these reasons, I decided to create a Facebook group and invite every Facebook friend of mine in the Stuyvesant High School network.

As would be the case with any data collection method, there are several statistical concerns with regard to using a "Facebook blast" as mentioned above. Most notably, this method creates an inherent self-selection bias, as students who are unsatisfied with their score on the essay component may be less willing to report it. However, such a bias should not affect the outcome of a test for correlation; even in the worst case scenario in which all students below a certain score choose not to report, the test would still show whether or not there was

a correlation among students with scores above a certain level, albeit with a slightly smaller sample size and score range.

Another possible concern regarding the use of a Facebook blast is the possibility that my friends on Facebook do not represent a random sample of Stuyvesant students, and may have, for example, higher essay scores on average. For the same reason as the self-selection bias, this should not affect the outcome of a linear regression to a large extent.

One final issue I considered before sending off the Facebook blast was what to do about students who took the test multiple times. A concern I had was that including all scores for each individual may lead to an issue involving independence of variables. However, I decided to ask for all scores from each person so that I could have as large a sample size as possible. This approach would also allow me to perform further tests based on individual students' score changes, if I so desired.

After making these decisions, I sent out the Facebook blast. Within a few weeks, I had collected 115 data points.

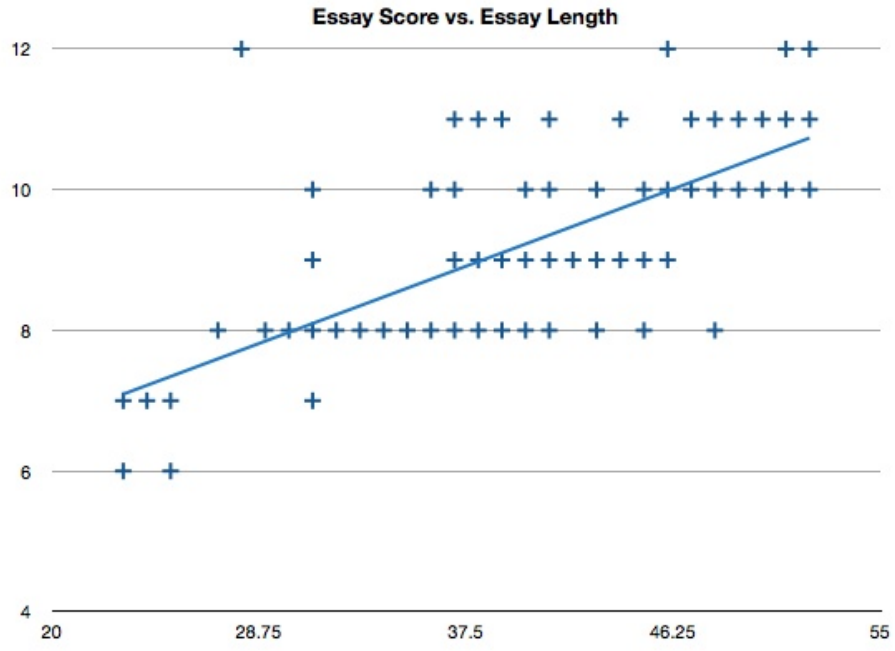
Number	Length	Score
1	24	7
2	23	7
3	49	10
4	38	8
5	31	9
6	44	11
7	29	8
8	48	10
9	43	8
10	45	9
11	32	8
12	50	10
13	43	8
14	52	12
15	42	9
16	33	8
17	51	11
18	41	8

Number	Length	Score
19	51	10
20	50	11
21	51	11
22	51	11
23	31	8
24	51	12
25	49	10
26	39	9
27	31	7
28	42	9
29	43	10
30	41	9
31	39	8
32	31	8
33	46	9
34	40	9
35	51	11
36	51	12

Number	Length	Score
37	46	10
38	51	10
39	43	9
40	36	10
41	47	10
42	37	9
43	46	12
44	44	9
45	47	11
46	50	10
47	52	11
48	40	10
49	43	10
50	36	8
51	42	9
52	37	8
53	52	10
54	45	8
55	37	10
56	52	10
57	43	8
58	51	11
59	45	8
60	41	10
61	43	10
62	46	10
63	51	12
64	44	11
65	43	9
66	30	8
67	43	9
68	48	11
69	35	8
70	48	11
71	38	9
72	25	7
73	34	8
74	47	10
75	37	8
76	41	11

Number	Length	Score
77	40	8
78	31	8
79	38	11
80	32	8
81	46	9
82	45	9
83	29	8
84	39	9
85	48	8
86	37	8
87	50	11
88	41	11
89	23	6
90	29	8
91	31	10
92	47	11
93	41	9
94	24	7
95	51	10
96	45	10
97	39	11
98	52	10
99	49	11
100	47	10
101	25	6
102	36	10
103	52	10
104	50	11
105	30	8
106	37	11
107	37	8
108	47	10
109	30	8
110	33	8
111	38	8
112	27	8
113	34	8
114	46	12
115	28	12

A graphical view of this data is on the following page. Note that most points on the graphical display represent multiple data points, as many essays had the same length and score.



Without even performing any tests, I could see a very clear correlation arising between length and score. I also noticed that in nearly every case in which a student sent in multiple scores, the higher score corresponded to the longer essay. At this point I was almost certain as to what the result of the test would be, and was more curious as to how significant the result would be.

5 Linear Regression Test

The obvious test to perform on this data is a linear regression to determine whether or not there is a significant positive linear correlation between essay length and essay score. For this test, my alternative hypothesis is that $\beta > 0$, where β is the coefficient of essay length in the linear regression. This hypothesis essentially translates to the following: "For Stuyvesant students, longer essays on the SAT receive higher scores."

The computer output for this test is shown below.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.706185733
R Square	0.4986982895
Adjusted R Square	0.4942619912
Standard Error	1.0115732444
Observations	115

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	115.0301811088	115.03018111	112.41315466	1.19511E-18
Residual	113	115.6306884564	1.0232804288		
Total	114	230.6608695652			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4.2043767078	0.49581258291	8.4797700839	9.77053E-14	3.2220825879	5.1866708276
Essay Length	0.1255811846	0.01184448024	10.602506999	1.19511E-18	0.1021151338	0.1490472355

This result gives a p-value of 10^{-18}, which is for statistical purposes ≈ 0. At an alpha level of 0.01, we may reject the null hypothesis. There is very strong statistical evidence to conclude that for Stuyvesant students, longer essays on the SAT receive higher scores.

It is also interesting to note that we were able to get such an enormously significant result even with Essay 115 included. With a score of 12 in only 28 lines, Essay 115 is about as far from the regression line as is reasonably possible in this experiment. It is possible that this is due to very small handwriting, or that the participant who sent in this score was lying. With Essay 115 excluded, the p-value is closer to 10^{-22}.

6 Tests for Individual Improvement

As of yet, we have shown that as a general rule, longer essays receive higher scores. However, a question that is of greater interest to students taking the SAT is whether or not an individual student will benefit from writing more. While these may seem the same, there is an important distinction. What if,

for example, "smarter" students tend to write more on the SAT essay, and these same "smarter" students tend to score higher? A test answering the first question would show a correlation between long essays and high scores, while a test answering the second question would show no significant difference between a student's scores when he writes more or less, because in this hypothetical scenario his score is based primarily on his "smartness."

Then how can we create a test to answer this second question? We need to compare the different scores of individual students who took the SAT multiple times. As many participants sent in multiple scores, we can use these students as our sample and compare their scores on their shortest and longest essays. A table of students who sent in multiple scores is shown below.

Number	Shortest Essay Score	Longest Essay Score
1	8	12
2	8	8
3	8	11
4	6	10
5	10	11
6	6	10
7	9	9
8	8	11
9	7	8
10	10	10
11	8	11
12	8	10
13	10	10
14	9	11
15	9	12
16	8	9
17	8	10
18	7	7
19	8	9

The results of this table are astounding: not a single student who sent in multiple scores received a lower grade on their longest essay than on their shortest essay (though some students received the same score). We wish to test the alternative hypothesis $\mu_{L-S} > 0$, where μ_{L-S} is the average score difference between longest and shortest essays for an individual student. This hypothesis essentially translates to the following: "Stuyvesant students receive higher scores when they write longer SAT essays."

Since the sample size is 19, we are on the border of being able to invoke the Central Limit Theorem. As a result, we will perform two tests for this hypothesis: one that uses the Central Limit Theorem and one that does not. For the test that uses the Central Limit Theorem, we will use a t-test with matched pairs design. The computer output is shown below.

t-Test: Paired Two Sample for Means		
	Variable 1	Variable 2
Mean	8.157895	9.947368
Variance	1.362573	1.830409
Observations	19	19
Pearson Correlation	0.322158	
Hypothesized Mean Difference	0	
df	18	
t Stat	-5.28845	
P(T<=t) one-tail	0.000025	
t Critical one-tail	1.734064	
P(T<=t) two-tail	0.00005	
t Critical two-tail	2.100922	

This test gives a p-value of 2.5×10^{-5}. At an alpha level of 0.01, we may reject the null hypothesis. There is strong statistical evidence to conclude that Stuyvesant students receive higher scores when they write longer SAT essays.

For the test that does not use the Central Limit Theorem, we will use a

sign test with matched pairs design. As 14 of the students' scores increased as length increased and no student's score decreased, this test gives a p-value of $2^{-14} = 6.1 \times 10^{-5}$. At an alpha level of 0.01, we may reject the null hypothesis. There is strong statistical evidence to conclude that Stuyvesant students receive higher scores when they write longer SAT essays.

7 Implications of Results

In every formulation, this study shows that length plays a large role in determining score on the SAT essay. This result has numerous implications for the College Board.

First, change the process by which essay graders are chosen. The current method encourages conformity to a flawed set of base scores determined by the College Board. Essay graders should be selected based on past performance as a teacher or grader of similar essays, and should be allowed to grade even if their opinions vary from those of the College Board.

Second, relax the conditions under which graders score essays. Even five minutes would not be nearly enough time to decide the quality of an essay. While the reader may get a "total impression" from a speed-read, they may miss any subtleties the writer meant to convey. Anything measured in three minutes of reading an essay is not representative of an essay-writer's skill.

Finally, expand the amount of time allotted for writing the essay. Just as three minutes is not enough time to judge a good essay, 25 minutes is not enough time to write one. As it stands, the SAT essay component measures something far different than essay-writing skill (writing speed, it seems). Even the 40 minutes given per essay by the AP English exam allow for slightly more intelligent pieces to be written.

And the implication for students? Write more.

Notes

[1] Jay Mathews, "The SAT Grader Next Door" *Washington Post*, August 1, 2005

[2] *http://www.princetonreview.com*

[3] *http://professionals.collegeboard.com*

[4] Michael Winerip, "SAT Essay Test Rewards Length and Ignores Errors" *New York Times*, May 4, 2005

[5] *http://www.jewishworldreview.com*

[6] *http://www.fastweb.com*

[7] *http://hubpages.com*

[8] *http://school.familyeducation.com*