Do Changes in SAT Reading Passage Length Change Student Performance?

Do Changes in SAT Reading Passage Length Change Student Performance?

Jim Patterson
Monika Ulrich
Jessica Marini
Michael Hurwitz

COLLEGE BOARD

JULY 2026

© 2026 College Board. College Board, SAT, and the acorn logo are registered trademarks of College Board. PSAT/NMSQT is a registered trademark of College Board and National Merit Scholarship Corporation. Bluebook and PSAT are trademarks of College Board.

Contents

Executive Summary 1

Introduction 3

Research Design and Data 5

1. Observational Research Design and Data 5
2. Experimental Research Design and Data 6

Results 9

1. Observational Research Results 9
2. Experimental Research Results 10

Discussion 14

References 15

Appendix 17

I. Study Recruitment 17

II. Study Consent Form 18

III. Simulated Test Materials 20

- Literature 21
- History/Social Studies 23
- Humanities 24
- Science 26

IV. Model Results Detail in Tabular Form 28

Executive Summary

When College Board launched the digital SAT®, one of the more visible changes was to the format of the multiple-choice literacy section. The previous paper and pencil SAT Evidence-Based Reading and Writing (ERW) section featured longer reading passages with multiple questions pertaining to each, while the digital SAT Reading and Writing (RW) section contains shorter reading passages each paired with a single question. The structural change has attracted broad public attention, including from major national media outlets, with some claiming that the shift to shorter reading passages made the new SAT RW format less demanding. Those claims have been made without empirical support. We address these claims directly with two separate research methodologies and a sample of nearly 200,000 students. We find that reading passage length does not meaningfully affect student performance on SAT reading and writing questions.

An observational study examined approximately 90,000 students who naturally spanned the SAT format change, taking the paper and pencil SAT in fall 2023 and the digital SAT in spring 2024. Because the same students took both versions, differences in individual reading ability, background, and test-taking experience are automatically held constant. The correlation between the paper and pencil SAT ERW and digital SAT RW scores in this sample is very high (0.87); it is also equivalently high when compared to test-retest correlations in a similar group of nearly 100,000 students who retook the SAT and experienced no change in format (0.89). If changes in reading passage length had influenced SAT RW performance, we would have expected the paper-to-digital correlation to deviate from the high paper-to-paper baseline correlation among SAT retakers' scores. The correlation in scores between SAT formats also remained similar across students with low, medium, and high prior literacy achievement levels relative to baseline correlations.

An experimental study recruited 367 high school students to take a specially designed research test form in which they answered matched multiple-choice questions associated with both shorter and longer versions of the same reading passages. The study covered all four subject areas sampled on the SAT ERW and RW sections: literature, history/social studies, the humanities, and science. Passage content, text complexity, question skill, and estimated difficulty were held constant—only passage

length varied. The same students answered questions associated with both the shorter and longer passages, which controlled for individual differences. The results run counter to concerns about shorter reading passages being easier. In three of the four subject areas (history/social studies, the humanities, and science), students performed *better* or similarly on questions associated with longer passages than shorter ones. In the fourth area, literature, students showed marginally higher scores on shorter-passage items, although the effect size across the full sample was negligible (d = 0.13, below the threshold for practical significance); when disaggregated by prior literacy achievement level, the only meaningful effect size was specific to low achievement students (medium size; d = 0.63). No student subgroup by subject area or prior achievement showed a pattern of meaningfully better performance associated with the digital SAT's shorter passage length format.

These two distinct methodological approaches find that reading passage length does not meaningfully affect student performance on SAT reading and writing questions. The evidence from both the observational and experimental approaches consistently supports the validity and rigor of the digital SAT's format, reinforcing the substantial body of prior research published since the digital SAT's launch.

Introduction

College Board shifted its SAT Suite of Assessments from a paper and pencil format to a digital format beginning with international testing in spring 2023 and domestic testing in spring 2024. The change to digital testing enabled the implementation of a multistage adaptive testing design capable of accurately measuring students' attainment of the same set of knowledge and skills more quickly. As a result, the digital SAT lasts just over two hours, compared to three hours for the paper and pencil test, while maintaining the measurement precision of the longer test.

The Reading and Writing (RW) section of the digital SAT uses shorter reading passages than the previous paper and pencil version, the Evidence-Based Reading and Writing (ERW) section. Students taking the digital SAT RW section are given 54 passages of 25 to 150 words each and respond to one high-quality question associated with each passage. This new SAT RW format contrasts with the previous paper and pencil SAT's ERW section, which consisted of nine reading passages of 400 to 750 words and several multiple-choice questions about each passage (College Board, 2024a).

The format change from longer reading passages with multiple questions about each to shorter reading passages with a single question each was intended to better facilitate digital adaptive testing on a range of devices and was not designed to change the rigor of the SAT. College Board conducted extensive testing of SAT RW questions to ensure comparable cognitive complexity, comparable question difficulty, and assessment of the same skills and knowledge between the test versions. For nearly three years before launching the digital SAT, College Board collected and analyzed approximately 16 million responses from 940,000 students and showed that the RW questions presented the full range of challenge to the SAT population and satisfactorily distinguished students by performance level.

Since launching the digital SAT, College Board has published a significant amount of evidence in support of its decision to move to shorter RW passages, summarized in College Board (2023). That research shows that students' digital SAT scores are strongly related to their paper and pencil SAT scores—overall, in Math, and in Reading and Writing—as well as to other common education measures, such as high school GPA and average AP® Exam scores (Marini et al., 2022). The evidence also shows that digital

SAT RW questions are capable of eliciting cognitively complex, higher-order thinking from students in general (College Board, **2024b**, **2024c**) and from various student subpopulations (College Board, **2025a**, **2025b**, **2025c**).

Despite this body of evidence, the shorter reading passage length on the digital SAT has raised questions about whether the new format is less rigorous (e.g., **Finley, 2025**; **Bloomberg Editorial Board, 2025**). We provide new evidence to address these questions using both administrative data on nearly 200,000 SAT takers and data from an experimental study that directly tests the effect of passage length on performance. Both empirical approaches affirm that the rigor of the SAT has been maintained and that student performance on SAT-style multiple-choice literacy questions is not meaningfully affected by reading passage length.

Research Design and Data

Our research directly tackles questions raised in the discourse around reading passage length: *What is the effect of reading passage length on students' performance on the SAT, and do any effects vary with student characteristics?*

We address these research questions using two separate methodological approaches: (1) an observational study and (2) an experimental study. Both approaches rely on within-student comparisons, holding constant individual factors that are unobservable in studies that use cross-student comparisons. Use of within-student comparisons eliminates the confounding effects of individual reading ability, socioeconomic background, and test-taking experience.

1. Observational Research Design and Data

The observational study relies on a subset of students who took both the digital SAT—with shorter reading passages—and the paper and pencil SAT—with longer reading passages. This subset of nearly 90,000 students happened to span the changes in the SAT format such that they took the paper and pencil SAT in fall 2023 and the digital SAT in spring 2024. We refer to this group as the paper-to-digital SAT subsample. This overlap in SAT formats only existed during the 2023–24 academic year, as the paper and pencil SAT was retired when the digital SAT was launched. The students in this subsample were not recruited for a formal research study; their SAT-taking patterns occurred naturally.

We examine the correlation between SAT ERW (paper and pencil) and SAT RW (digital) scores for this paper-to-digital subset of students. While we expect this correlation to be high based on prior research (Marini et al., 2022), the correlation should be very close to the SAT test-retest correlation observed when reading passage length does *not* change. Thus, the paper-to-digital correlation from the 2023–24 subsample is compared to paper-to-paper SAT correlations among nearly 100,000 SAT retakers in the 2022–23 academic year who took the paper and pencil SAT in both fall 2022 and spring 2023. In these analyses, student factors are fully controlled for because the same students are taking and retaking the SAT.

Descriptive statistics for the 2023–24 paper-to-digital and 2022–23 paper-to-paper SAT retaking subsamples are provided in table 1. The composition of both subsamples is similar across demographics and prior academic achievement. One exception is the higher proportion of students self-reporting unknown race/ethnicity in the paper-to-digital sample. This shift coincides with the timing of the Supreme Court decision in the *Students for Fair Admissions v. Harvard* case related to race and college admission, which likely influenced some students to not report their race/ethnicity information.

Table 1. **Descriptive Statistics for Observational Study**

	Paper-to-Digital SAT Retaking Sample: Fall 2023 Paper/ Spring 2024 Digital		Paper-to-Paper SAT Retaking Sample: Fall 2022 Paper/ Spring 2023 Paper	
	n	**%**	***n***	**%**
Gender				
Female	46,889	53%	51,057	52%
Male	41,769	47%	45,869	47%
Race/Ethnicity				
Asian	18,127	20%	22,800	23%
Black/African American	4,829	5%	7,163	7%
Hispanic/Latino	11,318	13%	14,871	15%
Native Hawaiian/Other Pacific Islander	100	<1%	205	<1%
American Indian/Alaska Native	469	1%	315	<1%
Two or More Races	3,390	4%	4,240	4%
No Response	17,458	20%	4,668	5%
White	33,125	37%	43,091	44%
Urbanicity				
City	24,417	27%	25,755	26%
Rural	10,523	12%	11,430	12%
Town	2,546	3%	2,807	3%
Suburb	42,969	48%	46,708	48%
Prior Achievement on PSAT RW				
Low (<420)	5,657	6%	5,050	5%
Medium (420–600)	41,924	47%	45,968	47%
High (>600)	21,255	24%	23,455	24%
Missing	19,980	22%	22,880	24%

2. Experimental Research Design and Data

For the experimental study, we recruited high school students to participate in a study that would directly test the effect of passage length on performance in a manner that could control for all facets of reading passage length and complexity level as well as multiple-choice question difficulty and skill tested. Participating students took two simulated SAT RW modules in Bluebook™, College Board's digital test application. Across the simulated SAT RW modules, longer and shorter versions of the same reading passages were included in the four subject areas that appear in the digital SAT RW section: literature, history/social studies, the humanities, and science.

In total, 367 students—primarily high school juniors and seniors—participated in the study and formed the final analytic sample.[1] Students' prior literacy achievement—as represented by previous SAT Suite RW section scores—and performance on the simulated SAT RW modules were measurable on the same scale, given that the SAT Suite uses an item response theory (IRT) scoring method.

Prior Literacy Achievement and Simulated Question Difficulty Categories

Low Prior Achievement/Easy Questions:	RW section scores <420
Medium Prior Achievement/ Moderate-Difficulty Questions:	RW section scores 420–600
High Prior Achievement/Hard Questions:	RW section scores >600

After stratifying the sample by prior academic achievement on the PSAT/NMSQT®, PSAT™ 10, or SAT RW section, students were randomized into two treatment arms: test form A or test form B. All students, regardless of treatment arm, answered a total of 16 SAT-style multiple-choice questions associated with eight reading passages. All materials were the same between test forms A and B, with the only difference being whether the module of longer passages appeared before or after the module of shorter passages.[2]

The study was conducted in Bluebook via a customized "research test form" administration on February 5, 2026. All test content was accessible for test takers with disabilities, as is standard for SAT Suite tests, and participants had access to all the universal tools and accommodations options they would normally have had for the RW section during this activity.[3]

If reading passage length has no effect on multiple-choice question performance, all else equal, performance differences across questions related to shorter and longer reading passages would not be statistically significantly different from one another.

Descriptive statistics for the analytic sample are provided in table 2. Female and Asian participants are somewhat overrepresented. Most students attended public high schools and were in grade 11 or 12. Despite special efforts to recruit them for this study, students with low prior academic achievement in literacy were underrepresented relative to those with medium or high prior academic achievement levels.

[1] Study registration was completed by 699 students, 386 students participated in the experiment on the scheduled day, and 19 participants' data were removed based on incomplete and obviously low-effort responses (e.g., spending fewer than two minutes testing in either or both modules), resulting in an analytic sample of 367. See Appendix I for the study recruitment email and Appendix II for the consent form.

[2] See Appendix III for the reading passages and multiple-choice questions used in the experimental study as well as information about how test forms were scored.

[3] Participants (other than any with extended time accommodations) were given 45 minutes to complete the activity. At an average of 2 minutes and 30 seconds per question, we anticipated this to be enough time for the vast majority of participants to be able to read the passages and answer the questions without speededness (i.e., undue rushing due to time constraints). For comparison, students taking the digital SAT RW section operationally have an average of 1 minute and 11 seconds per question.

Table 2. **Descriptive Statistics for Experimental Study**

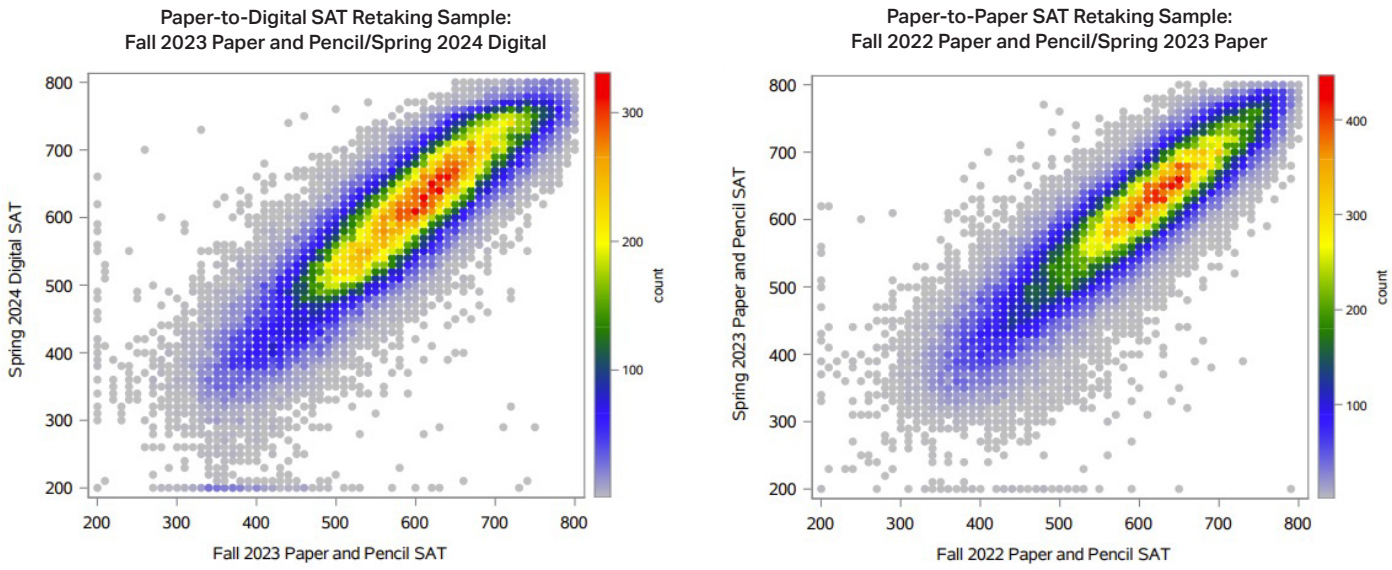
	n	%
Gender		
Female	208	57%
Male	157	43%
Race/Ethnicity		
Asian	111	30%
Black/African American	39	11%
Hispanic/Latino	72	20%
Native Hawaiian/Other Pacific Islander	0	0%
American Indian/Alaska Native	4	1%
White	110	30%
Two or More Races	18	5%
No Response	13	4%
Urbanicity		
City	142	39%
Rural	44	12%
Town	7	2%
Suburban	174	47%
High School Type		
Home School	1	<1%
Private	23	6%
Public	343	93%
Grade in High School		
10	33	9%
11	234	64%
12	99	27%
No longer in high school	1	<1%
Prior Achievement on SAT Suite RW		
Low (<420)	24	7%
Medium (420–600)	187	51%
High (>600)	156	43%

Results

1. Observational Research Results

Figure 1 shows a strong correlation between SAT ERW scores (paper and pencil, longer reading passages) and SAT RW scores (digital, shorter reading passages) for students who took both formats (left panel). Visually, the correlation is as strong as the typical correlation between SAT retakers' reading and writing scores when there are no format changes to the test (right panel).

Figure 1. **Plot of Reading and Writing Score Relationships for SAT Retakers**

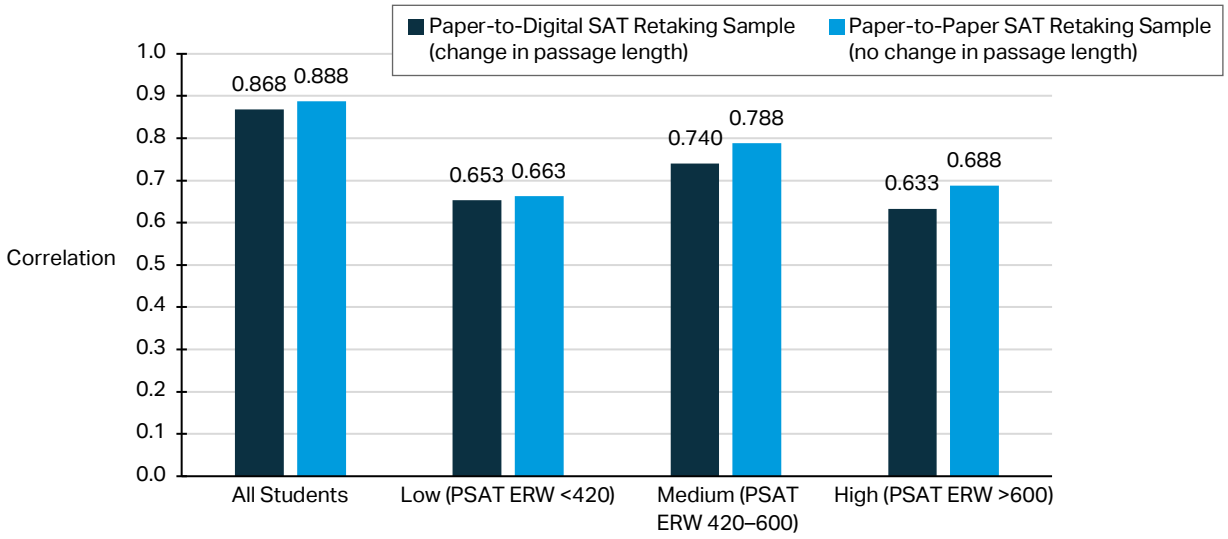


Statistically, the correlation between scores plotted in the left panel is 0.87 and the correlation between scores plotted in the right panel is 0.89.[4] The similar correlations suggest that reading passage length changes between paper and pencil and digital SAT tests had no meaningful effect on SAT RW scores.

[4] Other research has demonstrated this consistency in correlations in different samples of students who were recruited as part of validity and concordance research conducted prior to the launch of the digital SAT (Westrick et al., 2023; College Board, 2023).

We also examine whether there are any differences in the relationships in figure 1 by students' prior literacy achievement as measured by performance on the PSAT/NMSQT® or PSAT™ 10 Reading and Writing section. More than three-quarters of students in our SAT retaking samples in table 1 also have a PSAT score that we classify as signaling low, medium, or high prior literacy achievement. Figure 2 shows the correlations across SAT scores among these three prior achievement subgroups. In all cases, the relationships between fall 2023 paper and pencil SAT ERW scores and spring 2024 digital SAT RW scores are similar to typical retesting correlations when there are no format changes to the assessment.

Figure 2. **SAT Reading and Writing Score Correlation Among Retakers, by Retest Format and PSAT Score**



2. Experimental Research Results

Student performance on the experimental activity was calculated by counting each correct multiple-choice response as 1 "point" and counting each incorrect response as 0 "points." Thus, a given student could earn up to 16 points, as there were 16 multiple-choice questions to answer. Because eight of these questions were associated with each passage length (shorter, longer), the maximum score on shorter-passage questions was 8, as was the maximum score for longer-passage questions. Because two of these questions were associated with each passage length in each sampled subject area (e.g., shorter length, literature subject area), the maximum subject area score by passage length was 2.

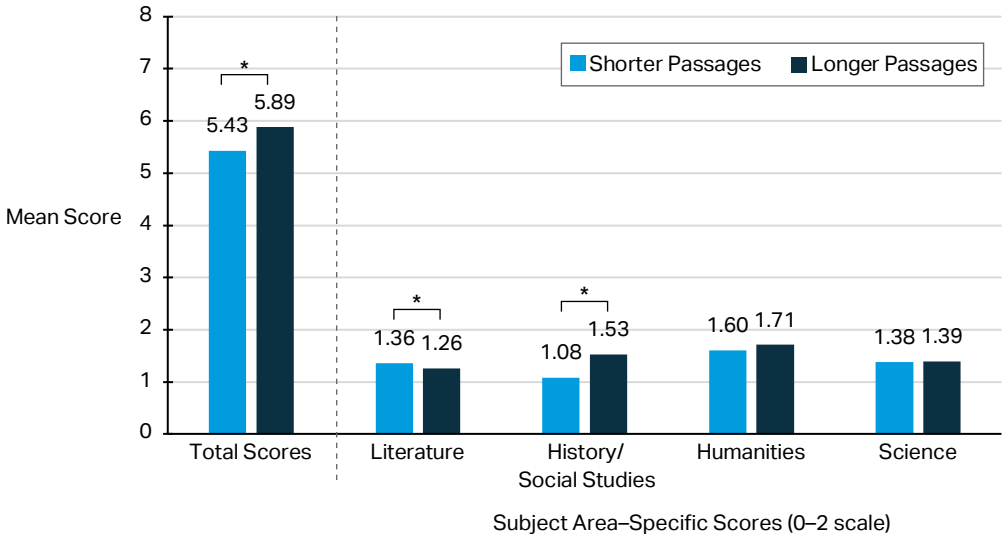
Student performance on both shorter- and longer-passage questions is shown in figure 3.[5] The average total score for all questions associated with shorter reading passages was 5.43 (out of a maximum of 8), while the average total score on questions associated with longer reading passages was 5.89. Based on paired *t*-tests, this difference is statistically significant in favor of higher scores for questions associated with longer reading passages.

[5] See Appendix IV for detailed results in tabular form.

Figure 3 also shows average performance on questions associated with the four subject areas. In history/social studies, the humanities, and science, patterns follow those of the total scores, with students performing better or similarly on longer-passage questions than on shorter-passage questions.[6] In literature, the mean score for the shorter-passage questions was statistically significantly higher than that for the longer-passage questions. The magnitude of mean difference by passage subject area and length was greatest for history/social studies (0.44) and smallest for science (0.01), with the magnitude of the mean difference for literature falling toward the lower end of the range (0.10).

We also calculate effect sizes to better understand the practical significance of the results displayed in figure 3. The standardized mean differences indicate a small effect size favoring higher scores for questions associated with longer reading passages in terms of total scores, a medium effect size favoring higher scores for questions associated with the longer passage in history/social studies, and no meaningful effect of passage length in literature, the humanities, or science. See Appendix table A2 for additional details.

Figure 3. **Mean Scores for Questions Associated with Shorter vs. Longer Reading Passages, by Total Scores and Subject Area–Specific Scores**



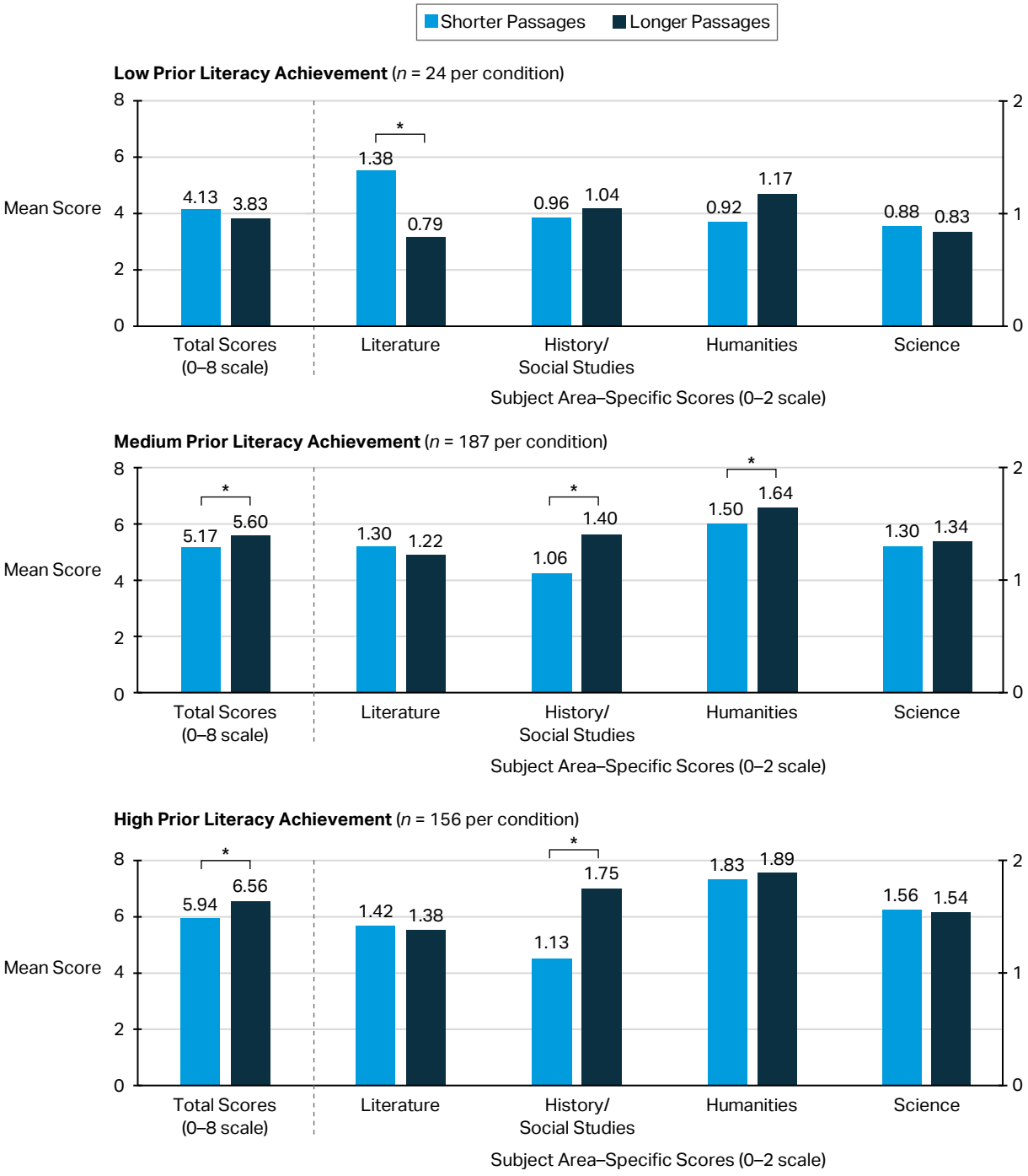
* Statistically significant difference between shorter and longer passages ($p < .05$)
Note: Total scores are on a 0–8 scale; subject area–specific scores are on a 0–2 scale.

We also investigate whether reading passage length is related to performance for students with different levels of prior academic achievement in literacy. Figure 4 shows performance on questions associated with shorter and longer passages for students with low, medium, and high prior literacy achievement. In terms of total scores, all three subgroups of students perform better or similarly on the questions associated with the longer passages. Examination of results by subject area reveals that performance is statistically significantly higher on questions associated with shorter passages only among students with low prior literacy achievement and only in literature. In all other

[6] There may be a ceiling effect for scores on the humanities questions, as the average scores are within one standard deviation and include 2.00 (or a perfect score) for both shorter and longer passage lengths.

subject areas and student subgroups, performance is higher or similar on questions associated with longer passages.

Figure 4. **Mean Scores for Questions Associated with Shorter vs. Longer Passages, by Prior Literacy Achievement, Total Scores, and Subject Area–Specific Scores**



* Statistically significant difference between shorter and longer passages ($p < .05$)
Note: Total scores are on a 0–8 scale; subject area–specific scores are on a 0–2 scale.

We also calculate effect sizes to better understand the practical significance of the results displayed in figure 4. The standardized mean differences reveal effect sizes that favor higher performance on questions associated with longer passages for several subgroups: small effects for medium prior literacy achievement students in total scores and on the history/social studies questions as well as for low prior literacy achievement students on the humanities questions; a medium effect for high prior literacy achievement students in total scores; and a large effect for high prior literacy achievement students on the history/social studies questions. No meaningful effects of passage length were detected in any other subject area or literacy achievement subgroup except for a medium effect favoring higher performance for low prior literacy achievement students on the literature questions associated with the shorter passage. See Appendix table A4 for additional details.

Discussion

The research literature on optimal reading passage length for comprehension remains inconclusive (College Board, 2024b).

While some studies have found that longer texts elicit qualitatively different comprehension responses and that shorter texts promote deeper processing and better retention, others have found no significant effect of passage length on comprehension across varying populations, question types, and text lengths.[7] The lack of consensus, as noted by Alderson (2000) and illustrated by the range of approaches from Derrick (1953) to Jalilehvand (2012), likely stems from wide variation in how studies define "short" and "long" as well as differences in participant age, ability level, and assessment methods, suggesting that optimal passage length is context dependent rather than absolute.

The present report shares two empirical approaches to understanding the relationship between reading passage length on the SAT Suite of Assessments and student performance in reading and writing. Both observational data and data collected through a carefully crafted experimental research design point to similar conclusions that affirm prior research showing no meaningful relationship between passage length and performance. If reading passage length affected SAT RW performance by reducing rigor, the paper-to-digital score correlations would diverge from the paper-to-paper baseline correlations, and experimental performance on shorter-passage questions would generally outpace performance on longer-passage questions. If anything, students performed *better* or similarly on questions associated with longer passages than with shorter passages both in overall performance and across performance in history/social studies, the humanities, and science. Among the four literacy subject areas, only literature showed a significant advantage for shorter passages, though the full-sample effect size was negligible; disaggregation by achievement level revealed a medium, meaningful effect specific to low achievement students. These results suggest that claims of decreased rigor in the SAT related to the use of shorter reading passages are not grounded in data or evidence.

[7] Studies by Mehrpour and Riazi (2004), Jalilehvand (2012), Bae and Lee (2018), and Grant (1980) found no significant effect of passage length on comprehension across varying populations and question types. Conversely, Newsom and Gaite (1971), Rothkopf and Billington (1983), Mesmer and Hiebert (2015), and Surber (1992) found negative associations between passage length and comprehension or retention, with Surber suggesting shorter texts may induce greater depth of processing.

References

Alderson, J. Charles. 2000. *Assessing Reading*. Cambridge University Press.

Bae, Minryoung, and Byungmin Lee. 2018. "Effects of Text Length and Question Type on Test-takers' Performance on Fill-in-the-Blank Items on Korean CSAT." *English Teaching* 73, no. 4 (Winter): 149–74. https://files.eric.ed.gov/fulltext/EJ1306306.pdf.

Bloomberg Editorial Board. "Dumbing Down the SAT Bodes Poorly for Education." *Bloomberg,* September 3, 2025.

College Board. 2023. *Digital SAT Research Summary* (September 2023). College Board. https://research.collegeboard.org/media/pdf/Digital%20SAT%20Research%20Summary.pdf

College Board. 2024a. *Digital SAT Suite of Assessments Technical Manual*. College Board. https://research.collegeboard.org/media/pdf/Digital%20SAT%20Suite%20of%20Assessments%20Technical%20Manual-FINAL.pdf

College Board. 2024b. *Assessment Framework for the Digital SAT Suite*, version 3.01 (August 2024). College Board. https://satsuite.collegeboard.org/media/pdf/assessment-framework-for-digital-sat-suite.pdf.

College Board. 2024c. *The Cognitively Complex Thinking Required by Select Digital SAT Suite Questions.* College Board. https://satsuite.collegeboard.org/media/pdf/digital-sat-cognitive-lab-report.pdf.

College Board. 2025a. *The Cognitively Complex Thinking Required by Select SAT Suite Questions: Evidence from Students with Specific Learning Disorders Affecting Reading (Dyslexia).* College Board. https://satsuite.collegeboard.org/media/pdf/digital-sat-cognitive-lab-report-sldr.pdf.

College Board. 2025b. *The Cognitively Complex Thinking Required by Select SAT Suite Questions: Evidence from Students with Attention Deficit Hyperactivity Disorder (ADHD).* College Board. https://satsuite.collegeboard.org/media/pdf/digital-sat-cognitive-lab-report-adhd.pdf.

College Board. 2025c. *The Cognitively Complex Thinking Required by Select SAT Suite Questions: Evidence from English Learners (ELs).* College Board. https://satsuite.collegeboard.org/media/pdf/digital-sat-cognitive-lab-report-el.pdf.

REFERENCES

Derrick, Clarence. 1953. "Three Aspects of Reading Comprehension as Measured by Tests of Different Lengths." PhD diss., University of Chicago. ETS Research Bulletin RB-53-08. https://doi.org/10.1002/j.2333-8504.1953.tb00039.x.

Finley, Allysia. 2025. "The College Students Who Can't Do Elementary Math." *Wall Street Journal*. November 30, 2025.

Grant, Norma Louise. 1980. "The Effects of Passage Length and Reading Mode on Retellings and Question Responses of Sixth Grade Children." EdD diss., University of Arizona. https://repository.arizona.edu/handle/10150/282204.

Jalilehvand, Maryam. 2012. "The Effects of Text Length and Picture on Reading Comprehension of Iranian EFL Students." *Asian Social Science* 8, no. 3: 329–37. https://doi.org/10.5539/ass.v8n3p329.

Mehrpour, Saeed, and Abdolmehdi Riazi. 2004. "The Impact of Text Length on EFL Students' Reading Comprehension." *Asian EFL Journal* 6, no. 3: 1–13.

Mesmer, Heidi Anne, and Elfrieda H. Hiebert. 2015. "Third Graders' Reading Proficiency Reading Texts Varying in Complexity and Length: Responses of Students in an Urban, High-Needs School." *Journal of Literacy Research* 47, no. 4: 473–504. https://journals.sagepub.com/doi/pdf/10.1177/1086296X16631923.

Newsom, R. S., and A. J. H. Gaite. 1971. "Prose Learning: Effects of Pretesting and Reduction of Passage Length." *Psychological Reports* 28, no. 1: 123–9. https://doi.org/10.2466/pr0.1971.28.1.123.

Rothkopf, Ernst Z., and M. J. Billington. 1983. "Passage Length and Recall with Text Size Held Constant: Effects of Modality, Pacing, and Learning Set." *Journal of Verbal Learning and Verbal Behavior* 22, no. 6: 667–81. https://psycnet.apa.org/doi/10.1016/S0022-5371(83)90395-X.

Surber, John R. 1992. "The Effect of Test Expectation, Subject Matter, and Passage Length on Study Tactics and Retention." *Reading Research and Instruction* 31, no. 3: 32–40. https://doi.org/10.1080/19388079209558086.

Appendix

I. Study Recruitment

Through a customized email solicitation in January 2026, College Board reached out to approximately 50,000 students who had previously taken the SAT, PSAT/NMSQT, and/or PSAT 10 tests and had opted in to being contacted by College Board for research opportunities and other purposes. The goal was to obtain a heterogeneous student sample of 300–500 high school juniors and seniors for the study. In addition to heterogeneity in terms of gender, race/ethnicity, geographic location, and other demographic characteristics, the sample was intended to be divisible into three similarly sized strata by prior literacy achievement on SAT Suite RW section scores. Anticipating that lower-achievement students may be less likely than their peers to participate in a study of their reading comprehension skills and knowledge, these students were overrepresented in recruitment efforts.

Subject: You're Invited: Join Our Reading Research Study and Earn a $50 Gift Card

Dear [firstname],

You're invited to take part in a short online research study to help us improve how we measure reading comprehension on future College Board assessments. As a thank you for your time and effort, you'll receive a **$50 gift card** for completing the study. This invitation is personalized—please don't forward it.

Study Details

- **Date:** Thursday, February 5 at 6 p.m. local time
- **Format:** Online through Bluebook™, College Board's digital testing app
- **Time commitment:** About 45 minutes
- **Device requirements:** You'll need a device with Bluebook already installed. You can use a Windows laptop or tablet, Mac laptop, iPad, or a school-managed Chromebook.
- **Personal Chromebooks aren't permitted.** Confirm that your device is supported by reviewing our approved devices page.

- **What you'll do:** You'll read a few short passages on different topics and answer a small set of reading comprehension questions. You may see passage formats or question types that differ from those used on current SAT or AP Exams. Your responses will help us enhance future assessments.
- **Incentive requirements:** To receive the $50 gift card, you'll need to complete all activities, which include downloading and setting up Bluebook in advance, submitting your responses, and staying for the full duration of the study. You may skip questions or stop at any time, but incomplete sessions won't qualify for the $50 gift card.
- **Confidentiality:** All information you provide will be used only for research. We may share overall findings publicly, but your individual responses will remain confidential and will never affect any of your past or future College Board scores.
- **Scores:** You will not receive any scores as part of this study.

Button: [Register Now]

Sign up by **January 23, 2026**. If you're selected to participate, we'll notify you via email by **January 29**.

If you have questions, email us at CBdigitalsupport@collegeboard.org or call us at 877-348-5728, Monday–Friday, 8 a.m.–8 p.m. ET

Thank you for considering this opportunity to support our research!

Sincerely,
College Board

II. Study Consent Form

Before being eligible to participate in the study, all participants (if age 18 or over) or a parent/guardian (if the participant was under 18) had to sign a consent form. This form described the nature of the study, the research purposes for which the study's data were being collected, the privacy protection participants would receive (i.e., their data would only be analyzed in the aggregate and would not be associated in reporting with personally identifiable information), the nature and amount of compensation participants would receive on successful completion of the activity, and the fact that participants could opt out of the study or any part of the activity at any time (but that successful completion was required to receive the compensation). Participants were also told that they would not receive official SAT scores as part of participating in the activity.

CollegeBoard

Student Research Group Agreement

By signing this agreement, the student identified below ("**Student**"), with consent of their parent/guardian ("**Parent/Guardian**"), agrees to Student's participation in Reading Comprehension Study, a research study for College Board ("**Study**"). The Study involves the Student providing feedback to College Board on SAT-style questions, including answering those questions to the best of their ability. The study will be conducted entirely online. The activity will take no more than forty-five (45) minutes, including setup, and on successful completion of the activity, payment will be made via digital payment platform Tremendous. Student will receive a link from Tremendous to the email address provided; this link can be used to redeem payment in the form of a bank transfer, PayPal deposit, or a gift card of choice—Tremendous has over 300 gift card options.

Student and Parent/Guardian grant College Board and its designees, affiliates, agents, subcontractors, and licensees (collectively, "**College Board**") the right to use, transcribe, edit, reproduce, broadcast, publish, exhibit, publicize, and otherwise distribute, without compensation to Student and Parent/Guardian, any Student responses, statements, and comments Student makes during or in connection with the Study ("**Information**"). The rights hereby granted to College Board are perpetual and worldwide.

Student and Parent/Guardian acknowledge that College Board will rely on this permission and that College Board, in its sole discretion, may decide whether or not to use the Information. Student and Parent/Guardian will not assert a claim that the use of the Information is a violation of Student rights. Student and Parent/Guardian further understand and agree that they hereby waive all rights and claims to ownership of the College Board materials in which the Information may appear.

The Study session will be conducted entirely online and will not be recorded. The session will occur at a time of Student's choosing within the day and time frame specified by the provided Study materials. Participation in the Study does require that Student download, install, and run Bluebook, College Board's digital testing application, on a desktop or laptop computer, iOS-based tablet, or Chromebook and that Student have at least some internet access before, during, and immediately following the activity. Student may be sent, via email, a brief survey seeking feedback on the activity; completion of this survey is voluntary (i.e., not required to receive the indicated compensation), but doing so is greatly appreciated.

In addition, Student and Parent/Guardian acknowledge that any information and materials that are disclosed or otherwise made available to Student and Parent/Guardian in connection with the Study ("**Confidential Information**") are highly confidential and proprietary to College Board and agree (i) to keep them strictly confidential, (ii) not to disclose to or discuss with any third party, and (iii) not to use for any purpose other than to participate in the Study.

Student and Parent/Guardian understand that College Board is offering to pay Student based on the research activity a US $50 gift card, provided that such payment is permissible under applicable laws and regulations, and the policies and regulations

of Student's or Parent/Guardian's employer, if any. Student and Parent/Guardian acknowledge and agree that College Board is not responsible for determining whether Student and/or Parent/Guardian institution's policies and regulations or applicable laws and regulations preclude Student from participating in the Study or receiving such payment. Student and Parent/Guardian will not consider this agreement an offer to provide this payment if Student and/or Parent/Guardian is prohibited from accepting such payment.

This Student Research Group Agreement is the full and complete understanding between College Board, Student, and Parent/Guardian. Student and Parent/Guardian each represent they have had adequate time to read this document carefully and to ask any questions that they may have.

Please Print:

Name of Participant	Signature	Date
Name of Parent/Guardian	Signature	Date
Student Street Address, City, State		
Student Email address		

III. Simulated Test Materials

College Board developed four pairs of stimulus passages, with each pair consisting of passages of highly similar content but differing lengths, "shorter" and "longer."

The shorter passage in each pair consisted of an actual digital SAT Suite passage taken from the practice environment (but not from an intact practice test) and, consistent with the digital test specifications, was up to 150 (standard, six-character) words in length. Collectively, these passages sampled from all four subject areas covered by the digital (and paper-based) SAT literacy section—literature, history/social studies, the humanities, and science—and represented the three text complexity levels called for in digital RW stimuli: grades 6–8, grades 9–11, and grades 12–14. (A highly similar range was used for paper and pencil ERW stimuli during that test's operational life.)

The longer passage in each pair was then developed in one of two ways, depending on the passage's subject area. The longer literature passage was developed by appending additional contiguous text from the original source text to the shorter passage. The longer history/social studies, humanities, and science passages, by contrast, were initially created using a generative AI tool (ChatGPT 5), with the tool being prompted to expand the length of the original (shorter) passage without substantially changing its content, text complexity, or other features and while ensuring that all information and ideas in the shorter version were carried over to the longer version in some form; overlap mappings were also produced to make it easier to verify that the longer passage contained all the information in the shorter passage. College Board subject matter experts then vetted the generated content for comparability and made manual edits as

necessary to ensure content accuracy and bring the two texts more closely in line with each other on the desired characteristics. In all cases, the text common to both versions ("shorter" and "longer") is **boldface** in the "longer" passage version. Passage word counts are standardized and based on every six characters (letters, numbers, spaces, punctuation) constituting a "word."

The associated multiple-choice test questions are meant to be closely parallel in terms of skill/knowledge testing point and estimated difficulty (easy, medium, hard) such that -A1 and -B1 as well as -A2 and -B2 are highly similar but assess participants' understanding of differing details, meaning that answering either the -A or -B questions first shouldn't clue the answers to the other, parallel questions. The keyed response (*key*) to each question is marked with an asterisk (*).

The passages and questions displayed below for literature, history/social studies, humanities, and science are only approximations of their appearance in College Board's Bluebook digital testing app. (The included metadata elements [e.g., key, skill/knowledge element] weren't presented to participants.)

LITERATURE

Literature – Shorter Version

The following text is from Anthony Trollope's 1855 novel *The Warden*. The narrator is describing a rectory, the residence of a clergyperson.

> There was an air of heaviness about the rooms which might have been avoided without any sacrifice of propriety; colours might have been better chosen and lights more perfectly diffused; but perhaps in doing so the thorough clerical aspect of the whole might have been somewhat marred; at any rate, it was not without ample consideration that those thick, dark, costly carpets were put down; those embossed, but sombre [wallpapers] hung up; those heavy curtains draped so as to half exclude the light of the sun.

Text Complexity Level: Grades 12–14

Standardized Word Count: 85

LIT-A1. How does the narrator describe the rectory's carpets?

A. Thick, dark, and expensive*

B. Old, stained, and cheap

C. New but poorly made

D. Pretty but too delicate

Skill/Knowledge Element: Central Ideas and Details (explicit information)

Estimated Difficulty: Easy

LIT-A2. What does the text most strongly suggest about the rooms in the rectory?

A. They have many qualities the narrator finds pleasing, but they lack useful features.

B. They have been furnished with the intention of maintaining a sense of appropriateness.*

C. They are decorated in a style that to the narrator seems overly ornate.

D. They are meant to feel spacious but are in reality too small.

Skill/Knowledge Element: Central Ideas and Details (implicit information)

Estimated Difficulty: Hard

Literature – Longer Version

The following text is from Anthony Trollope's 1855 novel *The Warden*. The narrator is describing a rectory, the residence of a clergyperson.

> Dr Grantly was blessed with a happy, thriving family. There were, first, three boys, now at home from school for the holidays. They were called, respectively, Charles James, Henry, and Samuel. The two younger (there were five in all) were girls; the elder, Florinda, bore the name of the Archbishop of York's wife, whose godchild she was: and the younger had been christened Grizzel, after a sister of the Archbishop of Canterbury. The boys were all clever, and gave good promise of being well able to meet the cares and trials of the world; and yet they were not alike in their dispositions, and each had his individual character, and each his separate admirers among the doctor's friends. . . .

The two little girls Florinda and Grizzel were nice little girls enough, but they did not possess the strong sterling qualities of their brothers; their voices were not often heard at Plumstead Episcopi; they were bashful and timid by nature, slow to speak before company even when asked to do so; and though they looked very nice in their clean white muslin frocks and pink sashes, they were but little noticed by the archdeacon's visitors. . . .

And now let us observe the well-furnished breakfast-parlour at Plumstead Episcopi, and the comfortable air of all the belongings of the rectory. Comfortable they certainly were, but neither gorgeous nor even grand; indeed, considering the money that had been spent there, the eye and taste might have been better served; **there was an air of heaviness about the rooms which might have been avoided without any sacrifice of propriety; colours might have been better chosen and lights more perfectly diffused; but perhaps in doing so the thorough clerical aspect of the whole might have been somewhat marred; at any rate, it was not without ample consideration that those thick, dark, costly carpets were put down; those embossed, but sombre [wallpapers] hung up; those heavy curtains draped so as to half exclude the light of the sun:** nor were these old-fashioned chairs, bought at a price far exceeding that now given for more modern goods, without a purpose. The breakfast-service on the table was equally costly and equally plain; the apparent object had been to spend money without obtaining brilliancy or splendour. The urn was of thick and solid silver, as were also the tea-pot, coffee-pot, cream-ewer, and sugar-bowl; the cups were old, dim dragon china, worth about a pound a piece, but very despicable in the eyes of the uninitiated. The silver forks were so heavy as to be disagreeable to the hand, and the bread-basket was of a weight really formidable to any but robust persons. The tea consumed was the very best, the coffee the very blackest, the cream the very thickest; there was dry toast and buttered toast, muffins and crumpets; hot bread and cold bread, white bread and brown bread, home-made bread and bakers' bread, wheaten bread and oaten bread; and if there be other breads than these, they were there; there were eggs in napkins, and crispy bits of bacon under silver covers; and there were little fishes in a little box, and devilled kidneys frizzling on a hot-water dish; which, by the bye, were placed closely contiguous to the plate of the worthy archdeacon himself. Over and above this, on a snow-white napkin, spread upon the sideboard, was a huge ham and a huge sirloin; the latter having laden the dinner table on the previous evening. Such was the ordinary fare at Plumstead Episcopi.

And yet I have never found the rectory a pleasant house. The fact that man shall not live by bread alone seemed to be somewhat forgotten; and noble as was the appearance of the host, and sweet and good-natured as was the face of the hostess, talented as were the children, and excellent as were the viands and the wines, in spite of these attractions, I generally found the rectory somewhat dull. After breakfast the archdeacon would retire, of course to his clerical pursuits. Mrs Grantly, I presume, inspected her kitchen, though she had a first-rate housekeeper, with sixty pounds a year; and attended to the lessons of Florinda and Grizzel, though she had an excellent governess with thirty pounds a year: but at any rate she disappeared: and I never could make companions of the boys. . . .

On the whole, therefore, I found the rectory a dull house, though it must be admitted that everything there was of the very best.

Text Complexity Level: Grades 12–14

Standardized Word Count: 731

LIT-B1. How does the narrator describe the rectory's curtains?

A. Worn and slightly tattered

B. Colorful and cheerful

C. Costly and fancy

D. Heavy and light limiting*

Skill/Knowledge Element: Central Ideas and Details (explicit information)

Estimated Difficulty: Easy

LIT-B2. What does the text most strongly suggest about the furnishings in the rooms in the rectory?

A. They represent the owner's profession in a way that the narrator finds extravagant but oddly charming.

B. They signal the owner's high social status in a manner that the narrator views as unsurprising but distasteful.

C. They reflect deliberate choices that the narrator understands but doesn't necessarily agree with.*

D. They embody compromises that the narrator initially objects to but ultimately recognizes were inevitable.

Skill/Knowledge Element: Central Ideas and Details (implicit information)

Estimated Difficulty: Hard

HISTORY/SOCIAL STUDIES

History/Social Studies – Shorter Version

Maryam Hosseini and her team used aerial images of well-mapped areas of Boston to train a computer program to identify sidewalks. When it was tested on images of Washington, DC, it reliably identified sidewalks in the vast majority of cases and even whether sidewalks were cobblestone or brick. Hosseini believes the program will improve on this already-strong performance as it gets used more.

Text Complexity Level: Grades 6–8

Standardized Word Count: 66

HSS-A1. Based on the text, Hosseini would most likely recommend that she and her team take which step to improve their computer program?

A. They perform testing in additional cities.

B. They obtain new aerial images of cities.

C. They thoroughly rewrite the program.

D. They collect data from additional users.*

Skill/Knowledge Element: Central Ideas and Details (explicit information)

Estimated Difficulty: Medium

HSS-A2. Which choice best describes the function of the underlined phrase in the text as a whole?

A. It emphasizes the effectiveness of the computer program Hosseini and her team developed.*

B. It names the central problem that the computer program Hosseini and her team designed was meant to solve.

C. It reveals what the computer program Hosseini and her team produced discovered about Washington, DC.

D. It describes a feature that Hosseini and her team intend to add to the computer program they made.

Skill/Knowledge Element: Text Structure and Purpose (part-whole relationships)

Estimated Difficulty: Medium

History/Social Studies – Longer Version

Maryam Hosseini and her team used aerial images of well-mapped areas of Boston to train a computer program to identify sidewalks. Their goal addressed a basic data gap: most U.S. cities still lack complete, up-to-date sidewalk maps, which makes it hard for planners to evaluate walkability or design accessible routes. The group's open-source tool—developed across MIT, New York University, and other collaborating institutions—learns to recognize pedestrian infrastructure and then converts those detections into sidewalk polygons and a topologically connected network that routing engines can use. Training began with thousands of images from places where sidewalk maps already existed so the model could learn what sidewalks, crosswalks, and footpaths look like from above.

The training data did include Boston, but the team also drew on nearby Cambridge as well as New York City and Washington, DC. Using areas that were already well mapped let the researchers evaluate how much of the pedestrian network their model identified. In internal tests, the system recognized about nine-tenths of sidewalks and crosswalks in Boston and Cambridge, a level of performance strong enough to be useful for planning, accessibility audits, and gap-finding. Because the approach is image based, it can be applied anywhere suitable aerial imagery exists.

When it was tested on images of Washington, DC, it reliably identified sidewalks in the vast majority of cases. The team reports similarly high accuracy when implementing the framework across Boston, Cambridge, New York, and Washington. The model is designed to be adaptable: Coding used in the process is openly available, and cities can run the same application on their own imagery to create a first-cut sidewalk layer, and then refine it with local knowledge. This means the tool can scale from neighborhoods to whole metropolitan regions without the need for custom, city-by-city reengineering.

The project does more than draw lines on a map. **Even whether sidewalks are cobblestone or brick** can be inferred when the sidewalk network is linked to a complementary model that classifies surface materials from widely available street-level photographs. That companion effort, built by members of the same research collaboration, tags each sidewalk segment with a material label (for example, brick, cobblestone, concrete), information that matters for wheelchair users, people pushing strollers, and anyone sensitive to vibration or uneven surfaces. In Washington, DC, the team has combined these material tags with curb-ramp and crossing information to calculate segment-by-segment accessibility scores, demonstrating how the mapped network supports richer analyses.

Hosseini believes the program will improve on this already strong performance as it gets used more. That confidence reflects both the project's open-source strategy and the fact that these sorts of applications tend to improve when additional examples are used in training. As more cities run the application and contribute corrections, the model's ability to generalize should advance, and the resulting sidewalk data can be kept current as aerial imagery is refreshed. In practical terms, that means planners and advocates can move from small-scale, hand-drawn sidewalk lines toward consistent, city-scale pedestrian maps that help answer simple questions—Can you walk to that bus stop safely? Is there a continuous route to the clinic?—with much greater certainty.

Text Complexity Level: Grades 6–8

Standardized Word Count: 580

HSS-B1. The text indicates that before they could train their computer program, Hosseini and her colleagues first had to do which of the following?

A. Identify materials commonly used to make sidewalks

B. Learn the techniques involved in taking aerial images of sidewalks

C. Find a city for which sidewalk locations were already known*

D. Create a map of the sidewalks of Washington, DC

Skill/Knowledge Element: Central Ideas and Details (explicit information)

Estimated Difficulty: Medium

HSS-B2. Which choice best describes the function of the underlined sentence in the text as a whole?

A. It presents an early finding that Hosseini and her team later revised.

B. It suggests that a test of the computer program was successful.*

C. It calls out an important limitation of the computer program.

D. It implies that Hosseini and her team were surprised by their findings.

Skill/Knowledge Element: Text Structure and Purpose (part-whole relationships)

Estimated Difficulty: Medium

HUMANITIES

Humanities – Shorter Version

Speech stenography (the real-time written recording of spoken words using quickly written symbols) has existed since antiquity. But musical stenography is a more recent development, described in treatises mainly published between 1768 and 1950 (such as Peter Kabala's 1940 work *Musical Shorthand*) before its techniques were supplanted by phonograph recording. Some methods of musical stenography repurposed techniques used in speech stenography in an attempt to gain advantages over conventional Western musical notation. In particular, just as the squiggles and dots of speech stenography would allow stenographers to rapidly document deviations from prepared speeches, it was thought that musical stenography would be able to capture spontaneous musical improvisations that differed from printed musical scores and would otherwise go unrecorded.

Text Complexity Level: Grades 9–11

Standardized Word Count: 141

HUM-B1. What point does the text make about speech stenography?

A. It was invented after musical stenography.

B. It has existed since ancient times.*

C. It was largely replaced by phonograph recording.

D. It developed most rapidly between 1768 and 1950.

Skill/Knowledge Element: Central Ideas and Details (explicit information)

Estimated Difficulty: Medium

HUM-B2. As used in the text, what does the word "capture" most nearly mean?

A. Preserve*

B. Earn

C. Accomplish

D. Accumulate

Skill/Knowledge Element: Words in Context

Estimated Difficulty: Medium

Humanities – Longer Version

Speech stenography (the real-time written recording of spoken words using quickly written symbols) has existed since antiquity. In the Roman world, secretaries and court reporters developed dense repertoires of abbreviated signs so they could write nearly as fast as people spoke, and medieval copyists preserved and elaborated those systems. By the eighteenth and nineteenth centuries, shorthand had become a mature craft, with textbooks, public demonstrations, and professional guilds. Its promise was simple and powerful: transform fleeting sound into durable script at the pace of performance.

But musical stenography is a more recent development, described in treatises mainly published between 1768 and 1950 (such as Peter Kabala's 1940 work *Musical Shorthand*) before its techniques were supplanted by phonograph recording. Writers across Europe and the Americas proposed "rapid" or "tachygraphic" notations that tried to do for melody and rhythm what shorthand did for speech. These inventors replaced the familiar noteheads, stems, and beams of the five-line musical staff with squiggles, lines, and dots drawn for speed. Nineteenth-century pamphlets and small books—some written by career stenographers, others by musicians with a technical bent—claimed their signs could be learned quickly, allow on-the-spot notation of complex passages, and be transcribed later into conventional musical scores.

Some methods of musical stenography repurposed techniques used in speech stenography in an attempt to gain advantages over conventional Western musical notation. Parliamentary stenographers such as Hippolyte Prévost designed systems that leaned on the ergonomics of fast pen strokes; other authors borrowed the rhetoric of shorthand manuals, promising "systems" that reduced the number of distinct marks, grouped related gestures, and exploited muscle memory. Treatises by figures such as Adalbert de Rambures presented side-by-side comparisons intended to show how a few compact signs might stand in for longer strings of staff notation. In the twentieth century, small-format booklets continued the genre. **Just as the squiggles and dots of speech stenography would allow stenographers to rapidly document deviations from prepared speeches, it was thought that musical stenography would be able to capture spontaneous musical improvisations that differed from printed musical scores and would otherwise go unrecorded.**

The rise of mechanical sound recording altered this landscape. From the late nineteenth century onward, the phonograph and its successors made it possible to preserve sound vibrations themselves rather than infer them through symbols. As discs and tapes became cheaper and more common, the case for text-based "rapid" musical capture weakened. Music stenography never achieved wide adoption, and by the middle of the twentieth century its manuals had largely retreated to libraries and private collections. Yet the impulse behind them—recovering improvisation, the unrepeatable flourish, the wayward tempo—remains a throughline in how musicians think about preserving performance. In that sense, these systems were not dead ends so much as provocative reminders that notation is always chasing sound.

Text Complexity Level: Grades 9–11

Standardized Word Count: 537

HUM-A1. How does the text characterize Peter Kabala's *Musical Shorthand*?

A. As a highly influential book on Western musical notation

B. As an ancient treatise on the rapid transcription of music

C. As an example of a work about musical stenography*

D. As a nearly forgotten pamphlet on twentieth-century music

Skill/Knowledge Element: Central Ideas and Details (explicit information)

Estimated Difficulty: Medium

HUM-A2. As used in the third paragraph, what does the word "conventional" most nearly mean?

A. Polite

B. Ceremonial

C. Trite

D. Standard*

Skill/Knowledge Element: Words in Context

Estimated Difficulty: Medium

SCIENCE

Science – Shorter Version

In an international collaboration, Elaine Ostrander, Bridgett M. vonHoldt, and other researchers probed the evolutionary history of size variation in modern dogs. Scientific consensus held that early dogs had large body mass and that a genetic driver of smaller size in some breeds (e.g., miniature schnauzers) developed only within the last 20,000 years as a result of selective breeding for characteristics favored by humans. Ostrander et al. assert that this explanation is flawed, having discovered that a mutation responsible for variants of *IGF1*, a gene found in many mammals that regulates production of insulin-like growth factor 1, is ubiquitous in domestic dog breeds.

Text Complexity Level: Grades 9–11

Standardized Word Count: 113

SCI-B1. Which finding, if true, would most directly support the researchers' assertion?

A. An additional mutation related to *IGF1* affects the development of characteristics other than body size in smaller dog breeds like miniature schnauzers.

B. The mutation related to *IGF1* influences body size and is found in 53,000-year-old genetic material from fossils of Siberian wolves (*Canis lupus campestris*), an ancestor of dogs.*

C. One variant of *IGF1* is found only in smaller dog breeds like miniature schnauzers and appears to have emerged no more than 20,000 years ago.

D. *IGF1* has been isolated in genetic material from fossils more than 20,000 years old of the red wolf (*Canis rufus*) and certain other species related to dogs.

Skill/Knowledge Element: Command of Evidence: Textual

Estimated Difficulty: Hard

SCI-B2. As used in the text, what does the word "held" most nearly mean?

A. Gripped

B. Retained

C. Believed*

D. Contained

Skill/Knowledge Element: Words in Context

Estimated Difficulty: Easy

Science – Longer Version

In an international collaboration, Elaine Ostrander, Bridgett M. vonHoldt, and other researchers probed the evolutionary history of size variation in modern dogs. Their focus was a question that cuts across archaeology, genetics, and breed history: Why do dogs span an unparalleled range of body sizes, and when did the genetic levers that make teacup breeds and giant mastiffs possible first arise? The team assembled genomes from both modern breeds and ancient canids and focused on candidate regions long suspected to modulate growth.

Scientific consensus held that early dogs had large body mass and that a genetic driver of smaller size in some breeds (e.g., miniature schnauzers) developed only within the last 20,000 years as a result of selective breeding for characteristics favored by humans. That narrative fit with the visible history of breed formation over the past few centuries and with the intuition that "toy" size emerged late as people selected for companionable smallness. It also matched an older research literature in which many size genes of modest effect were cataloged in breeds, while the precise, causative variant at the heart of the most important pathway remained elusive.

Ostrander et al. assert that this explanation is flawed, having discovered that a mutation responsible for variants of *IGF1*, a gene found in many mammals that regulates production of insulin-like growth factor 1, is ubiquitous in domestic dog breeds. Their analysis points to a single change near the *IGF1* gene that correlates with body mass across dogs and even across wild canids. The mutation occurs in two versions (alleles). Dogs carrying two copies of one allele tend to be smaller on average, those with two copies of the other tend to be larger, and mixed genotypes tend to fall between. Crucially, the small size–associated allele is not a recent invention: It appears widely in present-day dogs and is evident in ancient specimens.

The team's data suggest that rather than a "new" variant that arose only after domestication, a preexisting, evolutionarily older variant was later amplified in small breeds by human selection. That revision in thinking matters for how we read the archaeological record of dogs, how we model the early phases of domestication, and how we parse the relative roles of long-standing variation versus subsequent mutation in producing dramatic phenotypes. It also clarifies why earlier scans repeatedly flagged *IGF1* as a major contributor to size.

The larger program of research does not deny that other genes shape height and mass (many do, and some are breed specific), but it elevates one pathway—the *IGF1* gene—to a central place in the story. It also shows how international data sharing, ancient DNA, and targeted research can converge on a single, testable conclusion. For students of domestication, the upshot is a more continuous timeline, one in which modern breeds exploit variation that predates modern breeding and, in all likelihood, domestication itself.

Text Complexity Level: Grades 9–11

Standardized Word Count: 501

SCI-A1. Which finding, if true, would most directly support the scientific consensus about early dogs' size?

A. Dog fossils from 100,000 years ago showed a consistent pattern of dogs of various types having had large masses.*

B. Dog fossils from 50,000 years ago showed clear signs of significant size variations across dog types.

C. Dog fossils from 20,000 years ago showed no significant size variations across dog types.

D. Dog fossils from 10,000 years ago showed a consistent pattern of dogs of various types having had small masses.

Skill/Knowledge Element: Command of Evidence: Textual

Estimated Difficulty: Hard

SCI-A2. As used in the first paragraph, what does the word "probed" most nearly mean?

A. Prodded

B. Browsed

C. Examined*

D. Found

Skill/Knowledge Element: Words in Context

Estimated Difficulty: Easy

Scoring

College Board scored the study test forms against answer keys established by the lead researcher. For the purposes of this activity, each question was worth exactly one "point" for a correct (keyed) response and zero "points" for an incorrect response; no partial credit was awarded. For study purposes (only), each participant received three scores:

- A total score, ranging from 0 to 8, for all questions associated with the shorter passages
- A total score, again ranging from 0 to 8, for all questions associated with the longer passages
- A series of subject area–specific scores, ranging from 0 to 2, for questions associated with the shorter passage in each area (e.g., literature, shorter)
- A series of subject area–specific scores, ranging from 0 to 2, for questions associated with the longer passage in each area (e.g., literature, longer)

Participation in the study did not yield any official SAT scores for any participants or any scores reported to colleges or universities, nor did participants' involvement in the study affect any of their previously obtained SAT Suite scores.

IV. Model Results Detail in Tabular Form

Table A1 displays the mean number correct and associated standard deviations across all participants (*N* = 367) for shorter- and longer-passage questions both in total and within each subject area.

Table A1. **Descriptive Statistics for Total and Subject Area–Specific Scores, by Reading Passage Length**

	N	Mean Score	SD	Score Values	
				Minimum	Maximum
Total Scores					
Shorter Passages	367	5.43	1.33	0	8
Longer Passages	367	5.89	1.57	0	8
Subject Area–Specific Scores					
Literature, Shorter	367	1.36	0.53	0	2
Literature, Longer	367	1.26	0.62	0	2
History/Social Studies, Shorter	367	1.08	0.45	0	2
History/Social Studies, Longer	367	1.53	0.59	0	2
Humanities, Shorter	367	1.60	0.60	0	2
Humanities, Longer	367	1.71	0.50	0	2
Science, Shorter	367	1.38	0.65	0	2
Science, Longer	367	1.39	0.60	0	2

Table A2 shows the results of paired *t*-tests to statistically compare differences in responses by passage length for total scores and by subject area. Because the values in the "mean difference" column are obtained by subtracting scores on longer-passage

questions from scores on shorter-passage questions, negative numbers indicate participants' longer-passage scores were higher than their shorter-passage scores, while positive numbers indicate the reverse. Statistically significant mean differences are marked with an asterisk (*).

Table A2 also includes standardized mean differences and effect sizes. Using Cohen's[1] guidelines, standardized mean differences (*d*) from |0.20| to |0.49| are considered small effect sizes, standardized mean differences from |0.50| to |0.79| are considered medium effect sizes, and standardized mean differences equal to or greater than |0.80| are considered large effect sizes. Any effect size less than |0.20| is considered a difference of no practical significance. Negative values of *d* indicate that participant scores on questions associated with the longer passage(s) were higher; positive values indicate that participant scores on questions associated with the shorter passage(s) were higher.

Table A2. **Paired t-Tests and Standardized Mean Differences for Total and Subject Area–Specific Scores**

	N	Mean Difference (Shorter – Longer)	SD	95% CI	*t*-statistic	*p*-value	*d*	Effect Size
Total Scores	367	−0.46*	1.48	[−0.61, −0.31]	−6.01	< .001	–0.31	Small
Literature	367	0.10*	0.74	[0.02, 0.17]	2.46	.014	0.13	None
History/ Social Studies	367	−0.44*	0.70	[−0.51, −0.37]	−12.11	< .001	–0.63	Medium
Humanities	367	−0.11*	0.64	[−0.17, −0.04]	−3.27	.001	–0.17	None
Science	367	−0.01	0.79	[−0.09, 0.07]	−0.20	.843	–0.01	None

* Denotes a statistically significant result

Table A3 displays the mean number correct and associated standard deviations disaggregated by participants' prior literacy achievement for shorter- and longer-passage questions both in total and within each subject area.

[1] Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates.

Table A3. **Descriptive Statistics for Total and Subject Area–Specific Scores, by Reading Passage Length and Prior Literacy Achievement Level**

Prior Literacy Achievement Level	Category	*n*	Mean Score	SD	Score Values	
					Minimum	Maximum
Low	**Total Scores**					
	Shorter	24	4.13	1.75	1	7
	Longer	24	3.83	1.99	0	8
	Subject Area–Specific Scores					
	Literature, Shorter	24	1.38	0.65	0	2
	Literature, Longer	24	0.79	0.66	0	2
	History/Social Studies, Shorter	24	0.96	0.69	0	2
	History/Social Studies, Longer	24	1.04	0.69	0	2
	Humanities, Shorter	24	0.92	0.88	0	2
	Humanities, Longer	24	1.17	0.70	0	2
	Science, Shorter	24	0.88	0.74	0	2
	Science, Longer	24	0.83	0.70	0	2
Medium	**Total Scores**					
	Shorter	187	5.17	1.32	0	8
	Longer	187	5.60	1.55	1	8
	Subject Area–Specific Scores					
	Literature, Shorter	187	1.30	0.54	0	2
	Literature, Longer	187	1.22	0.66	0	2
	History/Social Studies, Shorter	187	1.06	0.44	0	2
	History/Social Studies, Longer	187	1.40	0.63	0	2
	Humanities, Shorter	187	1.50	0.62	0	2
	Humanities, Longer	187	1.64	0.53	0	2
	Science, Shorter	187	1.30	0.65	0	2
	Science, Longer	187	1.34	0.59	0	2
High	**Total Scores**					
	Shorter	156	5.94	1.03	2	8
	Longer	156	6.56	1.06	3	8
	Subject Area–Specific Scores					
	Literature, Shorter	156	1.42	0.51	0	2
	Literature, Longer	156	1.38	0.52	0	2
	History/Social Studies, Shorter	156	1.13	0.41	0	2
	History/Social Studies, Longer	156	1.75	0.43	1	2
	Humanities, Shorter	156	1.83	0.39	0	2
	Humanities, Longer	156	1.89	0.31	1	2
	Science, Shorter	156	1.56	0.58	0	2
	Science, Longer	156	1.54	0.52	0	2

APPENDIX

Table A4 shows the results of paired *t*-tests to statistically compare differences in responses by passage length and prior literacy achievement for total scores and by subject area. Because the values in the "mean difference" column are obtained by subtracting scores on longer-passage questions from scores on shorter-passage questions, negative numbers indicate participants' longer-passage scores were higher than their shorter-passage scores, while positive numbers indicate the reverse. Statistically significant mean differences are marked with an asterisk (*). Table A4 also includes standardized mean differences and effect sizes.

Table A4. **Paired t-Tests and Standardized Mean Differences for Total and Subject Area–Specific Scores, by Prior Literacy Achievement Level**

	Prior Literacy Achievement Level	*n*	Mean Difference (Shorter − Longer)	SD	95% CI	*t*	*p*-value	*d*	Effect Size
Total Scores	Low	24	0.29	2.14	[−0.61, 1.19]	0.67	.510	0.14	None
	Medium	187	−0.43*	1.62	[−0.66, −0.19]	−3.61	<.001	−0.26	Small
	High	156	−0.62*	1.10	[−0.80, −0.45]	−7.08	<.001	−0.57	Medium
Literature	Low	24	0.58*	0.93	[0.19, 0.98]	3.08	.005	0.63	Medium
	Medium	187	0.08	0.79	[−0.03, 0.19]	1.39	.166	0.10	None
	High	156	0.04	0.62	[−0.06, 0.14]	0.77	.440	0.06	None
History/ Social Studies	Low	24	−0.08	0.88	[−0.46, 0.29]	−0.46	.647	−0.09	None
	Medium	187	−0.34*	0.73	[−0.45, −0.24]	−6.38	<.001	−0.47	Small
	High	156	−0.62*	0.57	[−0.71, −0.52]	−13.41	<.001	−1.07	Large
Humanities	Low	24	−0.25	0.94	[−0.65, 0.15]	−1.30	.207	−0.26	Small
	Medium	187	−0.13*	0.72	[−0.24, −0.03]	−2.55	.012	−0.19	None
	High	156	−0.06	0.46	[−0.13, 0.01]	−1.57	.118	−0.13	None
Science	Low	24	0.04	0.91	[−0.34, 0.43]	0.22	.824	0.05	None
	Medium	187	−0.03	0.84	[−0.15, 0.09]	−0.53	.600	−0.04	None
	High	156	0.01	0.71	[−0.10, 0.13]	0.23	.822	0.02	None

* Denotes a statistically significant result