CollegeBoard | HumRRO HUMAN RESOURCES RESEARCH ORGANIZATION

The Complex Thinking Required by Select SAT® Items: Evidence from Student Cognitive Interviews

CollegeBoard and HumRRO
August 2020

Prepared by College Board and HumRRO

College Board

Jim Patterson, English language arts/literacy, lead author
Steve Boxer, mathematics
Dona Carling, mathematics
Paula Cunningham, psychometric support
Katina Marshall, English language arts/literacy
Andrew Schwartz, mathematics
Bill Trapp, mathematics

HumRRO

Richard Deatz
Sheila R. Schultz
Roland S. Reyes
Anne Woods

About College Board

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

For further information, visit **collegeboard.org**.

About HumRRO

The Human Resources Research Organization (HumRRO) is a results-oriented, non-profit organization with extensive experience in evaluation and assessment in a variety of areas including education, hiring and promotion, recruitment and selection, talent management, career planning, workforce development, and certification. We have just over 150 staff members, the majority of whom have advanced degrees in technical fields, such as educational measurement, assessment, and evaluation or industrial-organizational or quantitative psychology. HumRRO is an industry leader in designing, implementing, and evaluating education and credentialing programs and human capital systems. The thread that ties all these elements of our business together is a passion for and unrelenting focus on high-quality evaluation and assessment. We are an independent organization that has offered customized solutions tailored to our client's unique needs for over 65 years.

For more information, visit **humrro.org**.

© 2020 College Board. College Board and the acorn logo are registered trademarks of College Board. PSAT/NMSQT is a registered trademark of the College Board and National Merit Scholarship Corporation. SAT®, PSAT™ 10, and PSAT™ 8/9 are registered trademarks of College Board. HumRRO and the HumRRO logo are registered trademarks of the Human Resources Research Organization.

Suggested citation: College Board and HumRRO. 2020. *The Complex Thinking Required by Select SAT Items: Evidence from Student Cognitive Interviews*. New York: College Board.

Contents

Tables

Executive Summary

In the spring of 2019, College Board and HumRRO conducted a cognitive interview ("cognitive lab") study of a large cross section of test items from the SAT Reading, Writing and Language, and Math Tests. The study was designed to collect evidence from students of the SAT test–taking age and level of attainment (in this case, high school juniors) with regard to whether the studied items elicited complex cognition in accordance with the items' designs (i.e., their intended constructs). This evidence would serve as one component of College Board's multifaceted, ongoing effort to assess the validity of its flagship college admission test and the associated tests of the SAT Suite of Assessments.

College Board and HumRRO staff collaborated on a study design and implementation of that design, with close attention to rigor and consistency. A sample of ninety-nine cognitive interviews, drawn from ninety-five students across three geographically diverse sites (New York City, New York; Louisville, Kentucky; and Monterey, California) was analyzed in relation to sets of required/expected behaviors defined by College Board assessment experts. College Board staff coded interview transcripts with respect to whether the students offered a verbal demonstration of one or more of the required/expected behaviors associated with each item; these data were then tabulated and used to help determine whether the items were performing as intended to draw out from students complex cognition in accordance with their designs. Vignettes from the transcripts of students who both answered a given item correctly and demonstrated the requisite behavior(s) were selected to illustrate successful performance and to serve as a second data source with respect to the items' functioning. Literacy items were drawn from five types central to the design of the SAT Evidence-Based Reading and Writing (ERW) section: (1) Citing Textual Evidence (Reading Test), (2) Interpreting Words and Phrases in Context (Reading Test), (3) Analyzing Quantitative Information (Reading Test), (4) Development (across four subtypes; Writing and Language), and (5) Effective Language Use (across two subtypes; Writing and Language). Math items were drawn from five broad areas central to the design of the SAT Math Test: (1) Algebra, (2) Functions, (3) Geometry, (4) Ratios, Proportions, and Percentages, and (5) Statistics and Probability.

Two main statistics proved particularly illuminating with regard to the degree of correspondence between student performance and intended constructs: (1) the number and percentage of students who demonstrated all required behaviors (for ERW items) or one or more expected behaviors (for Math items) and (2) the arithmetic difference between the number of students who answered a given item correctly and the number of students who also demonstrated all required (ERW) or one or more expected (Math) behaviors. These statistics provide synoptic insight into whether the items studied performed as intended and thereby elicited complex cognition from students. These statistical measures were supplemented by vignettes from student participants. While necessarily more selective, these vignettes, taken from student verbalizations as they worked through the studied items, offer insight into student thought processes and served as both a check on and confirmation of the broader-based statistical findings.

In the end, both sources of data—statistics and vignettes—converged, offering strong evidence that the SAT items selected for study were readily capable of eliciting from students the sorts of complex behaviors one would expect students having attained college and career readiness to be able to demonstrate. While some caveats are in order, particularly with respect to certain item types and approaches, the study's results, on the whole, offer an important piece of validity evidence in support of the SAT's richness and value.

Introduction

As one part of its continual and multifaceted effort to evaluate and enhance the construct validity of its SAT® Suite of Assessments, College Board, in conjunction with HumRRO, an independent research organization, conducted a cognitive interview ("cognitive lab") study involving several SAT Reading, Writing and Language, and Math item types in the spring of 2019. The main purpose of the lab was to investigate whether these item types, which were chosen for study because of their centrality to key design elements of the SAT, are tapping into the cognitive processes that College Board has claimed for them—in other words, that these items' intended constructs are being enacted in practice by students taking the SAT and that the items are thereby eliciting demonstrations of complex skills and knowledge needed for college and career readiness and success.

Students of SAT test–taking age and educational attainment (specifically, high school juniors) participating in the study were recruited from three sites across the United States (New York City, New York; Louisville, Kentucky; and Monterey, California) to engage in a cognitive interview process as they worked through a series of test items, elucidating for the interviewer their approach to answering each item and the items collectively. Transcriptions of 102 recorded interviews were made, and these data sets were analyzed and coded by College Board content and assessment experts in relation to a set of required behaviors associated with each item.

This report presents two main sources of data from the study. First, for each studied item, tabulations were made of the numbers and percentages of participating students (1) demonstrating each required behavior, (2) demonstrating all required behaviors, (3) answering correctly, and (4) answering correctly while also demonstrating all required behaviors as well as (5) the arithmetic difference between (a) the number of students answering correctly and (b) the number of students answering correctly and also demonstrating all required behaviors. Of these tabulations, (2) and (5) are particularly important as construct validity evidence. For (2), a high number/percentage indicates that many students were enacting an item's intended construct, while a low number/percentage indicates fewer students were enacting the intended construct. For (5), a low number suggests students must demonstrate the required behaviors (i.e., enact the intended construct) to answer a given item correctly, while a high number suggests that students may be able to find a shortcut around the intended construct. Second, illustrative vignettes for select items were drawn from the transcripts of students answering correctly and demonstrating all required behaviors. These vignettes serve to illuminate students' thinking processes and help show the richness of the items' cognitive demands. Together, these results suggest that, in the main, the studied items worked as intended and, as a result, elicited complex cognitive processes from students, thereby offering evidence for the construct validity of the SAT Reading, Writing and Language, and Math Tests and supporting the conclusion that the SAT assesses the kinds and levels of thinking required for postsecondary readiness and success.

Although the present study directly addresses only SAT items, the findings are likely to be broadly applicable to the other tests of the SAT Suite of Assessments: PSAT/NMSQT®, PSAT™ 10, and PSAT™ 8/9. This is because the four tests of the SAT Suite have a common design, with appropriate variations for student age and attainment. That said, cognitive lab study of PSAT/NMSQT, PSAT 10, and PSAT 8/9 items using members of the tests' student populations (high school sophomores and juniors and eighth and ninth graders, respectively) would serve to test this supposition.

Purpose

The 2016 redesign of the nationally recognized SAT college admission test was based on an extensive review of the best available evidence regarding essential college and career readiness and success outcomes. Subsequent evidence, such as that from curriculum survey data (College Board 2019) and feedback from independent subject matter experts tasked with reviewing test items, has validated key design decisions undergirding the test and supported the continued inclusion of particular item types on the Reading, Writing and Language, and Math Tests. To obtain additional evidence, and that of a different character than had been previously obtained, College Board, with significant support from the independent research organization HumRRO, planned and conducted a cognitive lab in the spring of 2019 to assess the performance of a range of items with students of SAT test–taking age and educational attainment (in this case, high school juniors). The primary aim of the study was to use student think-aloud data to determine to what extent the items were eliciting the sorts of complex cognition that the SAT's designers had intended them to elicit—the kinds and levels of cognition required for college and career readiness and success in reading, writing, language, and math.

Evidence for the proposition that the items were measuring what they were intended to measure (i.e., evidence in support of the items' construct validity) would come from students demonstrating behaviors expected by the tests' designers, while evidence against the proposition would come from students failing to demonstrate those behaviors and/or demonstrating behaviors inconsistent with the items' design. For example, the design of Interpreting Words and Phrases in Context vocabulary items found on the SAT Reading Test requires that students read and comprehend the passage context in which tested words or phrases appear and then use passage-based reasoning to answer such items correctly. If, however, students are able to answer these items simply from prior vocabulary knowledge without reference to the associated passage context, the items are eliciting only low-level recall and thus aren't performing as intended, which would prompt College Board to gather more evidence and to consider refinements to the test design.

Methodology

As specified in the *Standards for Educational and Psychological Testing*, when test items are developed, they should be reviewed for clarity, relevance to the construct, and construct-irrelevant content (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014). Although this review requirement is often met by having experts rate each item on designated criteria, experts' ratings don't include a review and analysis of response processes. According to Tourangeau and Rasinski (1988), respondents recall or retrieve relevant information, make a judgment about that information, and select a response—all of which is based on their interpretation or comprehension of the item. Because the respondent's interpretation of an item directly affects test validity and is critical to the inferences that can be made about assessment results, the response process also should be reviewed to ensure students' interpretation is the same as what was intended.

To understand what students were thinking as they engaged with given items (i.e., the students' response processes) and whether, as intended by the constructs, students were able to provide appropriate demonstrations of complex skills and knowledge, College Board and HumRRO implemented a cognitive interview (lab) approach (Leighton 2017; Peterson, Peterson, and Powell 2017; Willis 2005; Ericsson and Simon 1993) to determine the effectiveness of select items relative to their intended constructs. Cognitive interviewing is a multistep, evidence-based qualitative approach that identifies sources of confusion in assessment items and assesses validity evidence based on content and response processes.

In general terms, College Board and HumRRO staff worked together to define formally the intended construct of a given item type (i.e., what the item type was intended to measure and how), gathered data about students' interpretations of and responses to examples of the item type, and compared the students' interpretations and responses to a defined set of behaviors associated with enactment of the intended construct.

More specifically, College Board content and assessment experts in English language arts/literacy and math who are deeply knowledgeable regarding the SAT and its item types formally defined the intended constructs for each of several item types and enumerated required/expected and optional student behaviors associated with each of those types.[1] Mock test forms, including answer sheets, were developed and presented to students, who were prompted to think aloud as they answered each item. Transcripts of the output of the think-aloud portion of the cognitive interviews with students were then analyzed and coded in relation to the required/expected behaviors.[2] The more frequently that students

[1] As explained in more detail in the Results section, below, "required" is used for Evidence-Based Reading and Writing items because the behaviors describe mandatory components associated with answering items correctly and as intended, whereas "expected" is used for Math items because the behaviors describe one or more strategies that students are supposed to demonstrate but that may still result in an incorrect answer (e.g., because of computational error).

[2] Students also answered postexperience interview questions, but these results weren't coded for this study. Interviewer notes were also not analyzed for this report.

demonstrated required/expected behaviors, the stronger the evidence would be that the items were measuring their intended constructs; conversely, the less frequently students demonstrated required/expected behaviors, the weaker such evidence would be.

College Board and HumRRO staff worked collaboratively on the research questions and study design. Using these research questions as a guide, HumRRO developed two separate but related draft cognitive interview protocols, one for the SAT Reading Test and one for the SAT Writing and Language Test; College Board staff developed a parallel protocol for the SAT Math Test. Interviewers were to use these protocols to guide and standardize cognitive interviews. Each protocol included sample probes for the interviewer to use when elaboration or clarification from students was needed. The protocols allowed interviewers to record students' thought processes as well as nonverbal cues (e.g., body language, sighs). To facilitate standard interviewing procedures, HumRRO prepared instructions to accompany each interview protocol. The instructions included a brief introduction that stated the purpose of the interview, provided an overview of the interview and the types of questions that would be asked, and indicated how students' responses would be used. HumRRO provided the draft protocols, along with accompanying instructions, to College Board staff for review and incorporated their feedback into the final protocols. Protocols were internally pilot tested to assess time requirements and clarity of interview instructions. Interviewers received appropriate training in the methodology. Each student was given a monetary incentive—a $100 Amazon gift card—for their time and effort.

HumRRO staff worked with College Board staff to develop a plan to recruit a total of 150 eleventh-grade students. The students were recruited across three geographically diverse samples: New York City, New York; Louisville, Kentucky; and Monterey, California. A sample of fifty students was recruited per location, with a sample of fifty students targeted to be interviewed per content area across locations. To the extent possible, each sample was to include males and females from urban, suburban, and rural settings and reflect student achievement across the SAT score range, as determined from student performance on previously administered College Board assessments (e.g., PSAT/NMSQT, PSAT 10).

HumRRO and College Board collaborated on recruitment approaches and materials. The recruitment method was an announcement of the study in a pop-up window when the College Board website was accessed. The announcement included five questions to determine students' interest in participating in the study and availability as well as their eligibility (i.e., grade, proximity to one of the three study sites, and previous SAT Suite test participation). Students who answered no to any of the questions were thanked for their interest but informed they weren't eligible to participate in the study. Students who answered yes to all the questions were then asked to input their first and last names, their phone number, the name and location of their high school, their College Board username (if they knew it), and their email address. These student-provided data were used to attempt a match to records in the College Board database; demographics and test score history were drawn from this database, and only students whose provided data could be matched to data in the College Board database were deemed eligible, as the study required evidence of their previous test scores.

HumRRO staff worked with College Board staff to develop a script for use when contacting potential participants. The script included the purpose for conducting the interviews, the types of questions students would be asked regarding select SAT items, and how College Board planned to use the information. The script also included confirmation of students' eligibility and availability to participate in the interview as well as gift card reimbursement information. Upon receipt of the lists of eligible participants, HumRRO staff contacted the sampled students through targeted mass emails, phone calls, and text messages. After communication was established and participation confirmed, HumRRO randomly assigned students to participate in the Reading, Writing and Language, or Math cognitive interviews. After initial confirmation, HumRRO requested and collected documentation from students, including a nondisclosure agreement (NDA), a consent form acknowledging voluntary participation in the study, and confirmation of the interview time, date, and location. HumRRO obtained parent or guardian consent for students under eighteen years of age. Student interest was strong in the immediate New York City vicinity but less so in Louisville, Kentucky, and Monterey, California. If interested and available, students in Louisville and Monterey were therefore allowed to participate in interviews for multiple tests.

HumRRO recruited and scheduled a total of 133 students to participate in the cognitive interviews across content areas: 44 Reading interviews, 43 Writing and Language interviews, and 46 Math interviews. A total of 98 students across the three locations participated in 102 interviews.[3] Overall, almost two-thirds (64%) of the students were female, and the majority (88%) attended public schools. Over half (61%) of the students attended schools that were in large cities, followed by schools in the suburbs (17%). Most students were Asian American (37%), followed closely by White (28%) and Black or African American students (20%), and then followed by Hispanic students (6%) and American Indian or Alaska Native students (4%). Compared to the population of high school students who took the SAT operationally in March 2019, the student sample studied here included more females (64% versus 54%), fewer White students (28% versus 53%), more Asian American students (37% versus 12%), more Black or African American students (20% versus 10%), fewer Hispanic students (6% versus 18%), and more American Indian or Alaska Native students (4% versus 0.3%). The sample was also notably higher achieving, as measured by prior SAT Suite test scores, than the March 2019 operational sample.[4]

Table 1 presents the demographic characteristics of the participating ninety-eight students.

[3] One student in Monterey and three students in Kentucky were allowed to complete two interviews each, owing to the lower-than-desired turnout at those sites.

[4] Each student had either a prior SAT or PSAT/NMSQT score. The mean total scaled score for the study sample is 1238.16 (n = 98), while that for the March 2019 SAT group is 1135.05 (n = 298,707). Being juniors, the sampled students hadn't all taken the SAT before the study; for most of them, their most recent score was actually on PSAT/NMSQT. Those for whom the most recent score was the SAT had a very high mean: 1397.08 (n = 24) in comparison to the SAT group in March 2019 (1135.05, above). For the students who had PSAT/NMSQT scores, their mean was 1186.62 (n = 74); this compares with the mean total scaled score on the fall 2018 PSAT/NMSQT of 970.32.

Table 1: Participant Demographic Characteristics

	New York		Louisville		Monterey		Overall[a]	
	n	%	*n*	%	*n*	%	*n*	%
Gender								
Male	17	29	8	47	10	43	35	36
Female	41	71	9	53	13	57	63	64
Total	**58**	**100**	**17**	**100**	**23**	**100**	**98**	**100**
School Type								
Public	50	86	15	88	21	91	86	88
Independent	7	12	2	12	1	4	10	10
Charter	1	2	0	0	1	4	2	2
Total	**58**	**100**	**17**	**100**	**23**	**100**	**98**	**100**
School Area								
Large City	48	83	5	29	7	30	60	61
Medium City	0	0	2	12	3	13	5	5
Small City or Town	1	2	1	6	6	26	8	8
Suburban Area	9	16	4	24	4	17	17	17
Rural Area	0	0	5	29	3	13	8	8
Total	**58**	**100**	**17**	**100**	**23**	**100**	**98**	**100**
Race/Ethnicity								
American Indian or Alaska Native	2	3	1	6	1	4	4	4
Asian American	21	36	4	24	11	48	36	37
Black or African American	16	28	1	6	3	13	20	20
Hispanic	3	5	0	0	3	13	6	6
Native Hawaiian or Other Pacific Islander	2	3	0	0	0	0	2	2
White	13	22	10	59	4	17	27	28
Not Reported	1	2	1	6	1	4	3	3
Total	**58**	**100**	**17**	**100**	**23**	**100**	**98**	**100**

[a] Four students participated in two interviews each.

Of the 133 interviews scheduled, HumRRO and College Board staff conducted 102 cognitive interviews, for an overall participation rate of 77%.[5] The organizations conducted thirty-two Reading interviews (31%), thirty-one Writing and Language interviews (30%), and thirty-nine Math interviews (38%). Of these, three (one for each test) weren't used,[6] resulting in an analyzed sample of ninety-nine transcribed

[5] Some students across the locations were unable to participate in their scheduled interview.

[6] A Reading interview recording from a Louisville student was inaudible. The audio file for a Writing and Language interview with a New York student was lost. A transcript for a Math interview with a New York student failed to be transcribed.

interviews (thirty-one Reading, thirty Writing and Language, and thirty-eight Math) from ninety-five students.

HumRRO staff conducted all Reading and all Writing and Language interviews, while College Board staff conducted all Math interviews. The interviews were conducted during March and April 2019. In addition to interviewers taking notes, all interviews were recorded. A single, trained College Board or HumRRO staff member conducted each cognitive interview using the protocol as a guide to ensure all relevant information was captured about the student's interpretation of each item and their thought process when answering it. In the end, thirty-one Reading, thirty Writing and Language, and thirty-eight Math interviews were analyzed.

The Reading interview consisted of three passages, with a total of twenty-two test items. Students were given sixty-five minutes to answer the Reading items. The Writing and Language interview included four passages, with a total of forty-four test items, but only thirteen of these items were targeted for thinking aloud. Students were given ninety minutes to answer the Writing and Language items. The Math interview included sixteen no-calculator (NC) items and eighteen with-calculator (WC) items. The Math interview was scheduled for one hour, fifty minutes; students were given thirty-five minutes to answer the no-calculator items and forty-five minutes to answer the with-calculator items. Students could take a short break between the no-calculator and with-calculator sessions, if needed.

Following the data collection, HumRRO provided College Board with audio recordings, transcripts, and other records and materials associated with the study. After consultation with HumRRO, College Board elected to perform a preliminary analysis of the data in 2019 and to undertake a full analysis (this report) during the first half of 2020.

College Board content and assessment experts manually analyzed the transcripts of the think-aloud portion of each student's interview to code each item relative to its associated behavior(s). These behaviors were considered "required" for the ERW items and "expected" for the Math items.[7] This difference arose from the nature of the behaviors identified by the College Board teams in the process of defining the items' intended constructs. ERW behaviors were defined as necessary components of answering in the intended fashion, so all behaviors were to be demonstrated by all students. Math behaviors, by contrast, were defined as potential strategies for answering successfully, meaning that students were expected to demonstrate at least one such behavior but not necessarily all the identified behaviors. Staff members also determined each student's answer to each item from the answer document students filled out and then calculated how many and what percent of students answered each item correctly. Note that because the sets of ERW required behaviors included determining the correct answer, students had to answer correctly to demonstrate all required behaviors; this wasn't the case for Math, where students could demonstrate a proper process (i.e., one or more expected behaviors) and still obtain an incorrect answer (e.g., through a computational error). From this effort—which involved examining

[7] Although, as previously noted, the research team identified optional behaviors associated with each item, these weren't analyzed for this report.

approximately 2,300 student-by-item interactions—team members tabulated the following for each item:

1. The number and percentage of students demonstrating each required (ERW) or expected (Math) behavior
2. The number and percentage of students demonstrating all required behaviors (ERW) or one or more expected behaviors (Math)
3. The number and percentage of students answering correctly
4. The number and percentage of students answering correctly while also demonstrating all required behaviors (ERW) or one or more expected behaviors (Math)
5. The arithmetic difference between the number of students answering correctly and the number of students answering correctly and demonstrating all required behaviors (ERW) or one or more expected behaviors (Math)

Of these, (2) and (5) were particularly important analytically in relation to the study's purpose. Calculation (2) serves as a measure of construct validity: when that number/percentage was high, students were deemed to be enacting the item's intended construct. Calculation (5) is a further check on construct validity by taking into consideration how students answered a given item. A low differential here suggests that students must demonstrate all required behaviors to answer the item correctly, while a high differential suggests that students may circumvent the intended construct (i.e., not perform one or more required behaviors) and still answer correctly.

These results are reported in various tables throughout the next section. Staff also identified and interpreted vignettes for select items from students who demonstrated all required behaviors for a given item. These vignettes also appear throughout the results section, below, and offer insight into students' thought processes as they approach the items. These vignettes serve as further evidence that SAT items are capable of eliciting complex cognition in accordance with the items' demands.

Before we turn to the results, some caveats are in order. First, like most if not all qualitative studies, this study deals with relatively small samples of students—although in this case nearly one hundred students were sampled and almost seventy items were analyzed across three tests. Second, the gender and racial/ethnic makeup of the samples (noted above) isn't fully representative of the SAT test–taking population. Third, owing to the difficulty of obtaining enough students for this time-consuming study requiring in-person participation, students weren't selected in part based on achievement level as established by prior testing. The participating sample was higher achieving than typical for the SAT test–taking population.[8] Fourth, determining whether a given student verbalization demonstrated or failed to demonstrate a particular behavior inevitably involves some judgment on the part of raters; the numbers/percentages in the tables below thus admit to some degree of error. Fifth, students' demonstrated ability to

[8] The relatively high achievement level of the samples is arguably not a true limitation, however. The main purpose of the study wasn't to determine *typical* performance on items but rather to assess whether the items were capable of eliciting the complex cognitive processes their designers intended them to elicit. Higher-achieving students are more likely to exhibit this high level of thinking and to have the metacognitive skill to be able to verbalize their thought processes.

verbalize their thought process isn't the same as their ability to carry out complex thinking; in other words, some students may have undertaken required/expected behaviors for which the coders didn't give them "credit" owing to lack of verbal evidence. Finally, while generally of high quality, the interview recordings and subsequent transcripts have some limitations. Some students were hard to hear, a few gaps in recordings occurred, some transcripts had garbling or "inaudible" stretches, and (as indicated in the footnotes above) a small number of recordings were unusable. These issues affected only a small proportion of the work, and, except for the three lost/unusable recordings, in most cases enough context was available to circumvent problematic stretches in the transcripts.

Results

Evidence-Based Reading and Writing (ERW)

Student responses to twenty-two SAT Reading Test items and thirteen SAT Writing and Language Test items were studied using the cognitive lab methodology. The Reading Test and Writing and Language Test items represented skills and knowledge in five areas of central importance to and prominence in the tests, as discussed in the subsequent sections and as summarized in table 2.

Table 2: Breakdown of ERW Items by Test and by Category

Reading		Writing and Language	
Category	**Number of Items**	**Category**	**Number of Items**
Citing Textual Evidence	6 (plus 5 associated inference items)	Development (across Proposition, Support, Focus, and Quantitative Information types)	9
Interpreting Words and Phrases in Context	6	Effective Language Use (across Precision and Style and Tone types)	4
Analyzing Quantitative Information	5		

For each item type, a set of required behaviors was defined. In order to enact the item type as intended, each student was expected to demonstrate each associated required behavior. Doing so would also result in a correct answer because one or more behaviors associated with each item type required a correct answer to be enacted. As discussed more fully below, the approach used in Math deviates somewhat from the above, owing to the fact that the Math behaviors are expected strategies, which in many cases represent alternate, even mutually exclusive, approaches to answering an item efficiently; moreover, the performance of one or more expected strategies doesn't guarantee the proper outcome—that is, the right process can still produce the wrong result. The analytical method employed in this report for ERW and Math accounts for these differences and yields metrics that are broadly comparable with respect to whether students enacted the items' intended constructs.

READING

Citing Textual Evidence

The ability to use textual evidence effectively is critical to successful reading comprehension (as well as writing, speaking, and presenting about texts) (M. Liben 2020). With an adequate command of textual evidence, students are able to support their claims and interpretations and to make their points more convincing or persuasive. The capacity to cite textual evidence is highly rated by postsecondary instructors as a prerequisite for success in first-year, credit-bearing college courses (College Board 2019) and is fundamental to successful comprehension of and communication about texts in K–12, college, and workforce training settings (Gormley and McDermott 2015; Fisher and Frey 2015; Hart Research Associates 2018).

Developers of large-scale assessments of reading have traditionally inferred students' ability to use textual evidence by the responses they give to conventional comprehension items. That is, if students answer a given item correctly, they presumably (barring item flaws) used textual evidence to get there; if they don't answer a given item correctly, they presumably failed to find the appropriate textual evidence and/or to use textual evidence effectively.

As part of the redesign of the SAT, College Board wanted to assess students' evidence use directly. The SAT Reading Test's Citing Textual Evidence items take two forms. The more frequently used approach involves a pair of related items. The first item in the pair is a conventional inferential-level comprehension question, such as can be found on many multiple-choice assessments of reading, while the second item in the pair asks students to identify the textual evidence (typically, one of four sentence-length quotations from a passage) that best supports the answer to the previous question. The less frequently used (but still relatively common) approach takes the form of a standalone item that provides an inferential idea or conclusion in the stem and asks students to identify the textual evidence (again from a proffered range of options) that best supports that idea or conclusion. Performance on Citing Textual Evidence items contributes to a Command of Evidence subscore yielded (in combination with performance on other Reading Test and Writing and Language Test items) by the SAT.

To answer a Citing Textual Evidence item as intended, students are expected to demonstrate the following behaviors:

1. Read and demonstrate comprehension of the relevant portions of the associated passage (i.e., those associated with the inferential idea/conclusion and each evidence answer choice).
2. Draw a reasonable inference from the associated passage when answering the first of two questions in a Citing Textual Evidence pair (when applicable).
3. Use passage-based reasoning to determine the best evidence for an inferential idea/conclusion from the provided range of answer options.

Collectively, these behaviors represent a complex cognitive process involving close reading, inference making, and use of textual evidence.

Five of the studied Citing Textual Evidence items (items 2, 6, 9, 12, and 19) were part of a pair (associated with items 1, 5, 8, 11, and 18, respectively); the other

Citing Textual Evidence item (item 17) was a standalone item, with the inferential conclusion embedded in the stem of the item itself. These items were associated with passages in U.S. and world literature and history/social studies. The second required behavior above doesn't apply to the standalone item. Otherwise, students were expected to demonstrate all three of the behaviors in order for the items' intended construct to be fully enacted. However, students demonstrating one or more of the criteria could still show evidence of complex cognition.

Table 3 summarizes student performance on the six studied Citing Textual Evidence ("evidence") items. In this table (and throughout similar tables), darker cell shadings indicate higher numbers/percentages. In the five cases in which the Citing Textual Evidence item was paired with a preceding item, performance on that inference item is also noted. Sample (*n*) sizes vary, as noted, reflecting the fact that in certain cases the data were incomplete (e.g., an answer wasn't provided, the student didn't verbalize, or a gap in the transcript exists). In table 3 and subsequent tables like it, "item difficulty," where available, represents operational item difficulty (in terms of easy/medium/hard), not the difficulty of the item with respect to the sample group performance; "differential" refers to the difference between the number of students who answered the item (or, in some cases here, items) correctly and number of students who also demonstrated all required behaviors.

Table 3: Student Performance on Reading: Citing Textual Evidence Items

			Demonstrated Required Behaviors				Answered Correctly			Demonstrated All Behaviors and Answered Item(s) Correctly	
Item	**Content Area**	**Item Difficulty**	**1**	**2**	**3**	**All**	**Associated Inference Item**	**Evidence Item**	**Both Items**		**Differential**
2 *n* = 30	Literature	Hard	25 (83%)	8 (27%)	15 (50%)	6 (20%)	8 (27%)	16 (53%)	6 (20%)	6 (20%)	0
6 *n* = 30	Literature	Hard	21 (70%)	18 (60%)	10 (33%)	9 (30%)	18 (60%)	14 (47%)	9 (30%)	9 (30%)	0
9 *n* = 31	Social Science	Med	23 (74%)	16 (52%)	15 (48%)	15 (48%)	18 (58%)	20 (65%)	17 (55%)	15 (48%)	2
12 *n* = 31	Social Science	Med	25 (81%)	18 (58%)	21 (68%)	18 (58%)	18 (58%)	23 (74%)	18 (58%)	18 (58%)	0
17[a] *n* = 30	Social Science	Med	20 (67%)	N/A	18 (60%)	18 (60%)	N/A	21 (70%)	N/A	18 (60%)	3
19 *n* = 30	Social Science	Med	18 (60%)	16 (53%)	16 (53%)	14 (47%)	21 (70%)	22 (73%)	16 (53%)	14 (47%)	2

[a] Item 17 is a one-part Citing Textual Evidence item (i.e., one with the inferential idea/conclusion embedded in the stem); thus, there's no paired item, and behavior 2 is inapplicable.

The data in table 3 suggest, in general, that the Citing Textual Evidence items elicited complex cognition and that the items worked as designed. In thirteen out of seventeen cases (behaviors 1–3 across the six items, behavior 2 excepted for item 17, which isn't part of a pair), a majority of students demonstrated the required behavior. At the same time, for only two items (items 12 and 17) did a majority of students demonstrate both or all three required behaviors, although just under a majority of students demonstrated all required behaviors for items 9 (fifteen of thirty-one students, 48 percent) and 19 (fourteen of thirty students, 47 percent). Items 2 and 6 were statistically hard items that only about half of

students answered correctly, and only about a quarter of the sample answered item 1 (the inference item associated with item 2) correctly, making it impossible for a majority of students to demonstrate all associated behaviors (since behaviors 2 and 3 are tied in part to successful item performance). The difference between the students who answered both items correctly (or the single item, in the case of unpaired item 17) and the students who answered the item(s) correctly and demonstrated all required behaviors ranged from zero to three, with zero (three items) being the modal difference. This result further suggests that the items performed as intended, given that the students who answered the item(s) correctly typically also demonstrated all required behaviors.

Vignettes from select transcripts illustrate patterns of complex thought from students answering Citing Textual Evidence items correctly while also demonstrating all required behaviors.[9] For items 5 and 6, a pair of difficult items associated with a literature passage, students must first ascertain that the main character, a talented but monetarily poor artist, sees maintaining his high artistic standards as a burden that harms him financially and then find the textual evidence the best supports that inference. After ruling out the three distractors for item 5, student 24RNY uses reasoning based on the passage to conclude that the keyed response is the correct answer.

> And then D, [maintaining high artistic standards is a] "laborious undertaking that does not provide suitable compensation." I think that might be it because up in the last couple of paragraphs, last paragraph, [the main character] is kind of like ranting about [whether] waiting for a long time to find fame [. . .] is [. . .] really worth it because he won't be able pay for his rent or anything that is valuable. So, I think it's D.

After this, student 24RNY turns to the accompanying Citing Textual Evidence item. The student selects the best answer, D, by reading through the passage lines cited in the answer options and deciding that those in D best signal that the artist sees his standards as hard to maintain because they keep him from making the easy money he knows he could obtain by being a fashionable artist.

> And then D, it says [in] line[s] 74–77, "Why do I worry, and toil like a learner over the alphabet, when I might shine as brightly as the rest, and have money, too, like them?" I think it would be D because in the whole paragraph he is just complaining about how following his professor's advice [to uphold high standards] would be not financially good for him. So, and that last sentence is just like why doesn't he just follow [the fashionable artists] and follow the money. So, yeah. The answer's D, yeah.

While the above vignette exemplifies a serial process to answering paired items, some students approached the pairs in a less linear way. In paired items 18 and 19, both of medium difficulty, students must reach two conclusions from a social science passage on human brain activity: first, that a particular part of the brain (the amPFC) in a research study participant is most likely to experience an increase in activity when exposed to a scenario in which a real protagonist interacts

[9] In the vignettes that follow throughout the report, a few small edits for readability have been made "silently," while larger edits have been noted in brackets.

with real people who were childhood friends of the participant (as opposed to scenarios with less real-life, personal relevance to the participant), and, second, which of four sets of lines from the passage best supports the answer to the first question. Student 2RCA, who exhibited all three required behaviors, offers clear insight into their thought process. The student notes that items 18 and 19 are connected and will "work on these two together." After reading the stem to and answer options for item 18, the student notes:

> "The greatest increase in activity in the amPFC of a research subject's brain [would] most likely be" . . . "activity." So, we're focusing about activity in amPFC. Therefore, we must first look back to the passage about what does amPFC do. So, this part of line 33 says, "When exposed to scenarios featuring George Bush—a famous real person—the brain [involved] the amPFC [anterior medial prefrontal cortex] and the PCC [precuneus and posterior cingulate cortex]." So, therefore, I suspect that amPFC is relative and related to actions about a real person.

With some idea already of the nature of the correct answer to 18, student 2RCA reads the stem and answer options to item 19, the Citing Textual Evidence item, and considers option C, the keyed response.

> Therefore, looking to line[s] 62[–64], which is answer C, "As predicted, the activation in the amPFC and PCC [was] indeed proportionally modulated by the degree of relevance to the characters described." And I think that directly connects to the question in 18, which is the increase in activity in amPFC of a research subject's brain.

After returning to item 18, the student reaches a passage-supported conclusion about the relationship between amPFC activity and interaction with real, known people. In the process, the student rules out a tempting distractor (option D) also involving a real person but one whom the participant has only met, as opposed to having been childhood friends with.

> And "high personal relevance" is their friend or family, so therefore, I think interaction with real people, childhood friends of the subject's [answer option B] makes most sense, since it directly connects to the line [about] their friends or families. . . . So, is told about a real person—D, "is told about a real person the subject has previously met," does not give us any degree of relevance that this person has with a subject of the experiment.

The student then identifies the correct answers to both items 18 and 19.

> So, therefore, B should be the correct answer for number 18, and C should be the correct answer for number 19. That will be my answer for my answer sheet.

While responding to this item pair, student 2RCA uses a recursive process to work through both items. The student begins with a general understanding of the nature of the key to the inference item (item 18), uses the options in the evidence item (item 19) to help solidify an understanding, and returns to the inference item, finalizing the correct answers to both items.

The lone single-part Citing Textual Evidence item (item 17) studied elicited complex thought processes as well. As previously noted, in this format the inferential idea or conclusion appears in the item's stem, and students must determine the evidence from the passage that best supports that idea or conclusion. In item 17, a medium-difficulty item associated with the social science passage discussed above, students must determine which option "best supports the claim that there are important similarities between how the brain responds to scenarios involving real people and how it responds to those involving fictional people." As with the other studied Citing Textual Evidence items, four answer options—each quoting a short segment of the passage—are provided. In approaching this question, student 3RKY reasons carefully through each option in turn, testing each against the idea asserted in the item stem.

> So, I'm going to go back to lines 21 to 24 [option A]. "Common to both types of [situation was some] level of [mental] activity in parts of the brain, such as the hippocampus, that are at work when we in general recall facts or events." That one is highlighting similarities so I'm going to keep that in my mind while I'm reading other answers.
>
> [Lines] 29 to 32 [option B]. "However, there [were] a few striking finer distinctions in activity relative to the two scenarios and these depended on the type of character involved." The question is asking about supporting the claim that there are similarities. And that one's discussing differences, so it's not B.
>
> C. [Lines] 72 to 77. "You are familiar with their basic behavioural features as human beings." That whole section is talking about—well, I didn't read the whole section. But, by contrast, your mind is not equally familiar with fictional characters. That one's making contrast, and the question's asking about similarities, so it's not C.
>
> And then D. [Lines] 81 to 88. "You may have read all the books about a fictional character, but the amount of information you have gathered [about that character] is still [definitely limited] compared" to people who are real people. So, 81 to 88 is not correct because it's once again referencing differences, when the claim that the question is referring to is about their similarities. So, my answer is A.

As table 3 indicates and as the vignettes suggest, the Citing Textual Evidence items studied were able to call forth complex thought processes from students. These processes included reading closely, drawing inferences, and supporting inferences with textual evidence.

Interpreting Words and Phrases in Context

Vocabulary knowledge and the ability to apply vocabulary strategies, including the use of context clues, to determining the meaning of words and phrases in context, have a close association with students' reading achievement and with college and career readiness more generally. Tier two words and phrases (Beck, McKeown, and Kucan 2013)—those commonly found in readings, especially more complex readings, across a range of subject areas but infrequently in everyday speech—have particular importance to secondary success and postsecondary readiness because they help unlock the meaning of the sorts of texts read in challenging

middle school and high school courses and in college and workforce training programs (D. Liben 2020).

One aim of the redesign of the SAT was to increase the relevance of the test to both K–12 classroom practice and to students' lives and educational aspirations. To this end, College Board eliminated testing of so-called SAT words. Though there had never been a formal list of tested vocabulary, low-frequency words and phrases appeared in some items in previous iterations of the SAT, and third-party test preparation had promoted the memorization of the definitions of such words and phrases as critical to success on the exam. In place of obscure words and phrases that for many students would likely be encountered only on test day, College Board put a strong focus in the redesign on high-utility academic (tier two) words and phrases—those words and phrases students would likely encounter again and again in coursework in both secondary and postsecondary classrooms as well as in careers and life.

The SAT associates Interpreting Words and Phrases in Context items focused on tier two words and phrases with nearly every Reading Test passage, and performance on these items contributes to a Words in Context subscore yielded (in combination with performance on other Reading and Writing and Language items) by the SAT.

To answer an Interpreting Words and Phrases in Context item as intended, students are expected to demonstrate the following behaviors:

1. Read and demonstrate comprehension of the local context (the sentence or at most a paragraph) in which the focal word or phrase appears.
2. Use passage-based reasoning to select the answer option whose meaning most nearly captures how the focal word or phrase is used in context, which may optionally involve evaluating one or more of the distractors.

Together, these behaviors require students to demonstrate a complex understanding of how select tier two words and phrases are used in the contexts in which they appear. In the process of demonstrating these behaviors, students may exhibit use of context clues, an understanding of connotation and shades of meaning in relation to synonyms and closely related words and phrases, and other vocabulary knowledge and strategies.

Table 4 summarizes student performance on the six studied Interpreting Words and Phrases in Context items. Sample (*n*) sizes vary, as noted, reflecting the fact that in certain cases the data were incomplete.

Table 4: Student Performance on Reading: Interpreting Words and Phrases in Context Items

Item	Content Area	Item Difficulty	Demonstrated Required Behaviors			Answered Correctly	Demonstrated Both Behaviors and Answered Correctly	Differential
			1	2	Both			
3 *n* = 30	Literature	Med	26 (87%)	16 (53%)	16 (53%)	18 (60%)	16 (53%)	2
4 *n* = 31	Literature	Med	26 (84%)	22 (71%)	22 (71%)	26 (84%)	22 (71%)	4
7 *n* = 31	Social Science	Hard	16 (52%)	8 (26%)	8 (26%)	12 (39%)	8 (26%)	4
10 *n* = 31	Social Science	Easy	16 (52%)	16 (52%)	16 (52%)	31 (100%)	16 (52%)	15
16 *n* = 30	Social Science	Easy	18 (60%)	18 (60%)	18 (60%)	29 (97%)	18 (60%)	11
20 *n* = 30	Social Science	Easy	22 (73%)	21 (70%)	21 (70%)	26 (87%)	21 (70%)	5

The data in table 4 suggest, in general, that the students demonstrated complex behaviors when answering Interpreting Words and Phrases in Context items and enacted the items' intended design. In eleven out of twelve cases (behaviors 1 and 2 across the six items), a majority of students demonstrated the required behavior. In five out of six cases (items 3, 4, 10, 16, and 20), a majority of students demonstrated both required behaviors for the item; the exception (item 7) was a statistically hard item that students in the study answered correctly at a moderately low rate (39 percent), which precluded a majority from demonstrating both required behaviors (since behavior 2 is tied in part to successful item performance). In four cases (items 3, 4, 7, and 20), the difference between the students who answered the items correctly and those who answered correctly and also demonstrated both required behaviors ranged from two to five. In two other cases, however, the gap was considerably higher: fifteen for item 10 and eleven for item 16. At least some of this gap was likely a product of the extreme ease of the items for participating students, which seemed, from the transcripts, to discourage introspection (chiefly affecting behavior 2). Many students reached for a synonym to the focal words ("exact," as in "exact date"; "operate," as in "how does the brain operate") without much explicit reasoning, although students did reread the sentence-level context and often substituted answer options into it, suggesting the items have at least some level of text dependency. In the main, the data in table 4 suggest that the Interpreting Words and Phrases in Context items performed as expected and elicited complex cognition, particularly (but not only) where the items were moderately difficult or difficult.

Vignettes from students in the study who answered correctly and demonstrated both required behaviors associated with Interpreting Words and Phrases in Context items offer evidence of the complex cognitive demands associated with these items. Student 3RKY (also quoted above for their response to a Citing Textual Evidence item) uses a range of knowledge and skills to determine the meaning of the word "alien" as it's used in the context "No matter how much we know about the world of a fictional character there will still be something alien and inscrutable to us about that world." Item 20, a statistically easy item associated

with a social science passage, offers "inconsistent," "foreign," "extraterrestrial," and "complex" as answer options, with "foreign" being the keyed response. The student begins by rereading the immediate, sentence-level context in which the word appears and then uses reasoning and vocabulary knowledge and skill, including an appreciation of connotation and shades of meaning, to discern the intended meaning of "alien."

> "Inconsistent" [option A]. I wouldn't say "inconsistent." The connotation for "alien" for me that I get is it's "foreign" [option B] to us. That we don't know. So, B is seeming like the correct answer for me right now. "Extraterrestrial" [option C]. It's not talking about literally from another planet. So, I'm gonna go ahead and cross out C. [Option] D, "complex." While this may be complex, I don't think "complex" is the answer because it's just talking about something that we won't necessarily—no matter how much we know about that [fictional] world, there's still gonna be stuff that we don't know. And so "foreign" seems more fitting with something that we don't know than "complex." It doesn't necessarily—for me that sentence just made me think about how intricate these ideas are. Just a matter of basic "Do I know it or not?" So, I'm going to go with B, "foreign."

Item 3, a medium-difficulty item associated with a literature passage, is unusual (but not unique for the SAT) in testing the meaning of the same word—in this case, "fashionable"—appearing multiple times in the same passage. To answer the item correctly, students must discern that "fashionable" in this context most nearly means "trendy" and not "stylish," "modern," or "conventional" in reference to a character's portrayal of "fashionable artist[s]," "the fashionable style," and "fashionable little pictures and portraits [made] for money." Students need to recognize that the character takes a dim view of this sort of fashionableness and those who aspire to it and also attend to the connotations and shades of meaning of the answer options.

Student 4RCA approaches item 3 by rereading each sentence from the passage in which "fashionable" appears and then summarizing the point of view reflected.

> So, by these definitions, the professor associates "fashionable" with society that you think is too bold and garish.

The student then uses vocabulary knowledge and an understanding of both the context and the point of view to infer the exact intended meaning for "fashionable," which the student confirms by looking at the answer options.

> So, looking at the choices, I would say that "fashionable" most nearly means—not by looking at the answer choices, just by the passage—"fashionable" probably would mean "trendy" or something that appeals to society but not to your own inclinations or what you would like to do. So, looking at the choices, B matches perfectly, which is "trendy."

Student 17RNY employs a similar but more holistic approach to assessing the intent behind the use of "fashionable."

> Looking back at the passage, every time the professor uses "fashionable," he says it in mostly a negative tone, like, "You don't wanna become

> like everyone else. You wanna become your own person and not just something to look up at someone who has technique and stuff."

The student then muses over the match between the connotations and shades of meaning of the answer options and the character's use of "fashionable."

> So, I wouldn't necessarily say "stylish" [option A] because that's positive in saying, "Oh, you don't wanna be just a stylish artist." There's nothing really wrong with being stylish. And then choice B says "trendy." And it could be "trendy" because he doesn't—the professor doesn't want Tchartkoff to follow trends. He wants him to be his own individual. [Option] C says "modern." But there isn't anything that would say anything modern about what the professor is saying. And [option] D, "conventional." That doesn't really make sense because it doesn't fit in with the rest of the paragraph and how he doesn't want him to follow others. So, I would say B, "trendy."

Item 7, the statistically hardest item included in the studied sample and one associated with a social science context, involves determining that "curiosity" most nearly means "oddity" and not "concern," "question," or "wonder" in the sentence "False memories can sometimes be a mere curiosity, but other times they have real implications." In approaching the item, student 1RCA first rereads the sentence and then the answer options. The student notes that "mere" is a key qualifier of "curiosity" pointing to the intended meaning.

> Now, it's the modifier "mere" that changes what "curiosity" means. "Curiosity" on its own would have something different, but the fact [of] the inclusion of "mere" makes it seem like it's not important.

The student uses this understanding, derived from vocabulary knowledge and a close reading of the context, along with the strategy of substituting the answer options in for the tested word, to process the options and pick the correct answer.

> I can eliminate "question" [option B], and that would seem to me because there's nothing really questionable about it. It seems like it's more of an emotion. I could see how "concern" [option A] would fit in that same area if you were to exchange "curiosity" for "concern": "False memories can sometimes be a mere concern." But that doesn't sound right in the way that it should be in the context, so I can eliminate "concern" because there wouldn't be a problem. [Option] C, "oddity." Same method. Go to line 7. "False memories can sometimes be a mere oddity." That seems to fit more in terms of "curiosity," as it seems that it seems more like an irregularity. [Option] D, "wonder," has the same context as "oddity" of something strange, but "oddity" seems to fit more with the context of false memories than "wonder." The connotation of "wonder" is a little grayer. "Oddity" makes it seem a lot less common. I will be bubbling C for 7.

Student 1RCA's response exhibits many complex elements. The student uses both vocabulary knowledge and a close reading of the sentence in which "curiosity" appears to recognize that "mere" significantly influences the target word's meaning, indicating that the intent is to signal something fairly trivial. The

student also understands that while "concern" scans in the sentence, "mere curiosity" suggests the opposite of a concern ("there wouldn't be a problem"). The student also determines that "curiosity" in this context likely means "more like an irregularity," even though "irregularity" isn't an answer option. In addition, the student differentiates between "oddity" and "wonder" by noting that while both signal "something strange," "wonder" doesn't make much sense when talking, as the passage does, about false memories. Finally, the student uses the technique of substitution—inserting answer options for the tested word—to "sound out" the alternatives in context.

As indicated by table 4 and suggested by the above vignettes, students answering the studied Interpreting Words and Phrases in Context items called on vocabulary knowledge and skills as well as sophisticated passage-based reasoning to determine the most likely meaning of a range of tier two words situated in rich contexts. Among the attributes students demonstrated were word/phrase knowledge, an understanding of connotation and shades of meaning among words with similar denotations, and the ability to use context clues and other vocabulary strategies to infer and verify definitions.

Analyzing Quantitative Information

The ability to analyze data conveyed in informational graphics, such as tables, graphs, and charts, and to draw meaningful connections between these data and information and ideas conveyed in words is integral to successful comprehension of texts in numerous disciplines, including the natural and social sciences (Shanahan and Shanahan 2020). In turn, comprehension of such disciplinary texts is required for success in K–12 and in college and workforce training programs (whether in the latter as part of general education requirements or many majors/minors/certification programs) (Bain 2012; Moje, Stockdill, and Hornak 2019; College Board 2019).

One goal of the SAT redesign was the inclusion of informational graphics, chiefly in the forms of tables and graphs displaying quantitative data, as part of the Reading Test (and Writing and Language Test). This inclusion enhances the test's congruence with academic and real-world reading requirements in various disciplines, as multimodal texts (texts including words and other elements, such as informational graphics) are a key feature of how knowledge is constructed and how information and ideas are conveyed in various fields, including science and social science (Shanahan and Shanahan 2020).

Three basic "levels" of Analyzing Quantitative Data items appear on the SAT Reading Test. These levels reflect varying cognitive demands and are associated with varying depth of knowledge (DOK) levels. At the most basic level (associated with DOK 1), students must locate particular data from one or more informational graphics. At the middle level (DOK 2), students must reach a reasonable interpretation of data from such graphics. At the highest level (DOK 3), students must both locate and/or interpret data from such graphics and use the information thus gathered in concert with information and ideas from an associated test passage. All informational graphics on the Reading Test (and Writing and Language Test) appear in conjunction with topically related passages and are typically drawn from the published research studies on which the passages are based.

The SAT associates Analyzing Quantitative Data items with all social science Reading passages and with many science Reading passages, and performance on these items contributes to a Command of Evidence subscore yielded (in combination with performance on other Reading and Writing and Language items) by the SAT.

To answer an Analyzing Quantitative Information item as intended, students are expected to perform the following behaviors, which, as noted below, vary in some cases depending on the item's DOK level:

1. Read and demonstrate a general understanding of one or more informational graphics (all DOK levels).
2. Read and demonstrate comprehension of one or more portions of the associated passage containing information pertinent to the graphic(s) and the question (DOK level 3).
3. Locate data required to answer the question from one or more graphics (all DOK levels).
4. Offer a reasonable interpretation of data from one or more graphics relevant to answering the question (DOK levels 2 and 3).
5. Draw a reasonable connection between relevant data from one or more graphics and one or more portions of the associated passage containing information pertinent to the graphic(s) and the question (DOK level 3).

Collectively, these behaviors require students to demonstrate a range of knowledge and skills associated with using data in informational graphics, either alone or in conjunction with the information and ideas in associated passages. These behaviors range from the relatively straightforward activity of locating pertinent data to the more complex activity of drawing reasonable interpretations of data to the highly complex activity of synthesizing data with textual content to reach an understanding not obtainable from either source alone.

Table 5 summarizes student performance on the five studied Analyzing Quantitative Information items. Sample (*n*) sizes vary, as noted, reflecting the fact that in certain cases the data were incomplete.

Table 5: Student Performance on Reading: Analyzing Quantitative Information Items

				Demonstrated Required Behaviors								
Item	**Content Area**	**Item Difficulty**	**DOK**	**1**	**2**	**3**	**4**	**5**	**All**	**Answered Correctly**	**Demonstrated All Behaviors and Answered Item(s) Correctly**	**Differential**
13 *n* = 31	Social Science	Easy	1	19 (61%)	N/A	31 (100%)	N/A	N/A	19 (61%)	31 (100%)	19 (61%)	12
14 *n* = 31	Social Science	Med	2	22 (71%)	N/A	22 (71%)	17 (55%)	N/A	17 (55%)	22 (71%)	17 (55%)	5
15 *n* = 31	Social Science	Easy	3	25 (81%)	14 (45%)	26 (84%)	12 (39%)	6 (19%)	6 (19%)	29 (94%)	6 (19%)	23
21 *n* = 30	Social Science	Easy	1	23 (77%)	N/A	25 (83%)	N/A	N/A	23 (77%)	25 (83%)	23 (77%)	2
22 *n* = 29	Social Science	Easy	2	21 (72%)	N/A	27 (93%)	20 (69%)	N/A	20 (69%)	27 (93%)	20 (69%)	7

The data in table 5 suggest—with some caveats—that the students demonstrated complex behavior when answering Analyzing Quantitative Information items and enacted the items' intended design. In twelve out of fifteen cases (behaviors 1–5, as applicable, across the five items), a majority of students demonstrated the required behavior. If we momentarily exclude behavior 3—locating data—from consideration on the grounds that this behavior represents a comparatively simple cognitive activity, we find that in seven out of ten cases, a majority of students demonstrated a given required behavior. In four out of five cases (items 13, 14, 21, and 22), a majority of students demonstrated all required behaviors for the item (including behavior 3).

Item 15, for which fewer than half of students demonstrated behaviors 2, 4, and 5 and for which only six students demonstrated all five behaviors, deserves scrutiny. Unlike the three previously discussed Reading Test items not eliciting all required behaviors from a majority of students, item 15 wasn't a statistically difficult item operationally, nor did students in this study struggle to answer it—in fact, 94 percent answered it correctly. One factor in the low proportion of students demonstrating all required behaviors is simply the fact that five behaviors—more than for any other Reading Test item—were required; given that, as a group, fewer than half the students demonstrated behaviors 2, 4, and 5, it was inevitable that the proportion of students demonstrating all five behaviors would be low. However, likely a more important factor (and one that at least partially subsumes the one previously discussed) is that the item itself isn't as synthetic as it ostensibly is. Whereas DOK 3 Analyzing Quantitative Information items should require students to use, in roughly equal measure, both passage and informational graphic(s) as sources for the answer, item 15 is likely answerable from each source independently. That is, both the passage and the figure cited in the item stem each contained enough information to enable many students to answer the item correctly, meaning that students didn't necessarily have to demonstrate the full range of activity in order to reach the correct answer.

Two items evince substantial gaps between the number of students who answered correctly and the number who both answered correctly and demonstrated all required behaviors. Not surprisingly, one of these is the previously discussed item 15. The other is item 13, a statistically easy DOK 1 item for which a majority of students demonstrated both required behaviors but which was answered correctly by a higher proportion of students—in fact, by all participating students. The low difficulty and cognitive demand of this item likely meant that students didn't feel compelled to verbalize their reasoning, or even to use much reasoning. In this case, students could—and, per the transcripts, fairly often did—determine the values represented by the two bars in the associated bar graph and match those values to the keyed response without elaborating on their thought process or indicating a clear understanding of what the values or the figure represented, which depressed their demonstration of behavior 1.

Vignettes from students in the study who answered correctly and demonstrated all required behaviors required for a given Analyzing Quantitative Information item strongly suggest that the DOK 2 and 3 items (even item 15) were able to elicit complex cognitive processes. Since the DOK 1 items (items 13 and 21) involve

relatively low-level behaviors, the following discussion focuses on the two DOK 2 and one DOK 3 items (items 14 and 22 and item 15, respectively).

Item 14, a medium-difficulty DOK 2 item associated with a social science passage, asks students to determine which conclusion about participants in a study who have ordinary memory (as opposed to those with highly superior autobiographical memory, or HSAM) is supported by data in two bar graphs. Figure 1 shows the mean proportion of words in a word list memory test recalled by members of the ordinary-memory and HSAM memory groups, while figure 2 shows the mean proportion of "critical lures" (words similar to but not actually included in the word list test) that were (wrongly) recalled by members of the two groups. Student 8RNY addresses item 14 first by accurately determining the values represented in the two figures and then, after reading the answer options, by establishing the proper relationship between those values.

> The mean proportion [of words included in the word list test recalled by the group with ordinary memory] is in between 0.6 and 0.7, and what is not included [i.e., the mean proportion of critical lures falsely recalled by the group with ordinary memory] is between—is that over 0.7 or 0.7? That's over 0.7. . . . "They often recall words that neither were included on the list nor were critical lures [option A]. . . . They were allowed more time to complete the test than [were] the study subjects with HSAM [option B]. . . . They recalled a greater proportion of critical lures than included words, on average" [option C]. Okay, and "They confused critical lures for included words approximately 50 percent of the time, on average" [option D]. I'm going to say C, because they did—the words not included, it's higher than the ones that they did include, so let's say C for that one.

Item 22, a statistically easy DOK 2 item associated with a social science passage, calls for a comparison between eight conditions represented across two bar graphs. Relative to four conditions, figure 1 displays the percent change in activity in the precuneus and posterior cingulate cortex (PCC), while figure 2 displays the percent change in activity in the inferior frontal gyrus (IFG). Across these eight conditions and two figures, students must determine the highest percent change in activity. Student 5RCA summarizes the task as "just looking at the highest one altogether"—that is, the highest bar in the two graphs.

> The greatest percentage change overall is 0.3 percent from figure 2, which is the IFG one. So, when I'm looking at that, it's [exposure to] "fictional characters" in an "interactive scenario" under IFG. "Fictional characters, interactive scenario." So, I know that it's not A because that says PCC [has the highest percent change] and it's not C because that says PCC. Interactive scenario, fictional characters. That is B. Therefore, the answer to number 22 is B.

As previously noted, item 15, the sole DOK 3 (synthetic) Analyzing Quantitative Information item studied, elicited the five required behaviors from only six students. As we discussed, this was likely due in large part to the vulnerability of the item to being answered from either the passage or figure 2 alone. Nonetheless, from the students who did demonstrate all five required behaviors, we can see clear evidence of complex cognition—the sort of cognition normally required from DOK 3 quantitative items on the Reading Test.

Item 15 asks students to use one of the passage's bar graphs (the one previously described indicating the mean proportion of critical lures "recalled" by members of two groups with different memory traits) to determine which assertion is supported in both sources about people with HSAM. The correct response, "They are about as susceptible to memory distortion as are people with ordinary memory," is supported in two different ways. First, the passage asserts that "all of the participants in both groups fell for the lures" and that "both groups also performed unreliably when shown photographs and fed information intended to make them think they'd seen details in the pictures they hadn't." Second, figure 2 illustrates that members of the group with ordinary memory and members of the group with HSAM wrongly recalled an equal or nearly equal proportion of critical lures (approximately 0.7).

Student 29RNY exemplifies the intended approach to the item, drawing on passage and graph in roughly equal measure. The student first notes what they think the nature of the right answer is and then works through each answer option, noting its support or lack thereof in the passage and/or in the graph.

> Okay. I believe I already know what it's looking for here. It mentions somewhere in the passage that the HSAM group is just as likely as the ordinary memory group to be baited in by the critical lures, I guess.
>
> So, let's read the answer choices. "They [people with HSAM] are characterized by an exceptional ability to recall minute details of daily events" [option A]. Not really supported by the graph there. Look at the other answer choices anyway. "They are almost as susceptible to verbal lures as they are to visual lures" [option B]. Not really. It mentions critical lures, but neither verbal nor visual either. "They are more skilled than people with ordinary memory in distinguishing false memories from true memories" [option C]. No, the graph does not support that conclusion, nor does the passage. "They are about as susceptible to memory distortion as are people with ordinary memory" [option D]. That's the answer I was looking for here. So, choosing D here, as it's supported by the graph and mentioned in the passage.

Note that the student is in some sense incorrect about answer option B, "They are almost as susceptible to verbal lures as they are to visual lures," because the passage, though not the graph, mentions that both groups of participants were equally flummoxed by a recall test involving photographs and fake information fed to them intended to test their memory of actual details. This observation doesn't, however, take away from the fact that student 29RNY evinced a complex synthetic set of behaviors in correctly answering item 15.

As the above discussion indicates, the responses from the students sampled to the Analyzing Quantitative Information items, particularly the DOK 2 and DOK 3 items, are, on the whole, indicative of the complex cognitive processes these items expect of test takers. Collectively, these items required students to demonstrate the ability to locate data, to make reasonable inferences from data, and to synthesize information and ideas from words and data, the latter two activities being moderately or highly complex cognitively. Although the most cognitively complex item in this category (item 15) lacked the full synthetic robustness that

one would expect DOK 3 items to exhibit, it still was able to draw out high-level thinking from a subset of the participating students and is suggestive of the capacity of such items, when more tightly constructed, to demand such thinking from all students.

WRITING AND LANGUAGE

Development

Development is one of six broad knowledge and skill areas sampled by the SAT Writing and Language Test. The four specific item types composing this area—Proposition, Support, Focus, and Quantitative Information—pertain to a range of knowledge and skills fundamental to the effective revision of prose text to enhance how it structures and conveys information and ideas. Proposition items deal with the "big ideas" framing text: main points, central claims, thesis statements, and the like. Support items address the effective use of facts, figures, quotations, details, and so on to flesh out propositions. Focus items involve revising text to improve its adherence to its intended topic at a paragraph and/or passage level. Quantitative Information items ask students to use data accurately and effectively from informational graphics such as tables and graphs to strengthen and clarify writing. Performance on Development items contribute to a Command of Evidence subscore (in combination with performance on select items from the Reading Test) and to an Expression of Ideas subscore yielded by the SAT.

To answer a Development item as intended, students are expected to demonstrate the following behaviors, which, as noted below, vary depending on which of the four Development item types is being addressed:

1. Read and demonstrate comprehension of the local (sentence- or paragraph-level) context in which a given Development question is situated.
2. Read and demonstrate comprehension of relevant portions of the passage, up to and including the whole passage, the scope being defined by the item stem.
3. Demonstrate an understanding of the framing language of the stem (i.e., whether to add, revise, retain, or delete information and ideas).
4. Demonstrate, at least indirectly, a conceptual understanding of essayistic composition, of revision as a general process, and of topic development as a specific process and outcome.
5. **Proposition:** Demonstrate an understanding of the key ideas, points, or claims of a paragraph or passage (depending on the scope of the question as established by the item stem) in the course of making an effective decision regarding adding or revising a statement intended to express that idea, point, or claim.
6. **Support:** Demonstrate an understanding of a given idea, point, or claim in a passage in the process of making an effective decision regarding adding, revising, or deleting information intended to clarify, elaborate on, provide evidence for, or exemplify that idea, point, or claim.
7. **Focus:** Demonstrate an understanding of the gist of a paragraph or passage (depending on the scope of the question) in the process of making an effective decision regarding adding, revising, retaining, or deleting information and ideas on the basis of relevance to the topic.

8. **Quantitative Information:** Demonstrate an understanding of an informational graphic and the data it contains in the process of making an effective decision regarding adding or revising a verbal representation of data from that graphic in the passage in order to make that representation clearer or more precise or accurate, in so doing drawing roughly equally on both passage and graphic.

Taken together, the behaviors listed above require students to demonstrate a range of knowledge and skills involving adding, revising, deleting, or retaining information and ideas on the basis of a clear understanding of content and context. In the process of answering the items, students must exhibit the complex cognitive processes of establishing and refining key points in essayistic writing, buttressing these points with relevant support, recognizing and eliminating deviations from the text's focus, and using data from informational graphics accurately and strategically to enhance the precision, clarity, and persuasiveness of writing.

Table 6 summarizes student performance on the nine studied Development items. Sample (*n*) sizes vary, as noted, reflecting the fact that in certain cases the data were incomplete.

Table 6: Student Performance on Writing and Language: Development Items

Item	Content Area	Item Difficulty	Demonstrated Required Behaviors						Answered Correctly	Demonstrated All Behaviors and Answered Correctly	Differential
Proposition											
			1	2	3	4	5	All			
3 *n* = 30	History/ Social Studies	Med	22 (73%)	22 (73%)	26 (87%)	26 (87%)	22 (73%)	22 (73%)	25 (83%)	22 (73%)	3
Support											
			1	2	3	4	6	All			
8 *n* = 30	History/ Social Studies	Easy	26 (87%)	26 (87%)	27 (90%)	28 (93%)	26 (87%)	25 (83%)	26 (87%)	25 (83%)	1
14 *n* = 28	Humanities	Med	20 (71%)	20 (71%)	21 (75%)	21 (75%)	19 (68%)	15 (54%)	20 (71%)	15 (54%)	5
18 *n* = 29	Humanities	Med	16 (55%)	16 (55%)	22 (76%)	28 (97%)	15 (52%)	13 (45%)	16 (55%)	13 (45%)	3
Focus											
			1	2	3	4	7	All			
6 *n* = 30	History/ Social Studies	Easy	28 (93%)	28 (93%)	28 (93%)	29 (97%)	27 (90%)	27 (90%)	28 (93%)	27 (90%)	1
Quantitative Information											
			1	2	3	4	8	All			
24 *n* = 30	Careers	Easy	27 (90%)	27 (90%)	27 (90%)	27 (90%)	24 (80%)	24 (80%)	25 (83%)	24 (80%)	1
25 *n* = 30	Careers	Easy	29 (97%)	29 (97%)	29 (97%)	29 (97%)	29 (97%)	29 (97%)	30 (100%)	29 (97%)	1
40 *n* = 30	Careers	Med	28 (93%)	28 (93%)	26 (87%)	28 (93%)	25 (83%)	24 (80%)	26 (87%)	24 (80%)	2
41 *n* = 30	Careers	Med	28 (93%)	28 (93%)	29 (97%)	29 (97%)	15 (50%)	15 (50%)	25 (83%)	15 (50%)	10

The data in table 6 suggest, with some caveats, that the students generally demonstrated cognitively complex behavior when answering Development items and enacted the items' intended design. In forty-four out of forty-five cases (behaviors 1–8, as applicable, across nine items), a majority of students demonstrated the required behaviors, and the sole exception was one case (behavior 8, item 41, Quantitative Information) in which exactly half the students demonstrated the behavior. In seven out of nine cases (item 3, Proposition; items 8 and 14, Support; item 6, Focus; and items 24, 25, and 40, Quantitative Information), a majority of students demonstrated all required behaviors for the item. In one of the remaining cases (item 41, Quantitative Information), exactly half of the students demonstrated all required behaviors, and in the remaining case (item 18, Support), just under 50 percent of the students demonstrated all required behaviors.

For eight out of nine items, the difference between the number of students who answered correctly and the number who both answered correctly and demonstrated all required behaviors ranged from one to five, with one (four cases) being the modal number. The aberrant item was item 41, a medium-difficulty item in a careers-related passage for which ten fewer students answered correctly and demonstrated all five required behaviors than answered correctly. This was nearly exclusively due to a relatively low rate of demonstration of behavior 8, which requires students to show use of both passage and (in this case) data in an associated table to reach the correct answer. This was largely a product of these students relying more heavily on the passage than the table for the answer, or less often, exclusively on the passage. A largely or wholly passage-based answer was possible in this case because students could surmise that only one answer option logically followed the associated sentence's reference to turnover rate, while the other options focused on wages or staff size. The relatively low number of students demonstrating the required behaviors collectively and behavior 8 individually relative to the number of students who answered correctly does point to a weakness in this item vis-à-vis its intended construct—one comparable to that found in the Reading DOK 3 Analyzing Quantitative Information item previously discussed.

Vignettes across the four Development item types from students in the study who answered correctly while demonstrating all required behaviors for a given item offer further evidence that these items elicited complex cognitive processes. The following sections address each of the four item types in turn.

Proposition

Proposition items involve adding, revising, or retaining the key points—main ideas, central claims, and the like—that lend substance and structure to a text. In a history/social studies passage about cities converting landfills into parks, students answering item 3, a medium-difficulty item, must determine the choice that best introduces the main idea of a paragraph. The paragraph goes on to list several benefits of such conversions: landfill land is inexpensive and widely available, and its conversion increases property values, supports residents' health and well-being, and reduces crime. After reading item 3's stem and answer options, student 8WLKY begins their approach to the item by providing their own summary of the gist of the paragraph.

> I thought the main idea of the paragraph was to tell the benefits of converting past landfills—or reclaiming landfills to use as park spaces in bigger cities and urban areas . . .

The student then evaluates each of the three distractors, dismissing each as inadequate as a statement of the main idea.

> I don't think A ["There is no official count of the number of parks built on landfills nationwide, but it could be as many as 1,000"] would be the answer because that doesn't really seem like a statement about a paragraph. It seems more like a fact you would include in the middle of the paragraph. To do this, I think I'd reread the paragraph again. I already have A eliminated. I think I would be able to eliminate D ["Americans generated 251 million tons of trash in 2012, only 34.5 percent of which was recycled"], too, because that seems more of a fact—not introducing the idea of the paragraph. I think for this answer—or for this question, I would also eliminate B ["For many environmental and logistical reasons, not all landfills can be repurposed as parks"] because I think it talks more about the reasons landfills can't be repurposed in later paragraphs, so then I think my answer for that one would be C ["Reclaiming landfills for park space offers multiple advantages to urban areas"].

In response to item 3, this student first distills the content of the paragraph into their own statement of the main idea against which the options can be checked. In doing so and in ruling out two options as mere facts, the student also demonstrates an abstract understanding of the function of main ideas in essayistic writing. The student then rejects the third distractor as within the scope of the passage but not of the paragraph. Finally, the student settles on the correct answer, which is a clear match to the student's own statement of the gist of the paragraph.

Support

In Support items, students must add, revise, retain, or delete information and ideas with the goal of providing relevant support—facts, figures, quotations, details, and the like—to ideas, points, and claims made by the writer of a given passage. To answer item 18, a medium-difficulty item from a humanities passage about Arthur Conan Doyle's decision first to kill off Sherlock Holmes and then, amid reader protests, to bring him back to life, students must determine which second example, among the four proffered answer options, is most similar to the example of "public outcry" already in the passage: "the author received an abundance of angry letters petitioning for Holmes's return." The correct answer—"'Keep Holmes Alive' Clubs formed"—is similar to the previous example in that it identifies an action people took to protest Conan Doyle's decision to stop writing Holmes stories. By contrast, "[the letters] were not enough to make him change his mind" is an outcome, not an example; "those writing to him came from all parts of society" is merely an elaborative detail; and "he turned his attention to other forms of writing" is simply a subsequent action.

In approaching this item, student 6WLNY, after reading the stem and options, rereads the sentence from the passage in which the first example ("angry letters") appears and in which the second, similar example is to appear. They then reiterate

that the goal is to provide a comparable example and evaluate the various options in light of that criterion.

> An "example." "Those writing to him came from all parts of society," I would think that—well, D is irrelevant, "[he turned his] attention to other forms of writing." A ["(the letters) were not enough to make him change his mind"] is not focused on the outcry, the public outcry. So that leaves B, "'Keep Holmes Alive' clubs [formed]" and C, "Those writing to him came from parts of all society."

After reducing the choices to two, the student repeats the goal and picks the correct answer.

> And I think that would—hmm, the "most similar." And "'Keep Holmes Alive' clubs formed."

In their approach to the item, the student recognizes that the goal is to find another example of an "outcry," which quickly leads them to rule out two options as not being examples. To make the final decision, the student refines their understanding of the stem's criterion to focus on the most similar example, which leads them to select the correct answer. The student might have determined that option C isn't an example of an "outcry" either, but the important point here is that they demonstrated understanding of the item and its complex requirements.

Focus

Focus items ask students to assess passage content in terms of relevance to the passage's topic and the local context in which the information appears. For item 6, a low-difficulty item associated with the passage about converting landfills into parks, students must consider the appropriateness of a potential sentence-length addition to the passage: "Interest in urban parks declined around the mid-twentieth century but has been revived during the past few decades." The paragraph in which the information would appear is the one previously discussed about the benefits to urban areas of converting landfills to parks. Students must first determine ("yes" or "no") whether the material should be added—it shouldn't be—and then determine which of the two "no" options is more germane—in this case, "No, because [the proposed addition] distracts from the focus of the paragraph by introducing irrelevant information."

After reading the stem to item 6 (including the proposed additional sentence), student 1WLCA rules out the three distractors in turn. The student first eliminates the two "yes" options that would add the sentence to the paragraph.

> Okay. A is "Yes, because it provides a detail that supports the main claim of the paragraph," which I'm going to cross off because I don't think it should be added because it does not provide a detail that supports the main claim. B is "Yes, because it effectively sets up the sentence that follows in the paragraph." It doesn't set up a sentence that follows in the paragraph—it is unrelated to what follows in the paragraph—so that's gonna get crossed off.

The student then evaluates the two "no" options, first ruling out one and then selecting the correct answer.

> C is "No, because it includes information that contradicts the main idea of the paragraph." That one I'm going to cross off because it doesn't contradict the main idea of the paragraph, but it does distract from it, which isn't quite right. And then D is "No, because it distracts from the focus of the paragraph by introducing irrelevant information." And I'm going to go with D, because it does introduce something that is irrelevant to the rest of the passage. So, it should not be added in.

The student's thinking aloud shows clear evidence of a conceptual and practical understanding of informational relevance in relation to an extended written context. In addition, the student is able to distinguish between information that contradicts other information presented and information that's merely tangential to the main point being made. In other words, in answering this item correctly, the student executes a complex sequence of first determining the baseline inappropriateness of the material potentially to be added and then, having successfully done that, ascertaining the correct reason for not including it.

Quantitative Information

Like their Reading Test counterparts, Writing and Language Quantitative Information items require students to deal with data in informational graphics, such as graphs and tables. For the Writing and Language Test, however, students must use the graphically displayed data to correct (or affirm the correctness of) the writer's verbal representation and, in some cases, to achieve defined rhetorical aims. Item 41, a medium-difficulty item associated with a careers passage on employee turnover, exemplifies the more conceptually complex end of the range of Quantitative Information items. Item 41 asks students to determine relevant (not just accurate) information from a table to illustrate a point, which is that a researcher "found that the turnover rate at the higher-paying club store" of two studied "was lower." Students must first determine from the table that company B, which pays an estimated average hourly wage of $17, is the "higher-paying club store" referred to in the sentence (as opposed to company A, which pays $10 per hour). Students then must decide that the sentence's point would best be supported with company B's annual full-time employee turnover rate—17 percent—and the fact that company B paid out less in turnover costs—an estimated $3,628 annually for each employee replaced (as opposed to company A's 44 percent turnover rate and $5,274 per-employee turnover cost).

Student 5WLNY begins their approach to the item by establishing the parameters of the correct answer from the associated table even before reading the options.

> So, the turnover rate at the higher paying club store. So, you're looking at the percentage, the turnover rate at the higher paying club store, which is company B, and 17%.

The student then addresses the four options, judging each in relation to the stem's criterion of relevance to supporting the point as well as the student's own sense of the right answer.

> "NO CHANGE" [option A] says "the firm's 67,600 full-time employees made an average of $17 per hour," but that's just saying that's how much they get paid. That's not supporting what is mentioned in the previous part of the sentence, where it's, like, "the turnover rate at the higher-

> paying [club] store[, however,] was lower," even though the values match up, like dollars and percent [i.e., company B pays $17 per hour and had a 17 percent turnover rate]. They're both 17. A is incorrect. "And its staff, 67,600 full-time employees, [was] significantly smaller" [option B]. This seems to be a really, really loosely related detail compared with the first part of the sentence. It's talking about the turnover rate, but B provides information about how many employees they have. So, that doesn't fit at all. But I'll leave it there. C, "17 percent, at a lesser cost of [$]3,628 [per] full-time employee"—that's the answer. D, "And it paid its full-time employees"—no. D is talking about pay just like A is, so they're both wrong. B, it's [the sentence is] talking about the turnover rate, but B brings up details about how many employees they employ. And that's extremely loose, that's very loosely related. Irrelevant. C.

As the student observes, each answer option is accurate per the table, but only one provides relevant information in support of the writer's point. When answering this item, the student keeps the writer's point—turnover rate at the higher-paying company—firmly in mind. As the student's response suggests, this item calls on students to demonstrate a series of complex behaviors: first, recognizing that relevance and not accuracy is key; second, integrating passage information with data displayed in a table; third, logically concluding that company B is the "higher-paying" company, as neither passage nor table outright says this; and fourth, determining that option C, which identifies the turnover rate and the associated turnover cost, is more relevant in context than the wages the company pays or the number of people the company employs.

Effective Language Use

Effective Language Use is another of the six broad knowledge and skill areas sampled by the SAT Writing and Language Test. The four specific item types composing this area—Precision, Concision, Style and Tone, and Syntax—pertain to a range of knowledge and skills fundamental to effective revision of prose text to enhance the rhetorical use of language in relation to the writer's purpose. (It's important to note that these items deal exclusively with the rhetorical aspects of word choice, not with grammar, usage, and mechanics.) Of these four types, two were selected for study. Precision items focus on using words and phrases to convey information and ideas in rhetorically effective, context-appropriate ways. These items aren't accompanied by stems; instead, students are expected to follow the overall test directions when answering these items, which refer in part to improving the quality of writing in passages. Style and Tone items draw on students' understanding of a passage's established style and tone to make selected text consistent with the established pattern (e.g., in the same register as the surrounding text) or to achieve specific rhetorical aims, such as creating or extending a sentence pattern (e.g., a series of short imperative sentences). These items may or may not be accompanied by stems, and both a stemmed and a stemless item were part of the studied sample. Performance on items in this area (both those types studied and those not) contribute to a Words in Context subscore (in combination with performance on select items from the Reading Test) and to an Expression of Ideas subscore yielded by the SAT.

To answer an Effective Language Use item as intended, students are expected to demonstrate the following behaviors, which, as noted below, vary depending on which of the two studied Effective Language Use types is being addressed:

1. Read and demonstrate comprehension of the local (sentence- or paragraph-level) context in which a given Effective Language Use question is situated.
2. Read and demonstrate comprehension of relevant portions of the passage, up to and including the whole passage (the scope being defined by the stem, if present; if a stem isn't present, by students' understanding of the task and the overall test directions).
3. Demonstrate an understanding of the framing language of the stem, if present (i.e., whether to add, revise, retain, or delete an expression); if a stem isn't present, demonstrate an understanding of the task and the overall test directions to make an effective revision decision.
4. Demonstrate, at least indirectly, a conceptual understanding of essayistic composition, of revision as a general process, and of using language in rhetorically effective ways as a specific process and outcome.
5. **Precision:** Demonstrate relevant vocabulary knowledge and an understanding of the local context in which an item is situated in the process of making an effective decision regarding revising an expression to improve exactness and/or context appropriateness.
6. **Style and Tone:** Demonstrate relevant knowledge of style and tone and an understanding of the local context in which an item is situated in the process of making an effective decision regarding revising an expression either to maintain or improve the consistency of the passage's established style and tone or to achieve a rhetorical purpose specified in the stem.

Considered as a group, the above behaviors require students to demonstrate a range of knowledge and skills involving revising or retaining word choice in relation to the writer's purpose, the content being conveyed, and the general and local contexts in which items are situated. In the process of answering items of the two studied types, students must exhibit the complex cognitive processes of understanding the contexts in which the items are located and using that understanding, along with vocabulary knowledge and skill, to make strategic choices about which words and phrases most skillfully accomplish the writer's purpose.

Table 7 summarizes student performance on the four studied Effective Language Use items. Sample (*n*) sizes vary, as noted, reflecting the fact that in certain cases the data were incomplete.

Table 7: Student Performance on Writing and Language: Effective Language Use Items

Item	Content Area	Item Difficulty	Demonstrated Required Behaviors						Answered Correctly	Demonstrated All Behaviors and Answered Correctly	Differential
Precision											
			1	2	3	4	5	All			
4 *n* = 30	History/ Social Studies	Easy	17 (57%)	17 (57%)	4 (13%)	18 (60%)	16 (53%)	4 (13%)	24 (80%)	4 (13%)	20
34 *n* = 29	Careers	Med	20 (69%)	20 (69%)	1 (3%)	16 (55%)	19 (66%)	1 (3%)	21 (72%)	1 (3%)	20
Style and Tone											
			1	2	3	4	6	All			
11 *n* = 30	History/ Social Studies	Med	21 (70%)	21 (70%)	19 (63%)	23 (77%)	13 (43%)	11 (37%)	21 (70%)	11 (37%)	10
21 *n* = 29	Humanities	Med	25 (86%)	25 (86%)	19 (66%)	24 (83%)	5 (17%)	5 (17%)	8 (28%)	5 (17%)	3

The data in table 7 offer a mixed picture with respect to whether the students answering the questions generally demonstrated complex behaviors when answering the Effective Language Use items and thereby enacted the items' intended design. In sixteen out of twenty cases (behaviors 1–6, as applicable, across four items), a majority of students demonstrated the required behaviors. In no case, however, did a majority of students demonstrate all required behaviors for a given item. Even if we exclude from discussion item 21, which the sample of students, as a whole, answered correctly at a low rate (which precludes the possibility that a majority of students could demonstrate all behaviors, since behavior 6 involves choosing the correct answer), none of the other three items, all of which were answered correctly by a majority of students, had a majority of students demonstrating all required behaviors—and item 34, a Precision item, had only one student demonstrate all required behaviors.

The two Precision items failed to meet the threshold because of the low rate at which students demonstrated behavior 3, which requires articulation of at least the general purpose of the item. This articulation proved to be tricky at least partly as a consequence of the fact that the items by design lacked stems. Item 11, the Style and Tone item that a majority of participating students answered correctly, also lacked a stem but met the threshold for behavior 3, so there may be something in the nature of the Precision items that was conceptually unclear to the students, even as they answered both Precision items correctly at high rates. Item 11 itself failed to elicit demonstration of behavior 6 from a majority of students, due primarily to the relatively small proportion of students offering a clear item type–based (style and tone) rationale for their choice of answer. Given the above, it's not surprising that the three items discussed above evinced substantial gaps between the number of students who answered the item correctly and the number who both answered correctly and demonstrated all required behaviors.

Taken together, these findings suggest some reason for concern about how these three studied items function, if, as seems reasonable, one would expect that students ought to be able to demonstrate an understanding of what the items

are asking (in the case of Precision items) and to be able to give a rationale for their answer choice in terms of the content focus of the item (in the case of the stemmed Style and Tone item).

Nonetheless, vignettes from responses to the two studied Effective Language Use item types provided by students who answered correctly and demonstrated all required behaviors offer further evidence that these items are capable of eliciting complex cognitive processes. The following sections address each of the two item types in turn.

Precision

Precision items require students to use their vocabulary knowledge and skills and an understanding of context and the writer's purpose to make an effective word choice. Answer options are typically a series of synonyms and/or related words, and the best answer is the one that communicates most clearly and appropriately in context. Item 4, a statistically easy item based on the previously discussed history/social studies passage about landfill conversion, is typical of the type: students must determine whether widely available landfill land is "prolific," "abundant," "magnanimous," or "excessive." The four options are semantically related, but only "abundant" makes good sense in context. "Prolific" suggests fruitfulness, inventiveness, or productivity, which doesn't fit here. "Magnanimous" suggests generosity, which is nonsensical in this situation. "Excessive" used to describe the availability of the land from at least ten thousand closed municipal landfills in the United States makes sense in isolation but carries a negative judgment opposed to the point of view expressed in the passage. Student 6WLNY, whom we quoted earlier, was among the four students who demonstrated all five required behaviors, though their explanation of the item's intent (behavior 3) indicated merely that the issue was whether "prolific," which appears as an underlined portion in the passage, should be changed. In other respects, the student showed cogent reasoning and strong vocabulary knowledge while working through the item.

> "Magnanimous" [option C], just eliminate that from the get-go because that makes no sense. How can land be "generously kind and great"? "Prolific" [option A] implies that it is producing a lot of things—like a prolific author, prolific researcher, or something—and this doesn't really, it's not really relevant. So, you'd take off "magnanimous" and "prolific," so that leaves us with "abundant" [option B] and "excessive" [option D]. "Excessive" is kind of a—is an opinionated word because [the passage] says the U.S. has . . . at least 10,000 [closed] landfills and, well, what defines "excessive"? However, you can quantify "abundant" more because you can say, "There are lots of landfills." But somebody else might think, "Well, is that excessive?" So, I would go with option B, so it is also "abundant" because there's an abundancy of landfills that can be repurposed.

It should be noted that, unlike student 6WLNY, fourteen of the sampled students indicated that they didn't know the meaning or have a clear enough sense of "magnanimous" and/or "prolific" to be confident in ruling one or both of these options in or out. Since vocabulary is expressly part of the skill being measured, this doesn't represent an inherent threat to the item's validity, and the answer

options are all at the same general level of diction, but it likely did have some impact on students' ability to demonstrate the required behaviors.

Style and Tone

For Style and Tone items, students must maintain a tonal or stylistic pattern established in a given passage. As is the case with both items 11 and 21 in the studied sample, many Style and Tone items assess consistency of diction relative to the level of formality of the overall passage. For item 21, a medium-difficulty item associated with the humanities passage on Arthur Conan Doyle and Sherlock Holmes, students must recognize that the passage is relatively formal and that it's more tonally appropriate to say that Conan Doyle "relented" to pressure to resurrect Holmes rather than say that Conan Doyle "caved," "sold out," or "loosened up." Note that this item, unlike the other studied Style and Tone item (item 11), included a stem ("Which choice best maintains the tone of the passage?"), though, as noted earlier, a majority of students was able to articulate at least some sense of each item's purpose, stem or no. Student 1WLKY was one of the five students to demonstrate all five required behaviors for item 21, providing tone-based arguments rooted in an understanding of the broader context in which the tested word/phrase appears.

> "After nearly eight years [of pressure from fans], however, Conan Doyle caved." "Caved" doesn't sound right here so it's not A. B, "sold out," is not the right term he would use in this kind of passage. C, "loosened up," is also not consistent with the tone of the paragraph. So, I'm going to go with D, "relented." That's the most formal way of answering the question.

As the above clearly indicates, the student drew on both an understanding of the paragraph's tone and a developed vocabulary to make a reasoned choice about the most appropriate option. Answering the question correctly isn't simply a matter of a low-level understanding but rather requires a nuanced sense of both the local and global passage contexts in conjunction with the ability to apply vocabulary knowledge and skill.

Math

Student responses to thirty-four SAT Math Test items were studied using the cognitive lab methodology. The Math items represented knowledge and skills in five areas of central importance to and prominence in the test, as discussed in the subsequent sections and as summarized in table 8. SAT Math items include both ones for which a calculator is prohibited and ones for which a calculator is allowed, as well as both multiple-choice items and items for which students must generate their own response. Three items were evaluated twice because they drew on knowledge and skills in two of the five areas.[10] Note that unlike for the ERW items, no empirical item difficulty data are available for the Math items.

[10] Item WC2 was evaluated as both an algebra and a functions item. Item WC16 was evaluated as both a geometry and a ratios, proportions, and percentages item. Item WC12 was evaluated as both a ratios, proportions, and percentages item and a statistics and probability item.

Table 8: Breakdown of Math Items by Category

Category	Number and Types of Items
Algebra	9 total 4 no-calculator, 5 with-calculator 8 multiple-choice, 1 student-produced response
Functions	7 total 4 no-calculator, 3 with-calculator 7 multiple-choice
Geometry	6 total 3 no-calculator, 3 with-calculator 4 multiple-choice, 2 student-produced response
Ratios, Proportions, and Percentages	8 total 3 no-calculator, 5 with-calculator 5 multiple-choice, 3 student-produced response
Statistics and Probability	7 total 2 no-calculator, 5 with-calculator 8 multiple-choice, 1 student-produced response

Important methodological distinctions between the ERW and Math coding and analyses, and hence the results, must be observed. ERW required behaviors were conceived as components of successful item response relative to the construct. This means, first, that individual students were expected to demonstrate all behaviors and, second, that, as previously noted, students needed to answer correctly in order to demonstrate all the behaviors (since at least one behavior per item type was associated with determining the correct answer). Math behaviors, by contrast, represent sets of strategies—often mutually exclusive—that students as a whole would be expected to demonstrate in answering the items but that no individual student would demonstrate in totality; additionally, students could demonstrate one or more of the expected strategies and still answer incorrectly (e.g., due to a computational error). As a result, while in the ERW tables above the number and percentage of students in the sample demonstrating "all" behaviors was tabulated, for Math the number and percentage of students in the sample demonstrating "one or more" behaviors was tabulated. In addition, while in ERW the number/percentage of students demonstrating all behaviors and the number/percentage answering correctly and demonstrating all behaviors is necessarily the same (since answering correctly, as noted, is tied into the set of required behaviors), in Math the number/percentage of students in these two columns may differ.

Moreover, the meaning of Math's "differential" column is somewhat different from that of ERW's. While the ERW differential represents the difference between the number of students who answered a given item correctly and the number who also demonstrated all required behaviors (and thus also answered correctly), the Math differential represents the difference between the number of students who answered a given item correctly and the number who also demonstrated one or more required behaviors (and thus demonstrated at least one requisite strategy). The differential is still meaningful in Math and generally comparable in its interpretation to that in ERW because both the ERW and Math differentials offer indications of the number of students who answered a particular item correctly while sidestepping the intended construct (in the case of Math, not demonstrating use of at least one expected strategy).

ALGEBRA

Algebra knowledge and skills are prerequisite for many college degree programs and workforce training programs. The emphasis states place on algebra in their college and career readiness standards (e.g., California Department of Education 2013; Board of Education, Commonwealth of Virginia 2016; Minnesota Department of Education 2007) indicates this content is considered important for postsecondary readiness. Gaertner, Kim, DesJardins, and McClarty (2014), using national data sets, and Long, Iatarola, and Conger (2009), using a state database, concluded that completion of Algebra II in high school is a strong indicator of college readiness. In a study conducted with high school graduates, Hart Research Associates and Public Opinion Strategies (2005) found that the percentage of graduates who felt prepared for college and work was much higher for those who completed Algebra II than for those who didn't. An understanding of algebra is critical for success in a variety of pursuits, such as a career in the sciences, engineering, statistics, mathematics, or the social sciences (U.S. Department of Education 2018). The SAT Math Test samples content aligned to algebra knowledge and skills. Performance on these items contributes to the SAT Math test score and may contribute to the Heart of Algebra or Passport to Advanced Math subscores.

To answer the algebra items selected for the study as intended, students are expected to demonstrate one or more of the following behaviors, the specific requirements varying by item focus (as indicated below):

1. Solve a nonlinear equation.
2. Evaluate a linear or nonlinear algebraic expression.
3. Make connections between a graph and an equation by using data or patterns from a graph to identify an equation of a line.
4. Make strategic use of structure to solve a linear equation.
5. For a linear equation in two variables that represents a context, use the value of a known quantity to determine the value of an unknown quantity.
6. For a system of linear equations in two variables, demonstrate understanding of the conditions under which the system has no solution, a unique solution, or infinitely many solutions.

Collectively across a set of items, these behaviors represent complex cognitive processes requiring students to represent relationships symbolically, recognize structure in these symbolic representations, and demonstrate fluency in manipulating and evaluating these symbolic representations. In the process of answering the studied items, students must exhibit the complex cognitive processes of purposefully analyzing and applying algebraic structure and using the information in a symbolic representation to gain insight into the context that is represented.

Table 9 summarizes student performance on the nine studied algebra items on both no-calculator (NC) and with-calculator (WC) items. Sample (*n*) sizes vary, as noted, reflecting the fact that in certain cases the data were incomplete (e.g., from one or more students not providing an answer).

Table 9: Student Performance on Math: Algebra Items

Item	Demonstrated Expected Behavior(s)							Answered Correctly	Demonstrated One or More Behaviors and Answered Correctly	Differential
	1	2	3	4	5	6	One or More			
NC3 *n* = 37	34 (92%)	N/A	N/A	N/A	N/A	N/A	34 (92%)	32 (84%)	31 (84%)	1
NC6 *n* = 38	N/A	N/A	33 (87%)	N/A	N/A	N/A	33 (87%)	30 (79%)	28 (74%)	2
NC13 *n* = 38	35 (92%)	N/A	N/A	N/A	N/A	N/A	35 (92%)	20 (53%)	20 (53%)	0
NC16 *n* = 36	33 (92%)	33 (92%)	N/A	N/A	33 (92%)	N/A	34 (94%)	17 (47%)	17 (47%)	0
WC2[a] *n* = 38	N/A	N/A	34 (89%)	N/A	N/A	N/A	34 (89%)	34 (89%)	34 (89%)	0
WC6 *n* = 37	N/A	N/A	N/A	N/A	35 (95%)	N/A	35 (95%)	26 (70%)	26 (70%)	0
WC7 *n* = 38	N/A	N/A	N/A	N/A	N/A	26 (68%)	26 (68%)	16 (42%)	15 (39%)	1
WC8 *n* = 38	35 (92%)	35 (92%)	N/A	N/A	34 (89%)	N/A	35 (92%)	36 (95%)	35 (92%)	1
WC9 *n* = 37	N/A	N/A	N/A	27 (73%)	N/A	N/A	27 (73%)	33 (89%)	27 (73%)	6

[a] Also evaluated as a functions item

The data in table 9 suggest that participating students in general were able to demonstrate complex behaviors when answering algebra items and enacted the items' intended design. In all cases (behaviors 1–6, as applicable, across nine items), a majority of students demonstrated a given behavior. Consequently, for all nine items, a majority of students demonstrated one or more of the expected behaviors. For eight of the nine items, the difference between the number of students who answered the items correctly and the number of students who, in addition, demonstrated one or more of the expected behaviors was two or fewer, with the modal difference being zero (four cases). The exception was item WC9, where the differential was six. In answering this item, several students took an inefficient approach that required more time and effort yet yielded the correct answer for persistent students. Using these methods demonstrated competence without the insight to render the task a direct application of algebraic structure.

Vignettes from students in the study who demonstrated the expected behavior(s) associated with algebra items offer evidence of the cognitive challenge these items pose. Item NC3, for example, allows students to use the definition of absolute value to solve a nonlinear equation with greater efficiency by recognizing structure and using it to evaluate an algebraic expression, as done by student 10MCA:

> Firstly, I can realize there is no solution because absolute value must be positive, and the solution would not be true in any case.

By contrast, student 14MCA applies skills in a manner that displays competence but with less mastery and efficiency:

> So, I would pretty much change it to $P + 1$ because the absolute value of -1 is $+1 = -5$ and then I would move the $+1$ to the right side, which would give me $P = -1 - 5$. So, $-1 - 5$ would equal -6. So, $P = -6$. I messed up. In reality, I don't think there is a solution like with all the answer choices given because $-5 - 1$ would be -6 and then when you plug that into the absolute value equation, you end up with 6 and that does not equal -5 and -1 is obviously too small. And $6 - 1$ is 5 and the absolute of 5 is 5, not -5. So, I would say there is no solution to the equation. So, I would put D as a choice.

As another example, in solving item WC2, student 12MKY demonstrates expected behavior 3 by making connections between the graph and the equation, as well as showing competence by using additional strategies not specifically expected:

> So, on the graph, we have the y-intercept at (0, 1). It's a positive since it's going from bottom left to top right. I'm just gonna plug in 0 to the x's of the options of the answers and see if it gives me the y-value of 1. So, for A, $f(x) = 1/3x + 3$. One, I'm gonna type in 1/3 then multiply by 0, and then that gives me 0. And then $0 + 3 = 3$, so the $f(x)$ for option A is 3. But that is not y, so A is not the answer.
>
> For B, $f(x) = x + 3$, so I will plug in 0 for x, so that will make it $0 + 3 = 3$. 3 is not 1, so B is not the answer. C, $3x + 1$, so I plug in the 0. $3(0) = 0$, and then $0 + 1 = 1$, so C could be the answer. For D, $4x + 1$. I plug in the 0 into the function. So that means it's $4(0) + 1$. $4(0) = 0$, so $0 + 1$, so that means . . . D could also be the answer. So that's C to D. And then, there's a point on the graph where the line crosses, and that point is (1, 4). So then I will simply plug in those numbers. So, the 1, the x value, I will plug into the C options, the $3x + 1$. So, $3(1) + 1 = 3 + 1 = 4$. Which is the y-value of the line representing the graph? So, C could, again, be the answer. And then if I do D, if I plug in the number, $4(1) + 1$, so that means there is $4 + 1 = 5$, but 5 is not a 4, so D is not the answer. So, the answer is C for number two.

FUNCTIONS

Mathematical functions appear in high school math courses such as Algebra I, Geometry, and Algebra II and in integrated curricular pathways that blend content from these courses. Research identifies a relationship between these high school courses and college readiness (Gaertner, Kim, DesJardins, and McClarty 2014; Hart Research Associates and Public Opinion Strategies 2005; Klepfer and Hull 2012; Long, Iatarola, and Conger 2009). Skills with functions are included within all state standards (e.g., California Department of Education 2013; Board of Education, Commonwealth of Virginia 2016; Minnesota Department of Education 2007). An understanding of functions is critical for success in a variety of pursuits, such as careers related to science, engineering, statistics, mathematics, or the social sciences (Gaertner, Kim, DesJardins, and McClarty 2014). The SAT Math Test samples skills in the area of functions that are deemed essential for college readiness as identified by a curriculum survey of postsecondary instructors (Kim, Wiley, and Packman 2012; College Board 2019). Performance on these items contributes to the overall Math test score and may contribute to the Heart of Algebra or Passport to Advanced Math subscores.

To be successful in answering the functions items selected for the study as intended, students are expected to demonstrate one or more of the following behaviors, the specific requirements varying by item focus (as indicated below):

1. Interpret and use symbolic notation.
2. Calculate input or output values.
3. Interpret a solution, constant, variable, factor, or term based on context.
4. Identify an equation that models a relationship between quantities.
5. Make connections between a table, an algebraic representation, or a graph of a function by using an alternative representation to solve a problem.

Considered as a group and across the selected items, the above behaviors require students to demonstrate a range of knowledge and skills involving creating, identifying, interpreting, and applying representations of relationships between quantities. In the process of answering the studied items, students must exhibit the complex cognitive processes of analyzing relationships between quantities, understanding the various representations of these relationships, and using the structure of a relationship to gain a deeper understanding of the mathematics of the relationship and the information it encodes about a context.

Table 10 summarizes student performance on the seven studied functions items. Sample (*n*) sizes vary, as noted, reflecting the fact that in certain cases the data were incomplete.

Table 10: Student Performance on Math: Functions Items

	Demonstrated Expected Behavior(s)								
Item	1	2	3	4	5	One or More	Answered Correctly	Demonstrated One or More Behaviors and Answered Correctly	Differential
NC1 *n* = 37	37 (100%)	36 (97%)	N/A	N/A	N/A	37 (100%)	34 (92%)	34 (92%)	0
NC4 *n* = 37	27 (73%)	N/A	N/A	N/A	N/A	27 (73%)	18 (49%)	17 (46%)	1
NC8 *n* = 38	30 (79%)	N/A	31 (82%)	N/A	N/A	32 (84%)	30 (79%)	29 (76%)	1
NC9 *n* = 38	35 (92%)	N/A	N/A	N/A	22 (58%)	35 (92%)	37 (97%)	37 (97%)	0
WC1 *n* = 38	32 (84%)	30 (79%)	N/A	31 (82%)	N/A	32 (84%)	33 (87%)	30 (79%)	3
WC2[a] *n* = 38	35 (92%)	N/A	N/A	35 (92%)	6 (16%)	35 (92%)	34 (89%)	34 (89%)	0
WC3 *n* = 38	34 (89%)	N/A	34 (89%)	32 (84%)	N/A	34 (84%)	31 (82%)	31 (82%)	0

[a] Also evaluated as an algebra item

The data in table 10 suggest that, in general, participating students were able to demonstrate complex behaviors when answering functions items and enacted the items' intended design. In fifteen out of sixteen cases (behaviors 1–5, as applicable, across seven items), a majority of students demonstrated a given behavior. For all seven items, a majority of students demonstrated one or more of the expected behaviors. For the seven items, the discrepancy between the number of students who answered the items correctly and the number of

students who both answered the items correctly and demonstrated one or more of the expected behavior(s) was three or fewer, with gaps of zero (four items), one (two items), and three (one item), which offers further evidence supporting the conclusion that these items generally can't be answered correctly without demonstrating certain behavior(s).

Vignettes from students in the study who demonstrated the expected behavior(s) associated with functions items offer evidence of the items' complex cognitive demands. Item NC8 requires students to directly identify the meaning of a term in a symbolically presented relationship. This is accomplished by student 4MKY:

> So 61, if you look at the equation, it's in the same form as the slope intercept equation $y = mx + b$, where 61 takes the place as m and m is the slope. In the context of this problem, it would be the change in the number of birds per year. The average decrease in the population per year from 1962 to k years after 1962.

In solving item WC3, student 4MKY also demonstrates rich use of and interconnections among expected behaviors 1, 3, and 4:

> So, we already know that no matter how many pages the book is, it's going to at least be 2 millimeters thick. So that means that if we were to make a graph of this, the y-intercept would be 2. So I'm looking for an answer choice that doesn't have—that, oh. The question states that the front cover and the back cover are each 2 millimeters thick. That means that if you had 0 pages, it would be at least 4 millimeters thick. Meaning that we're looking for an equation that has the value 4 that is not attached to any variable. Meaning that B, C, and D are out. Meaning that A, $f(n) = 4 + .1n$ is the answer.

Item WC2 requires students to interpret and use symbolic notation to identify an equation that models a graphical relationship. These behaviors are demonstrated by student 8MNY:

> So, I take the y-intercept, which is at (0, 1), and then you see it also intercepts at point (2, 7), so I'm going to—since all of these are linear form, I'm just going to find $y = mx + b$, and y is over 1, that's 7 over 2, so y is over—the 7 over $2x$ plus 1. Sorry, 6, y is 6, so it becomes 6 over 2. So, y is equal to $3x + 1$, which is choice C.

GEOMETRY

Geometry knowledge and skills appear across all state standards (e.g., California Department of Education 2013; Board of Education, Commonwealth of Virginia 2016; Minnesota Department of Education 2007) as either the focus of a dedicated course in traditional curricular pathways or as clusters of standards in the courses of integrated pathways. The aggregate geometry content in high school is comparable regardless of approach and consistent with the geometry assessed on the SAT: area and volume; properties of angles, parallel lines, and triangles and other polygons, including similarity and congruence, and deductive reasoning and proof; the Pythagorean theorem and right-triangle trigonometry; and properties of circles, including unit-circle trigonometry. Although research into the connection between geometry knowledge and skills and college and career readiness isn't as developed as for the analogous connections between algebra,

functions, ratio, proportions, and percentages, and probability and statistics and college and career readiness, geometry has many connections with these areas of math as well as connections with spatial reasoning, measurement, and modeling (Battista 2007). Performance on SAT Math Test items in geometry contributes to the Math test score and may also contribute to the Heart of Algebra or Passport to Advanced Math subscores.

To answer the collection of geometry items chosen for this study as intended, students are expected to demonstrate the following behaviors, the specific requirements varying by item focus (as indicated below):

1. Match a property to a definition.
2. Use different properties of the same figure.
3. Recognize and apply the connection between different figures.
4. Apply a definition of volume to find an unknown length.
5. Identify information needed to apply a theorem.
6. Identify a theorem or definition that provides a line between given and requested information.
7. Recognize that calculating volume is needed to find requested information.

Considered as a group across the set of items, the above behaviors require students to demonstrate a range of knowledge and skills involving applying the definitions of, properties of, and theorems about geometric figures to derive further knowledge about these figures, including measurements involving these figures. In the process of answering the studied items, students must exhibit the complex cognitive processes of recognizing and synthesizing properties of geometric objects to gain more knowledge of these figures and contexts the figures represent.

Table 11 summarizes student performance on the six studied geometry items. Sample (*n*) sizes vary, as noted, reflecting the fact that in certain cases the data were incomplete. (This was particularly an issue with item WC18—the last Math item studied—for which nine students provided no answer.)

Table 11: Student Performance on Math: Geometry Items

	Demonstrated Expected Behavior(s)										
Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**One or More**	**Answered Correctly**	**Demonstrated One or More Behaviors and Answered Correctly**	**Differential**
NC10 *n* = 38	35 (92%)	N/A	N/A	N/A	N/A	N/A	N/A	35 (92%)	35 (92%)	35 (92%)	0
NC11 *n* = 38	N/A	N/A	N/A	N/A	25 (66%)	N/A	N/A	25 (66%)	21 (55%)	20 (53%)	1
NC12 *n* = 37	N/A	33 (89%)	N/A	N/A	N/A	N/A	N/A	33 (89%)	31 (84%)	30 (81%)	1
WC4 *n* = 37	N/A	N/A	N/A	28 (76%)	N/A	N/A	N/A	28 (76%)	20 (54%)	19 (51%)	1
WC16[a] *n* = 35	N/A	N/A	32 (91%)	N/A	N/A	N/A	31 (89%)	32 (91%)	23 (66%)	23 (66%)	0
WC18 *n* = 29	N/A	N/A	N/A	N/A	N/A	24 (83%)	N/A	24 (83%)	25 (86%)	23 (79%)	2

[a] Also evaluated as a ratios, proportions, and percentages item

The data in table 11 suggest that, in general, participating students were able to demonstrate complex behaviors when answering geometry items and enacted the items' intended design. In all seven cases (behaviors 1–7, as applicable, across six items), a majority of students demonstrated the given behavior. Consequently, for all six items, a majority of students demonstrated one or both expected behaviors. For none of the six items was there a discrepancy of greater than two between those who answered the item correctly and those who both answered the item correctly and demonstrated at least one of the expected behaviors, and the modal difference was one (three cases); this outcome suggests that to answer the items correctly, students must demonstrate the expected behavior(s), further supporting the conclusion that the items performed as designed.

Vignettes from students in the study who demonstrated the expected behavior(s) associated with geometry items offer evidence of the complex cognition these items can elicit. Students demonstrated these behaviors across a range of cognitive complexity. For example, in solving item NC10, student 11MNY draws a figure to organize information (an optional and often helpful, purposeful behavior that helped many of the students complete the tasks) and then uses that figure to match a property to a definition:

> For 10 it says the lengths of the sides of triangle *DEF* are *DE* = 3; *EF* = 3, and *DF* = 5. What can be proven about this triangle? So, if I draw *DEF* quickly, on my paper, I see that *DE* is 3; *EF* is 3, and *DF* is 5. So, I know for a fact that this is isosceles because that means I have two of the angles are—two [of] the sides [are] the same.

Student 2MKY shows the behaviors of using different properties of the same figure (the expected behavior) and identifying information needed to apply a theorem (an additional behavior) in order to solve item NC12:

> In triangle *ABC* above, side *AC* is extended to point *D*. What is the value of $y - x$? So, you know that 180—there's 180° in a triangle, and you know two of the angles are 35 and 105 degrees, so if you add those together you get 140°. So, if you subtract that from 180, you get 40°. So, you know that angle x is 40°, and a line has 180° in it, so angle y has to be—well, you know that $x + y$ must be 180. You know that x is 40. So, $40 + y = 180$. Subtract 40 from both sides. You get $y = 140$. And you are finding $y - x$. So, that is $140 - 40$, which is 100°. So, the answer is C.

RATIOS, PROPORTIONS, AND PERCENTAGES

The concepts of ratio, proportion, and percentage are major themes of middle school mathematics. Proficiency in ratio, proportion, and percentage is also essential for success in many areas of high school mathematics, such as the study of functions, and is associated with college and career readiness as identified by curriculum surveys of postsecondary instructors (Kim, Wiley, and Packman 2012; College Board 2019). The National Center on Education and the Economy (2013) found that students pursuing two-year degree programs must be able to work with multistep problems involving ratios, proportional relationships, percentages, unit conversions, and complex measurement problems. Quantitative literacy

is part of participation in a democracy; it's important to employers, who need employees who can use mathematics outside of the classroom; and it's important not only for science, technology, engineering, and mathematics (STEM) fields but also for a wide range of college majors (Steen 2001). Students pursuing a variety of degree programs and careers must be able to work with multistep problems involving ratios, proportional relationships, percentages, unit conversions, and measurement. Therefore, the SAT Math Test includes questions that assess these important skills. Performance on ratios, proportions, and percentages items contributes to the Math test score and also to the Problem Solving and Data Analysis subscore.

To answer the sampled items on ratios, proportions, and percentages as intended, students are expected to demonstrate the following behaviors, the specific requirements varying by item focus (as indicated below):

1. Identify and represent a proportional relationship described within an item.
2. Differentiate between part-to-whole and part-to-part relationships and convert between these two types of relationships.
3. Identify the correct mathematical operations, such as multiplication and division, that are used to represent verbally described relationships.
4. Keep track of units associated with measurements.
5. Identify and represent the relationships between quantities to compute percentages and percent change.
6. Demonstrate proportional reasoning.

Considered as a group across the set of items, the above behaviors require students to demonstrate a range of knowledge and skills involving identifying when the ratio between two quantities is constant, analyzing proportional relationships, and analyzing part-to-whole and part-to-part relationships, including percentages. In the process of answering the studied items, students must exhibit the complex cognitive processes of analyzing, applying, and synthesizing basic proportional relationships to gain insights into these relationships and the contexts they represent.

Table 12 summarizes student performance on the eight studied ratios, proportions, and percentages items. Sample (*n*) sizes vary, as noted, reflecting the fact that in certain cases the data were incomplete.

Table 12: Student Performance on Math: Ratios, Proportions, and Percentages Items

Item	Demonstrated Expected Behavior(s)							Answered Correctly	Demonstrated One or More Behaviors and Answered Correctly	Differential
	1	2	3	4	5	6	One or More			
NC2 *n* = 37	28 (76%)	N/A	30 (81%)	N/A	N/A	N/A	28 (76%)	30 (81%)	29 (78%)	1
NC5 *n* = 38	N/A	N/A	8 (21%)	N/A	N/A	8 (21%)	11 (29%)	11 (29%)	8 (21%)	3
NC15 *n* = 37	17 (46%)	N/A	32 (86%)	26 (70%)	N/A	N/A	33 (89%)	28 (76%)	28 (76%)	0
WC5 *n* = 37	13 (35%)	N/A	29 (78%)	19 (51%)	N/A	N/A	30 (81%)	30 (81%)	29 (78%)	1
WC12[a] *n* = 37	N/A	N/A	13 (35%)	N/A	15 (41%)	N/A	15 (41%)	14 (38%)	12 (32%)	2
WC14 *n* = 38	N/A	N/A	31 (82%)	N/A	29 (76%)	N/A	32 (84%)	30 (79%)	29 (76%)	1
WC16[b] *n* = 35	N/A	N/A	30 (86%)	14 (40%)	N/A	N/A	30 (86%)	23 (66%)	23 (66%)	0
WC17 *n* = 35	26 (74%)	18 (51%)	33 (94%)	N/A	N/A	20 (57%)	33 (94%)	28 (80%)	28 (80%)	0

[a] Also evaluated as a statistics and probability item
[b] Also evaluated as a geometry item

The data in table 12 suggest that participating students were generally able to demonstrate complex behaviors when answering ratios, proportions, and percentages items and enacted the items' intended design. In thirteen of twenty cases (behaviors 1–6, as applicable, across eight items), a majority of students demonstrated the given behavior. A majority of students was able to demonstrate one or more expected behaviors for all but two of the eight items (NC5, WC12). The difference between the number of students who answered a given item correctly and the number who also demonstrated one or more expected behaviors was between zero and three (a gap of zero in three cases, a gap of one in three cases, a gap of two in one case, and a gap of three in one case).

Vignettes from students in the study who demonstrated the expected behavior(s) associated with ratios, proportions, and percentages items offer evidence that these items are capable of evoking complex cognition. In solving item NC2, student 11MCA efficiently identifies the proportional relationship presented in the item and solves the item with proportional reasoning combined with fluent use of the correct mathematical operations:

> The ratio of a:b is equivalent to the ratio of 2:3. If the value of a is 12, then what is the value of b? So, $a/b = 2/3$ so that's what I wrote down. If the value of a is 12 then what of b? So, I'll cross out a here and I'll put 12. So, $12/b = 2/3$. So, then I'll cross multiply. So, $2b = 36$. So, $b = 18$. That's D.

Student 1MNY successfully completes the challenging task presented in item WC12, correctly recognizing that the item was about identifying the appropriate percentages and the proportional relationships they represent and representing

these relationships by applying the correct relationship to the values given in a data display:

> So, okay. So, this is greatest number and the graph is percentage. So, let's see. It would—just eyeballing it. So, A is 10,000 about—about 75% in favor of the proposition. So, that would be 7.5 thousand. So, A is 7.5 thousand. B is 60%, and they have 17,000. So, I just want to do that math. So, 17,000—0.6. So, it's—okay. B is 10.2 thousand. C has, what, 55—they have 22%—I mean, 22,000. So, 55% times 22 is 21—I mean, 12.1 thousand. And D has 26, and they have exactly 50%. So, they would have 13,000. So, D is the largest. So, that's the answer.

STATISTICS AND PROBABILITY

Statistics and probability knowledge and skills are included in middle school and high school courses and are identified as important topics covered in high school courses as well as higher education courses by secondary and postsecondary instructors (Kim, Wiley, and Packman 2012; College Board 2019). Researchers haven't definitively stated which statistics and probability knowledge and skills are strong indicators of college readiness. The authors of the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report* argue that "statistical literacy is essential in our personal lives as consumers, citizens, and professionals" (Franklin et al. 2007, 3). Shaughnessy states that "there is perhaps no other branch of the mathematical sciences that is as important [as statistics and probability] for *all* students, college bound or not" (1992, 466; emphasis in original). Probability is an essential tool for statistics. It informs the questions that statisticians ask, it influences how they collect data, it may be part of their analysis of collected data, and it's used to help interpret results. The SAT Math Test samples content aligned to knowledge and skills from the domain of statistics and probability. Performance on these items contributes to the Math test score and may contribute to the Problem Solving and Data Analysis subscore.

To answer the sampled statistics and probability items as intended, students are expected to demonstrate the following behaviors, the specific requirements varying by item focus (as indicated below):

1. Compare distributions using measures of center and spread, where the distributions have the same mean and different standard deviations.
2. Analyze and interpret data represented in a scatterplot or line graph; fit linear, quadratic, and exponential models.
3. Using a model that fits the data in a scatterplot, compare values predicted by the model to values given in the data set.
4. With random samples, describe which population the results can be extended to; understand why a result can be extended only to the population from which the sample was selected.
5. Use one- and two-way tables, tree diagrams, area models, and other representations to find relative frequency, probabilities, and conditional probabilities; understand formulas for probability, and conditional probability in terms of frequency.

6. Calculate, express, or interpret the probability or conditional probability of an event using, for example, data displayed in a two-way table.

Considered as a group across the set of items, the above behaviors require students to demonstrate a range of knowledge and skills involving analyzing data to draw conclusions and make inferences. In the process of answering the studied items, students must exhibit the complex cognitive processes of identifying and describing trends in data and variations from these trends and calculating the probability of an event.

Table 13 summarizes student performance on the seven studied statistics and probability items. Sample (*n*) sizes vary, as noted, reflecting the fact that in certain cases the data are incomplete.

Table 13: Student Performance on Math: Statistics and Probability Items

	Demonstrated Expected Behavior(s)									
Item	**1**	**2**	**3**	**4**	**5**	**6**	**One or More**	**Answered Correctly**	**Demonstrated One or More Behaviors and Answered Correctly**	**Differential**
NC7 *n* = 37	26 (70%)	N/A	N/A	N/A	N/A	N/A	26 (70%)	7 (19%)	6 (16%)	1
NC14 *n* = 37	N/A	37 (100%)	N/A	N/A	35 (95%)	36 (97%)	37 (100%)	31 (84%)	31 (84%)	0
WC10 *n* = 38	N/A	38 (100%)	N/A	N/A	N/A	N/A	38 (100%)	37 (97%)	37 (97%)	0
WC11 *n* = 38	N/A	N/A	N/A	30 (79%)	N/A	N/A	30 (79%)	21 (55%)	19 (50%)	2
WC12[a] *n* = 37	N/A	28 (76%)	N/A	N/A	N/A	N/A	28 (76%)	14 (38%)	14 (38%)	0
WC13 *n* = 38	N/A	38 (100%)	36 (95%)	N/A	N/A	N/A	38 (100%)	32 (84%)	32 (84%)	0
WC15 *n* = 38	N/A	N/A	N/A	N/A	N/A	36 (95%)	36 (95%)	20 (53%)	20 (53%)	0

[a] Also evaluated as a ratios, proportions, and percentages item

The data in table 13 suggest that, in general, participating students were able to demonstrate complex behaviors when answering statistics and probability items and enacted the items' intended design. In all ten cases (behaviors 1–6, as applicable, across seven items), a majority of students demonstrated the given behavior. For all seven items, a majority of students demonstrated one or more of the behaviors expected for a given item. For none of the seven items was there a discrepancy of greater than two between those who answered the item correctly and those who both answered the item correctly and demonstrated one or more of the expected behaviors, and the modal difference was zero (five cases).

Vignettes from students in the study who demonstrated the expected behavior(s) associated with statistics and probability items offer evidence of the items' complex cognitive demands. In answering item WC12, student 8MNY demonstrates the expected behavior of analyzing and interpreting data in a scatterplot, while also identifying relationships and computing percentages, and the additional behavior of using the correct mathematical operations to represent relationships:

> [S]o, we're given the number of people who voted, and the percent that agree with it. So, to find the total amount, we just have to multiply by the decimal of the percent. So, for A we get that it's around 11,000 times 72%, somewhere around there. For B, we get 58%, so 0.58 times 17,000. For C, we get, it's 22,000 and they're giving us around 55. And for D we get 26,000 times 0.5, so D becomes 13,000 people who liked it. For C, we get 12,100. For A we get 0.72, 11,000, 11,000 times 0.72. We get 7,920. And for B, we get 0.58 times 17,000, which is 0.58 times 17,000, which is—three, no. 0.58 times 17,000. . . . 0.58 times 17,000 is 9,860. So, D has the most, with 13,000, which is choice D.

Student 12MKY uses a model and analyzes its fit to a scatterplot to solve item WC13:

> Okay, on the line of the best fit, it crosses—on 2003, it crosses the point 4,500, the number of home runs. But the actual dot passes 5,000, but it's below 5,400. So, I would say it's about 5,250 is the actual number. But then the line of best fit says it should be 4,500, so I'm going to use a calculator for this. 5,250 – 4,500 = 750, and the answer is C, it says 750, so I choose C.

Student 24MNY analyzes the data in a two-way table in several ways to understand the contextual meanings of the ratios in several answer options for item NC14, using this information to find the correct answer:

> So, I'm just gonna highlight "slate roof" and the "single story" on the graph. And so—the combination of both is 4. And, if it's looking at random, from everything, it has to be 4/48, right? And you could go divide that I think, simplify it a little more, but it's not showing it simplified. And if I picked 4 out of 15, I'd be showing how many single-story shale houses, but it's saying "if one of the houses," which is all, so it can't be B. C says 4 out of 14. They're showing out of slate houses, which can't be it. The 14 out of 4 is showing all the slate houses, which can't be it, so I pick A, 4/48.

Discussion

As suggested in the results section, above, two statistics are particularly valuable in helping answer the questions of whether the Evidence-Based Reading and Writing (ERW) and Math items studied performed as intended in eliciting complex cognitive processes from test takers. First is the number/percentage of test takers who demonstrated all required behaviors for a given ERW item or one or more expected behaviors for a given Math item. Given that the required behaviors associated with each item type or category (with one DOK 1–level exception for Reading) are collectively complex, students demonstrating the required/expected behaviors are thereby both enacting the item type/category as designed and demonstrating complex cognition. Second is the difference between the

number of students who answered each item correctly and the number who answered correctly and also demonstrated all required ERW behaviors or one or more expected Math behaviors. A low number here suggests that the only way to answer the item correctly is by demonstrating the required/expected behaviors, whereas a high number suggests that students, intentionally or not, can circumvent the item's design—in essence, are able to find a shortcut to answering that calls into question the claim that the item is functioning as designed and eliciting sophisticated thinking. Below, we discuss the ERW and Math items, in turn, on these two criteria. We also remind readers that the vignettes shared for both ERW and Math provide additional evidence that students are demonstrating complex cognitive behaviors in accordance with the items' designs.

Evidence-Based Reading and Writing (ERW)

Preceding sections presented the results for five ERW item types or categories: Citing Textual Evidence (Reading), Interpreting Words and Phrases in Context (Reading), Analyzing Quantitative Information (Reading), Development (Writing and Language), and Effective Language Use (Writing and Language). Per the two criteria listed above—number/percentage of students demonstrating all required behaviors for an item and difference in number between those who merely answered correctly and those who also demonstrated all required behaviors—the ERW items were found, in the main, to be successful.

Citing Textual Evidence (Reading). A majority of students demonstrated all required behaviors for only two of the six Citing Textual Evidence items or pairs of items. A near-majority did so for two other item pairs; only 20 percent and 30 percent of students answered both items correctly in the other two pairs, making it impossible for students to demonstrate all required behaviors (since a given item or item pair must be answered correctly for students to demonstrate all ERW required behaviors). The modal difference between the number of students answering the item/pair correctly and also demonstrating all behaviors was zero (three pairs), with the other differences being two (two pairs) and three (one item). Taken together, these results indicate high performance on the two criteria of interest once item difficulty is taken into account.

Interpreting Words and Phrases in Context (Reading). A majority of students was able to demonstrate both required behaviors for five of six studied items. For four of the six items, the difference between the number of students answering correctly and the number who also demonstrated both required behaviors was two (one item), four (two items), or five (one item). The two items demonstrating larger gaps—eleven and fifteen—likely did so in significant measure because the items' extreme ease for this group (answered correctly by all but one student and by all students, respectively) reduced the need to demonstrate careful reasoning in the think-aloud. In most respects, then, the items in this category performed well on the two criteria of interest here.

Analyzing Quantitative Information (Reading). A majority of students was able to demonstrate all required behaviors for four of five studied items. For the five items, the difference between the number of students answering correctly and the number of students who also demonstrated all required behaviors was two, five, seven, twelve, and twenty-three. As noted in the results section, above,

the item demonstrating the second-largest gap was an easy DOK 1 item that, by design, didn't elicit complex behaviors and whose low difficulty seemed to discourage student reflection. Also as noted, the item demonstrating the largest gap (twenty-three) was flawed in that it didn't require the level of synthesis its DOK 3 designation would suggest. Overall, the results here indicate high performance on the two criteria of interest.

Development (Writing and Language). A majority of students was able to demonstrate all required behaviors in seven out of nine studied items; in the other two cases, exactly half of students or just under half of students were able to do so. For eight of the nine items, the difference between the number of students who answered correctly and the number who also demonstrated all required behaviors ranged from one (four items) to five (one item), with differences of two (one item) and three (two items) also indicated. One item evinced a gap of ten, which, as previously noted, suggests a weakness in that item in that it didn't, as was intended, require an equal or nearly equal use of both passage and accompanying table to answer correctly. In other respects, however, the items in this category performed well on the two criteria of interest here.

Effective Language Use (Writing and Language). A majority of students was unable to demonstrate all required behaviors for any of the four items in this category. One Style and Tone item was difficult for the sampled students (answered correctly by only eight out of twenty-nine students), which, as discussed earlier, prevented a majority of students from demonstrating all required behaviors. The two Precision items failed on this criterion largely because students, in the main, didn't demonstrate an understanding of what the item was asking of them, even though both items were relatively easy (answered correctly by roughly three-quarters of students in each case). The remaining Style and Tone item failed on this criterion primarily because many students failed to offer a clear rationale for their answer choice in terms of style or tone. Not surprisingly, given these concerns, we find that the gaps between the number of students who answered correctly and the number who also demonstrated all required behaviors were high in three of four cases, ranging from ten to twenty; the exception was the Style and Tone item that few students answered correctly. The general failure on the two criteria of interest suggest some lack of transparency about these items' purpose or demands, and the issue merits further study.

Although there were certain areas of concern on the two criteria of interest discussed above, vignettes from the students who both answered a given item correctly and demonstrated all required behaviors provide strong evidence of complex cognition as well as enactments of the items' intended designs. While not all students were able to offer this sort of evidence, those who did indicate the value of these items in eliciting sophisticated thinking.

The student vignettes associated with each of the ERW item types reinforce the general impression from the statistics that the items are capable of eliciting complex cognition. These vignettes, by design, illustrate successful performance and don't obviate the fact that certain item types deserve further study with respect to their constructs and presentation, but they do lend further credence to the general conclusion that the studied ERW items are, in large measure, working as intended and calling forth sophisticated thought processes.

Math

Preceding sections presented the results for five Math item categories: Algebra; Functions; Geometry; Ratios, Proportions, and Percentages; and Statistics and Probability. Per the two criteria listed above—number/percentage of students demonstrating one or more expected behaviors for an item and difference in number between those who merely answered correctly and those who also demonstrated one or more expected behaviors—the Math items were found, in the main, to be successful.

Algebra. A majority of students demonstrated one or more of the expected behaviors for all nine studied items. Furthermore, the difference between the number of students who answered a given item correctly and the number who demonstrated one or more expected behaviors was two or fewer for eight of the nine items, with the remaining item having a gap of six. These data suggest these items' high performance in relation to the two criteria of interest.

Functions. For all seven items, a majority of students demonstrated one or more expected behaviors. For the seven items, the gap between the number of students answering correctly and the number of students who also demonstrated one or more expected behaviors was three or fewer. The items in this category thus performed well on the two criteria under discussion.

Geometry. For all six studied items, a majority of students demonstrated one or more expected behaviors. Furthermore, in no case was the difference between the number of students who answered correctly and the number who also demonstrated one or more expected behaviors greater than two. As a result, items in this category performed well on the two criteria of interest.

Ratios, Proportions, and Percentages. For six of the eight items, a majority of students demonstrated one or more expected behaviors. In addition, the differential between the number of students who answered correctly and the number who also demonstrated one or more expected behaviors was never greater than three. In terms of the two criteria of interest, items in this category performed well.

Statistics and Probability. A majority of students demonstrated one or more expected behaviors for all seven items studied, and for none of the items was there a difference greater than two between the number of students who answered the item correctly and those who also demonstrated one or more expected behaviors. In these respects, then, the items performed well relative to the two criteria under examination.

Admittedly, some items evinced potential shortcomings with respect to requiring complex cognitive behaviors (in addition to permitting inefficient and effortful but nonetheless successful paths to correct solutions). However, vignettes from the student sample in all five studied categories indicate that the items were capable of eliciting sophisticated thought processes from students and, along with the data related to the two criteria discussed previously, offer evidence that the items, in the main, are appropriately challenging and working as intended.

Conclusion

This large, fine-grained qualitative study of student understanding of and performance on a broad cross section of SAT Evidence-Based Reading and Writing (ERW) and Math items using the cognitive interview methodology provides substantial evidence in support of the claim that the SAT ERW and Math sections include items that elicit from students instances of complex cognition in accordance with the items' designs. This evidence comes in two forms:

- First, tabulations of verbalizations of student performance on items in relation to sets of required (ERW) or expected (Math) behaviors generally support the conclusion that the items called on students to demonstrate a range of cognitively complex activities and that these activities were required for answering the items.
- Second, vignettes from students who both answered a given item correctly and demonstrated all required behaviors (ERW) or one or more expected behaviors (Math) vividly illustrate the sophistication of thinking that the associated items across numerous areas are capable of eliciting.

Although the evidence recounted in this report strongly endorses the claim that SAT ERW and Math items are cognitively challenging, the analysis and discussion herein have also candidly acknowledged that the studied SAT items display some weaknesses relative to their intended constructs. These item types/categories call for further scrutiny from College Board staff and from independent subject matter experts. Nonetheless, the bulk of the evidence in the report argues loudly for the cognitive breadth and depth of the SAT and, ultimately, its value as an assessment of college and career readiness.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association, American Psychological Association, and National Council on Measurement in Education.

Bain, Robert B. 2012. "Using Disciplinary Literacy to Develop Coherence in History Teacher Education: The Clinical Rounds Project." *The History Teacher* 45, no. 4 (August): 513–32. **https://www.jstor.org/stable/i23265888**.

Battista, Michael T. 2007. "The Development of Geometric and Spatial Thinking." In *Second Handbook of Research on Mathematics Teaching and Learning: A Project of the National Council of Teachers of Mathematics*, edited by Frank K. Lester Jr., 843–908. Charlotte, NC: Information Age.

Beck, Isabel L., Margaret G. McKeown, and Linda Kucan. 2013. *Bringing Words to Life: Robust Vocabulary Instruction*, 2nd ed. New York: Guilford.

Board of Education, Commonwealth of Virginia. 2016. *Mathematics Standards of Learning for Virginia Public Schools.* Richmond: Board of Education, Commonwealth of Virginia.

California Department of Education. 2013. *California Common Core State Standards: Mathematics.* Sacramento: California Department of Education.

Clements, Douglas H., and Michael T. Battista. 1992. "Geometry and Spatial Reasoning." In *Handbook of Research on Mathematics Teaching and Learning: A Project of the National Council of Teachers of Mathematics*, edited by Douglas A. Grouws, 420–64. New York: Macmillan.

College Board. 2019. *College Board National Curriculum Survey Report 2019.* New York: College Board. **https://collegereadiness.collegeboard.org/pdf/national-curriculum-survey-report.pdf**.

Ericsson, K. Anders, and Herbert A. Simon. 1993. *Protocol Analysis: Verbal Reports as Data*, rev. ed. Cambridge: Massachusetts Institute of Technology.

Fisher, Douglas, and Nancy Frey. 2015. *Text-Dependent Questions, Grades 6–12: Pathways to Close and Critical Reading.* Thousand Oaks, CA: Corwin.

Franklin, Christine, Gary Kader, Denise Mewborn, Jerry Moreno, Roxy Peck, Mike Perry, and Richard Scheaffer. 2007. *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K–12 Curriculum Framework.* Alexandria, VA: American Statistical Association.

Gaertner, Matthew N., Jeongeun Kim, Stephen L. DesJardins, and Katie Larsen McClarty. 2014. "Preparing Students for College and Careers: The Causal Role of Algebra II." *Research in Higher Education* 55, no. 2 (March): 143–65.

Gormley, Kathleen, and Peter McDermott. 2015. "Searching for Evidence—Teaching Students to Become Effective Readers by Visualizing Information in Texts." *The Clearing House* 88, no. 6 (October): 171–7. **https://doi.org/10.1080/00098655.2015.1074878**.

Hart Research Associates. 2018. *Fulfilling the American Dream: Liberal Education and the Future of Work: Selected Findings from Online Surveys of Business Executives and Hiring Managers.* Washington, DC: Hart Research Associates.

Hart Research Associates and Public Opinion Strategies. 2005. *Rising to the Challenge: Are High School Graduates Prepared for College and Work?; A Study of Recent High School Graduates, College Instructors, and Employers.* Washington, DC: Hart Research Associates and Public Opinion Strategies.

Kim, YoungKoung, Andrew Wiley, and Sheryl Packman. 2012. *National Curriculum Survey on English and Mathematics.* New York: College Board.

Klepfer, Kasey, and Jim Hull. 2012. *High School Rigor and Good Advice: Setting Up Students to Succeed.* Alexandria, VA: National School Boards Association, Center for Public Education.

Leighton, Jacqueline P. 2017. *Using Think-Aloud Interviews and Cognitive Labs in Educational Research.* New York: Oxford University Press.

Liben, David. 2020. "The Importance of Vocabulary and Knowledge in Comprehension." In *The SAT Suite and Classroom Practice: English Language Arts/Literacy*, edited by Jim Patterson, 53–69. New York: College Board. **https://collegereadiness.collegeboard.org/pdf/sat-suite-classroom-practice-english-language-arts-literacy-c3.pdf**.

Liben, Meredith. 2020. "Close Reading, Textual Evidence, and Source Analysis." In *The SAT Suite and Classroom Practice: English Language Arts/Literacy*, edited by Jim Patterson, 31–51. New York: College Board. **https://collegereadiness.collegeboard.org/pdf/sat-suite-classroom-practice-english-language-arts-literacy-c2.pdf**.

Long, Mark C., Patrice Iatarola, and Dylan Conger. 2009. "Explaining Gaps in Readiness for College-Level Math: The Role of High School Courses." *Education Finance and Policy* 4, no. 1 (Winter): 1–33.

Minnesota Department of Education. 2007. *Minnesota Academic Standards: Mathematics K–12.* Roseville: Minnesota Department of Education.

Moje, Elizabeth Birr, Darin Stockdill, and Rebekah Hornak. 2019. "Essential Instructional Practices for Disciplinary Literacy, Grades 6 to 12." *Michigan Reading Journal* 52, no. 1 (Fall): 62–66.

National Center on Education and the Economy. 2013. *What Does It Really Mean to Be College and Work Ready?: The Mathematics and English Literacy Required of First Year Community College Students.* Washington, DC: National Center on Education and the Economy.

Oehrtman, Michael, Marilyn Carlson, and Patrick W. Thompson. 2008. "Foundational Reasoning Abilities That Promote Coherence in Students' Function Understanding." In *Making the Connection: Research and Teaching in Undergraduate Mathematics Education*, edited by Marilyn P. Carlson and Chris Rasmussen, 27–42. Washington, DC: Mathematical Association of America.

Peterson, Christina Hamme, N. Andrew Peterson, and Kristen Gilmore Powell. 2017. "Cognitive Interviewing for Item Development: Validity Evidence Based on Content and Response Processes." *Measurement and Evaluation in Counseling and Development* 50, no. 4 (October): 217–23. **https://doi.org/10.1080/07481756.2017.1339564**.

Shanahan, Cynthia, and Timothy Shanahan. 2020. "Disciplinary Literacy." In *The SAT Suite and Classroom Practice: English Language Arts/Literacy*, edited by Jim Patterson, 91–125. New York: College Board. **https://collegereadiness.collegeboard.org/pdf/sat-suite-classroom-practice-english-language-arts-literacy-c5.pdf**.

Shaughnessy, J. Michael. 1992. "Research in Probability and Statistics: Reflections and Directions." In *Handbook of Research on Mathematics Teaching and Learning: A Project of the National Council of Teachers of Mathematics*, edited by Douglas A. Grouws, 465–94. New York: Macmillan.

Steen, Lynn Arthur, ed. 2001. *Mathematics and Democracy: The Case for Quantitative Literacy*. Princeton, NJ: National Council on Education and the Disciplines.

Tourangeau, Roger, and Kenneth A. Rasinski. 1988. "Cognitive Processes Underlying Context Effects in Attitude Measurement." *Psychological Bulletin* 103 (3): 299–314.

U.S. Department of Education. 2018. *A Leak in the STEM Pipeline: Taking Algebra Early.* Washington, DC: U.S. Department of Education. **https://www2.ed.gov/datastory/stem/algebra/index.html**.

Willis, Gordon B. 2005. *Cognitive Interviewing: A Tool for Improving Questionnaire Design.* Thousand Oaks, CA: SAGE.