CollegeBoard

The Cognitively Complex Thinking Required by Select SAT® Suite Questions

Evidence from Students with Specific Learning Disorders Affecting Reading (Dyslexia)

College Board
June 2025

The Cognitively Complex Thinking Required by Select SAT Suite Questions:

Evidence from Students with Specific Learning Disorders Affecting Reading (Dyslexia)

June 2025

Jim Patterson | **Lead author; Reading and Writing section analysis**

Michael Gosche | **Math section analysis**

Jay Happel | **Sample analysis**

Beth Oxler, Georgina Keenan, Nancy Burkholder | **Editorial services**

Vidlet, Inc. | **Cognitive interviews**

Suggested Citation:
College Board. 2025. *The Cognitively Complex Thinking Required by Select SAT Suite Questions: Evidence from Students with Specific Learning Disorders Affecting Reading (Dyslexia).* College Board.

About College Board

College Board reaches more than 7 million students a year, helping them navigate the path from high school to college and career. Our not-for-profit membership organization was founded more than 120 years ago. We pioneered programs like the SAT® and AP® to expand opportunities for students and help them develop the skills they need. Our BigFuture® program helps students plan for college, pay for college, and explore careers. Learn more at **cb.org**.

Contents

Tables

Figures

Executive Summary

This report documents the findings of a think-aloud (cognitive lab) study conducted with students with specific learning disorders affecting reading (here, abbreviated SLDR; also known as dyslexia) as they answered a set of either SAT® Suite Reading and Writing or Math questions. The research goals were, first, to ascertain, via qualitative and quantitative means, whether these students with SLDR were able to demonstrate cognitively complex thinking in line with the question types' constructs and college and career readiness requirements and, second, to explore whether participants' performance on the questions or their postexperience reflections on the think-aloud activity would uncover any construct-irrelevant barriers to their success on such questions.

Fifteen high school juniors and seniors who indicated having SLDR and met other criteria were selected to participate in the Reading and Writing segment of the study, while an additional twenty-one such students participated in the Math segment. Each participant was asked to think aloud (narrate their thoughts) to a moderator supplied by vendor Vidlet, Inc., as they answered up to fifteen Reading and Writing or Math questions (selected to be broadly representative of the sections' domains) and to answer a standardized series of postexperience interview questions. Participants engaged with the test questions via Bluebook™, the custom-built testing application developed by College Board to administer the SAT Suite tests in their digital-adaptive formats, had access to the app's universal tools, and could, if desired, use a third-party screen reader. Within the constraints of selection criteria, small sample sizes, and the self-selection methodology, the resulting Reading and Writing and Math participant pools were generally diverse in terms of gender, race/ethnicity, grade in school, self-reported high school GPA (HSGPA), and self-reported SLDR impact on their test-taking ability.

The focal portions of the sessions, which were scheduled for roughly two hours and for which participants were compensated via gift card, were video recorded. The transcripts produced from these sessions were analyzed qualitatively and quantitatively by College Board subject matter experts relative to lists of predefined required (Reading and Writing) or expected (Math) behaviors, which operationally defined the questions' constructs by question type. The researchers performed coding in MAXQDA, a qualitative/mixed-methods research software

package, and tabulated results in Microsoft Excel. Each participant-by-question interaction was assigned one of up to five performance levels (PLs), with PL 1 representing the most successful performance (answering a given question correctly while also demonstrating all required behaviors [Reading and Writing] or at least one expected behavior [Math]) and PL 5 representing the least successful (answering a given question incorrectly and demonstrating no required or expected behaviors).

The College Board researchers analyzed the coded transcripts on three dimensions:

1. **Participant performance** was analyzed in terms of the number and proportion of correctly answered questions for which participants demonstrated appropriate cognitive behaviors. Vignettes (transcript excerpts) from select participants were used when available to illustrate demonstrations of the cognitively complex thinking elicited by the test questions.
2. **Question performance** was analyzed in terms of the number and proportion of correctly answering participants who also demonstrated appropriate cognitive behaviors.
3. **Participant perceptions** of the question-answering activity, in the form of responses to postexperience interview questions, were analyzed for both general themes and for any cases in which participants identified potential construct-irrelevant barriers to their success in the activity and to SAT Suite test taking more broadly.

The main metric used to assess participant performance was the *participant differential* (D_p). Mathematically, D_p represents the arithmetic difference between (1) the number of Reading and Writing or Math questions a given participant answered correctly and (2) the number of such questions for which the participant demonstrated all required behaviors (Reading and Writing) or at least one expected behavior (Math). Conceptually, D_p represents the "difference" between simply answering a given question correctly and doing so while also exhibiting appropriate behaviors. Because participants answered a variable number of test questions during the activity, the threshold for a "good" D_p was set at 70 percent, meaning that a given participant needed to demonstrate appropriate behaviors for at least 70 percent of the questions they answered correctly. Vignettes (transcript excerpts) from participants attaining PL 1 on each test question are provided when available and serve as a second source of evidence respecting participant performance on the questions.

The main metric used to assess question performance was the *question differential* (D_q). Similar to D_p, D_q represents, in mathematical terms, the arithmetic difference between (1) the number of participants answering a given question correctly and (2) the number of such participants who also demonstrated appropriate behaviors. Conceptually, D_q represents the "difference" between the number of participants who simply answered a given question correctly and the number who did so while also demonstrating appropriate behaviors. The threshold for a "good" D_q was again set at 70 percent, meaning, in this case, that for a given question, at least 70 percent of correctly answering participants also demonstrated appropriate behaviors.

Participant perceptions of the think-aloud activity were collected via a standardized set of postexperience interview questions. Responses to these questions were analyzed both for general themes and for indicators that participants had been affected by construct-irrelevant barriers and were thus impeded from demonstrating the full extent of their subject matter knowledge.

This report delineates three key findings:

- **Participant performance.** Nine of fifteen Reading and Writing participants (60 percent) and seventeen of twenty-one Math participants (81 percent) met or exceeded the threshold for a good D_p, providing evidence that students with SLDR are able to demonstrate cognitively complex thinking in line with the question types' constructs. Additionally, vignettes exhibiting PL 1 were obtained for fourteen of the fifteen Reading and Writing questions and for thirteen of the fifteen Math questions, providing additional support for the claim that students with SLDR can demonstrate cognitively complex thinking via SAT Suite test questions.
- **Question performance.** Ten of the fifteen Reading and Writing questions (67 percent) and twelve of the fifteen Math questions (80 percent) met or exceeded the threshold for a good D_q, providing evidence that, overall, the presented questions were capable of eliciting cognitively complex thinking from students with SLDR.
- **Participant perceptions.** No clear evidence of construct-irrelevant barriers not already addressed by the provision of testing accommodations emerged from participant responses to the postexperience interview questions or observation of participant question-answering behavior during the think-aloud activity.

The generalizability of the results of this study is limited by several factors, including the study's small sample sizes, the artificiality of the think-aloud methodology itself, and the possibility (though, as it turned out, likely not the reality) that some participants may have previously encountered the studied SAT Suite test questions as part of their normal test preparation activities.

The study's positive outcomes respecting students with SLDR must also be contextualized with the understanding that the results assume these students have access to appropriate accommodations during testing, including extra time/extra breaks and possibly assistive technology, such as text-to-speech. The fact that only one participant used a screen reader as part of the study may also suggest that this and similar tools may have relatively low penetration among students with SLDR, a matter that should be investigated further given that such technology might help students with documented text processing issues.

Section 1: Introduction

The following report presents the methodology, findings, and implications of a verbal protocol study conducted in 2024 by College Board, with support from vendor Vidlet Inc., involving samples of high school juniors and seniors who identify as having a specific learning disorder affecting reading (hereafter abbreviated as SLDR, and also known as dyslexia) as they thought aloud through a series of either SAT Suite Reading and Writing or Math questions.

The research goals of this study were twofold:

- Does evidence gathered from qualitative and quantitative analysis of transcripts from samples of high school juniors and seniors with SLDR support the conclusion that select SAT Suite test questions are capable of eliciting cognitively complex thinking from students with SLDR in line with college and career readiness expectations and the question types' constructs?
- Is evidence gathered from these transcripts and/or responses to postexperience interview questions suggestive of potential non-content-related (i.e., *construct-irrelevant*) impediments to the ability of students with SLDR to demonstrate the full extent of what they know and can do in the literacy and math domains of the SAT Suite tests? If so, have these impediments been addressed by the provision of testing accommodations, such as extra time?

In brief, this study, one of several verbal protocol studies of the SAT Suite conducted by College Board (College Board and HumRRO 2020; College Board 2024a, 2025a, 2025b), engaged samples of high school juniors and seniors in thinking aloud—verbalizing their thought processes—as they answered a series of either Reading and Writing or Math test questions selected from the official practice environment. Transcripts of these moderator-led sessions were produced and then analyzed for evidence of participants having exhibited cognitively complex behaviors associated with the various Reading and Writing and Math question types administered. Each participant-by-question interaction was evaluated for these behaviors as well as for whether the question was answered correctly or incorrectly, and then performance levels were assigned. Metrics called *differentials* were determined for each participant and for each Reading and Writing and Math test question, with the criteria for successful results being, respectively, that each participant demonstrated appropriate cognitive behaviors at least 70 percent of the time when answering questions correctly and that at least 70 percent of the time, participants answered a given question correctly while also demonstrating appropriate cognitive behaviors. Transcript vignettes (excerpts) exemplifying participants correctly answering a given question and exhibiting appropriate behaviors were identified whenever possible and served as a second source of evidence for this study. Responses to a standardized set of postexperience interview questions were also analyzed and served as an additional evidence source.

Document Preview

Section 2: Literature Review offers a brief overview of the research literature consensus on the validity of using a concurrent verbal protocol/think-aloud methodology as a means of gaining insight into cognitive processes that would otherwise be inaccessible or prone to retrospective or inferential bias. Section 3: Methodology details the method used to conduct the study and analyzes the enacted student samples along demographic lines. Section 4: Results presents the qualitative and quantitative findings obtained from the study, including summative metrics, question-by-question transcript vignettes, and analysis of postexperience interview question responses. Section 5: Discussion interprets the findings presented in the preceding section, draws conclusions and implications, and considers the study's limitations. Section 6: Conclusion briefly wraps up the body of the report. Following the references is an appendix containing the recruitment materials and excerpts of the verbal protocols used by College Board and Vidlet in carrying out the study's data collection.

Section 2: Literature Review

Verbal Protocols as Data in Social Science Research

The formal use of verbal protocols as a research tool to uncover otherwise unobservable cognitive processes extends back at least a century (Ericsson and Simon 1993). The scholarly consensus over the last half century has supported the use of verbal protocols as a data collection tool within a range of limitations and constraints, discussed more thoroughly below (Russo et al. 1989; Bainbridge and Sanderson 1995; Goos and Galbraith 1996; Branch 2013). Verbal protocol studies have illuminated participant thought processes in a wide range of areas, including business management (Isenberg 1986), marketing and consumer choice (Bolton 1993; Bettman and Park 1980), computer programming (Vessey 1986), engineering (Atman and Turns 2001), accounting (Biggs and Mock 1983), nursing (Haffer 1990), information systems (Nguyen and Shanks 2007), library science (Branch 2001), human geography (Lundberg 1984), and education (Suto and Greatorex 2008).

Education has, in fact, been one of the more fertile areas for verbal protocol studies in recent years. The appeal of the methodology to this field is intuitively obvious. Researchers, teachers, curriculum specialists, and other stakeholders are committed to developing and implementing instructional methods and materials that promote student learning, but such learning takes place, often silently and unobserved, in students' heads. Without some sense of how students themselves are engaging (or not engaging) with these methods and materials, we can't fully or fairly account for the success or failure of these interventions.

One foundational verbal protocol study in the education field was that of Pressley and Afflerbach (1995), who used and refined the approach in an effort to create a model of conscious mental processes enacted during reading. A particular area of focus for many literacy-related verbal protocol studies has been distinguishing the behaviors of more and less successful readers. For example, Kletzien (1991) employed verbal protocols to attempt to differentiate strategy use by high school–age students of higher and lower reading achievement levels as they engaged with successively more challenging expository passages. Kletzien found that both groups of participants used similar strategies but that those with better comprehension skills used more, and more varied, strategies as the texts became harder. Magliano and Millis (2003) used verbal protocol analysis to help develop a latent semantic analysis–based computerized reading comprehension measure. Drawing on prior work and their 2003 study, the researchers found that "good readers emphasize establishing coherence[,] and poor readers emphasize the contents of the current sentence" as they read (255). More recently, Cho et al. (2018) qualitatively and quantitatively analyzed the verbal responses of ten more and ten less successful online readers in an effort to determine how these two groups differed in their cognitive approaches to analyzing a controversial topic. The authors concluded that the more successful readers engaged in the work in ways "notably different" (215) from those of their less successful peers in terms of extent of source evaluation and application of metacognitive strategies related to successfully accomplishing the task.

Verbal protocol analysis has also been used successfully to explore participants' thought processes as they engage in math tasks. For instance, Goos and Galbraith (1996) used the methodology to determine that two high school seniors collaborating on a series of problems in an applied math course exhibited "differing, but complementary, metacognitive strengths" (255), which typically aided in their joint problem-solving. Montague and Applegate (1993) analyzed the verbal protocols from eighty-one middle school students, roughly a third of whom were selected randomly from pools of learning disabled, average-achieving, and gifted students in a large southeastern metropolitan district. The researchers found that when presented with a range of problems in math, students identified as gifted were more strategic in their solving approaches than students in the other two achievement groups; that perceived difficulty of math problems seemed to affect students' perseverance and cognition; and that "students with LD [learning disabilities] approach[ed] problem solving in a qualitatively different manner than their more proficient peers" (29). Özcan et al. (2017) also used verbal protocol analysis to examine math problem-solving approaches used by students, in this case sixty-nine sixth graders sampled across achievement levels. Among their findings, the researchers determined that those students who employed an incorrect process in solving a nonroutine math problem "mostly [did] operations aimlessly" and approached the word problem superficially (139–40).

As indicated above, the verbal protocol method has been employed successfully with students with learning disabilities. Özkubat and Özmen (2021) used think-aloud protocols as one tool to evaluate the math problem-solving skills of both sixth-grade students with learning disabilities and low- and average-ability students without such disabilities. Deshpande et al.'s (2021) small-scale examination of high school students' problem-solving abilities in geometry used think-alouds to illuminate cognitive and metacognitive strategies employed by students with and without learning disabilities. Similarly, Botsas (2017) used think-aloud protocols to explore the cognitive and metacognitive strategy use of fifth- and sixth-grade students with and without learning disabilities as they read both narrative and expository science texts.

Verbal protocol studies have also frequently been used to study the cognitive (and metacognitive) processes of language learners as they acquire a second or subsequent language or perform other academic tasks. Yayli (2010) employed both think-aloud and retrospective methods to investigate the reading-related cognitive and metacognitive strategies of proficient and less proficient readers enrolled in a university-level English language teaching department in Turkey. Bowles and Gastañaga (2022) used a think-aloud method as one approach to assessing the impact of various forms of written corrective feedback given to heritage language, second-language, and third-language university-level learners of Spanish on their short essays. Al-Maani et al. (2024) used think-alouds to examine the language learning strategies used by intermediate and advanced Jordanian English as a foreign language (EFL) college seniors as they performed reading, writing, and listening tasks.

Though obviously not exhaustive, the above overview of verbal protocol studies in literacy and math education establishes that the methodology has been used to examine a broad range of cognitive and metacognitive activities in an array of fields. Moreover, in educational research, this approach has been used successfully in both literacy and math (as well as in other subject areas) with numerous categories of students, including younger and older students, higher- and lower-achieving students, native language speakers and language learners, and students who are neurodivergent as well as students who aren't.

Verbal Protocols as Data in Research on the Designs of Large-Scale Standardized Assessments

Of particular relevance to the present study is the use of the think-aloud methodology to analyze and evaluate elements of the design of large-scale standardized assessments. One such study is that of Johnstone et al. (2006), who concluded that the cognitive lab methodology elicited useful information about construct-irrelevant barriers in math test design from several student population subgroups of educational concern, including students with learning disabilities, students with hearing impairments, and English learners, as well as from English-proficient students without disabilities. By contrast, the researchers found students with cognitive impairments lacked the requisite verbalization capacities during problem-solving. Of further note, the authors found the methodology yielded little data on the hardest math test items studied because of the difficulties participants had in simultaneously solving these problems and verbalizing their approaches. A similar study, this time by Johnstone et al. (2007), explored a variety of ways of making grade 8 reading items more comprehensible. Using a think-aloud methodology with recently promoted eighth-grade students, the team determined that the use of "non-construct vocabulary"—that is, undefined specialized subject area terms—could pose (correctable) barriers to student performance, while such interventions as reducing passage word counts and boldfacing key words didn't seem to influence achievement.

Threats to Verbal Protocol Validity and Reliability

Although the preceding account clearly establishes that verbal protocol analysis has been extensively used in social science research, including in education, serious concerns about the validity of the method have been raised over the years that require and have received fair-minded consideration and response.

One of the earliest and most influential critiques of verbal protocols as data came from Nisbett and Wilson (1977). Drawing from then-burgeoning critiques of introspection-based research methods, the authors posited three major conclusions:

1. "The accuracy of subjective reports [of higher-order thinking involving inferences] is so poor as to suggest that any introspective access that may exist is not sufficient to produce generally correct or reliable reports.
2. "When reporting on the effects of stimuli, people may not interrogate a memory of the cognitive processes that operated on the stimuli; instead, they may base their reports on implicit, a priori theories about the causal connection between stimulus and response. . . .
3. "Subjective reports about higher mental processes are sometimes correct, but even the instances of correct report are not due to direct introspective awareness. Instead, they are due to the incidentally correct employment of a priori causal theories" (233).

Rather than outright rejecting these concerns, Ericsson and Simon (1993) countered with a simple mental processing model that differentiates between information stored in a person's short-term memory (STM) and long-term memory (LTM). Specifically, the authors contended that "information recently acquired (attended to or heeded) by the central processor is kept in STM, and is directly accessible for further processing (e.g., for producing verbal reports), whereas information from LTM must first be retrieved (transferred to STM) before it can be reported" (11). In other words, participants in verbal protocol studies should be able to give accurate accounts of their cognition during or shortly after experiencing a stimulus, such as a novel task to be solved; by contrast, verbal accounts that depend on recall and interpretation of past stimuli (i.e., that require, in Ericsson and Simon's model, retrieval from LTM) are more prone to the kinds of validity errors that Nisbett and Wilson (1977) identified.

Subsequent researchers have further codified potential threats to the accuracy of verbal protocols as data sources. Bainbridge and Sanderson (1995), for example, identified several ways in which verbal reports can be distorted, with the aim of encouraging researchers to find ways to minimize or eliminate these risk factors. Potential distortion sources identified by Bainbridge and Sanderson include the following:

1. Altering the nature and performance of a task merely by asking for a verbalization
2. Placing participants under significant time pressure, which can lead to glossing over steps in cognition
3. Social and self-presentation biases leading participants to give what they think are expected or socially acceptable answers

4. Asking participants to verbally discuss processes (e.g., perceptual-motor skills) that are typically performed nonverbally and outside of conscious thought
5. Participants being unable to articulate everything they know about and can do with a given stimulus (e.g., a problem-solving task), meaning that "verbal protocol evidence may provide only a limited sample of the total knowledge available to the person being studied" (173)

Stratman and Hamp-Lyons (1994) conceptualized threats to the accuracy of verbal protocols as problems of *reactivity*, or the verbal protocol methodology itself altering the cognitive processes intended to be studied. Challenges identified by the authors include flawed verbalization directions given to participants; the difficulty participants often experience in simultaneously thinking and verbalizing; the impact on participants of hearing their own voices during verbalization; the impact of participants learning about themselves during the verbalization process (rather than simply reporting); and the possibility of experimenters inadvertently cueing expected or desired responses through their words or actions. Similarly, Kirk and Ashcraft (2001, 158–59) identified three sources of threat to verbal protocol accuracy: veridicality ("whether the verbal reports accurately reflected the underlying cognitive processes"), reactivity ("the possibility that the verbal report requirement may have altered the mental processing that normally occurs"), and demand-induced bias ("the possibility that aspects of the experimental procedures suggested to participants what kinds of verbal reports and solutions were expected").

The consensus among researchers has been to treat issues of (in Kirk and Ashcraft's formulation) veridicality, reactivity, and demand-induced bias seriously without abandoning the methodology. For instance, Leow and Morgan-Short (2004), echoing Ericsson and Simon and others, suggest that verbal protocol approaches be limited to eliciting "introspective, nonmetalinguistic verbalizations" (36)—that is, verbalizations made concurrent with task performance, rather than retrospectively after the task, and focused on description of behaviors rather than attempts at explanations about why certain behaviors were performed. The researchers' study specifically examined whether the act of thinking aloud altered performance on a reading task given to college-age students and found no such evidence when students in the think-aloud and control (non-think-aloud) conditions were compared statistically. By contrast, Kirk and Ashcraft (2001), in their study of adult use of strategies in the solving of simple arithmetic problems and who also employed a "silent" control group, found questionable veridicality and signs of reactivity. (We speculate, along the lines of Bainbridge and Sanderson's [1995] cautions quoted above, that this outcome may have resulted in part because the task—simple arithmetic with college-age participants—was too routine, and therefore too far out of conscious understanding, for meaningful verbal protocol analysis.) They advocate for a careful analysis of instructions given to participants to minimize potential bias in response and for the use of a nonverbalizing control group to serve as a baseline. Russo et al. (1989) similarly call for the use of "silent" control conditions, as they found it impossible to determine a priori using then-existing theory which tasks were likely to provoke reactivity in participants.

Concurrent and Retrospective Verbalizations

The preceding discussion and the general research consensus (e.g., Russo et al. 1989) suggest that concurrent verbal protocols are more trustworthy than are retrospective ones. This stands to reason, as it should be easier for participants to accurately verbalize in-the-moment cognition during task performance than re-create their thought processes sometime after the fact. In accordance, the present study relies on concurrent verbal protocols and emphasizes description of behaviors performed by participants rather than the motivations behind their behaviors.

Some researchers, however, have made a case for a hybridized approach, one that makes use of both concurrent and retrospective dimensions. Johnstone et al. (2006) advocated for such a blended approach, contending that it counterbalanced both the propensity of think-aloud verbalizations to be "incoherent" (2) and that of interviews to elicit potentially inaccurate retrospective explanations of behaviors already encoded into long-term memory.

While noting several concerns about the use of data requiring participants to retrieve information from long-term memory, Taylor and Dionne (2000) advocate for the value of retrospective debriefing (RD) in tandem with concurrent verbal protocols (CVP), which they found obtained "a richer account of problem-solving strategy than did either method used alone." Specifically:

> When problem solvers are requested to think aloud while solving a problem (CVP), and then to describe how they solved the problem (RD), CVP data can be used to provide data-based cues to guide the collection of RD data on a specific problem-solving event. . . . In turn, convergent information about the same event contained in the broader spectrum of RD data can be used by researchers to elaborate CVP data, which tend to focus on the control of the problem-solving process. . . . Equally important are instances in which CVP and RD data diverge. These divergent reports offer opportunities for critical examination and clarification of both the problem solver's knowledge and the CVP and RD methodologies. As a result of using the two methodologies as complementary data sources, the richness of data available on a particular event is enhanced. (417)

In addition to the precautions various authors already cited have offered to increase the validity and reliability of concurrent verbal protocols, Taylor and Dionne (2000) propose additional considerations for limiting threats to the accuracy of retrospective debriefings. These include keeping the focus of questions on neutral and complete reportage; conducting the interview as close as possible in time to the experience itself; stressing with participants the need for accuracy; limiting the number of tasks asked about; focusing when possible on specific, important moments in the verbal protocols; using probes carefully to flesh out detail and check researcher understanding without being leading; and keeping the focus on description rather than interpretation ("'what' and 'which' rather than 'why'"; 417).

Methodological Implications for the Present Study

In a number of ways, the present study closely attends to the critiques levied against and cautions raised concerning the use of verbal protocols as data. First, the study was designed primarily to elicit what Leow and Morgan-Short (2004, 36) referred to as "introspective, nonmetalinguistic verbalizations" by recording participants' concurrent reports of their behaviors while answering test questions. Second, the study was designed to gather retrospective debriefing data, in the form of standardized postexperience interviews with participants, as a secondary data source while paying heed to Taylor and Dionne's (2000) recommendations for limiting reactivity in questioning. Third, the initial instructions given to participants for the concurrent verbal protocols were kept as simple and nondirective (in Taylor and Dionne's words, as "infrequent and neutral"; 415) as possible, and interviewers were directed to prompt students only when they had lapsed into silence for a period of time or were clearly working without verbalizing. Fourth, the tasks posed by the SAT Suite test questions given to participants are sufficiently nonroutine to be likely to evoke conscious, accurate reports of inline processing as participants work through them. Finally, the present study was originally conceived as a follow-up to a previously published cognitive lab study involving a cross section of the SAT Suite test-taking population (College Board 2024a), which meant that the results of a "control" group of sorts would have been available for comparison to the results of this study. However, it proved logistically impossible to administer the same test questions by the same means to the participants in this study as it was to the participants of the prior study and impractical to add a new control group, so the present study has to stand on its own.

Section 3: Methodology

Test Question Selection

College Board subject matter experts began the research process for this study by identifying sets of SAT Suite Reading and Writing and Math test questions that would represent as many of the key skill/knowledge elements of the test sections' designs as possible. Because the designs of and specifications for all SAT Suite tests—the SAT, PSAT/NMSQT®, PSAT™ 10, and PSAT™ 8/9—are intentionally similar (College Board 2024b), the selected questions as sets could fairly be said to represent those encountered in the suite as a whole rather than in just one of the tests.

Consistent with the approach used in a prior cognitive lab study (College Board 2024a), the present study intentionally excluded questions from the Reading and Writing section's Standard English Conventions content domain. Although facility with the conventions of Standard English is highly valued in academic and career settings, the strongly rule-based nature of tasks in this domain makes these questions unlikely to elicit rich responses from students in a verbal protocol setting, and College Board makes no strong claim about the cognitive complexity of these questions. All other Reading and Writing content domains and all Math content domains were represented by multiple test questions in the question sample selected.

Fifteen Reading and Writing questions and fifteen Math questions were ultimately selected for study. These questions were drawn from actual SAT Suite item pools rather than developed specifically for this study and were therefore representative of questions students might encounter on test day. For logistical reasons, all questions used in the study were drawn from a linear (nonadaptive) version of an extant SAT practice test form that had recently been made available to students. This choice increased somewhat the risk that one or more participants would have encountered these questions previously as part of full-form test practice (a point returned to in this report's subsection on study limitations in Section 5: Discussion), but it also ensured that participants were presented with questions in combinations that could organically occur as part of authentic testing (or authentic practice, as the same procedures used to generate operational test forms are used to produce official full-length practice tests).

Collectively, the Reading and Writing and Math question samples represent a wide range of content domains, skill/knowledge testing points, subject areas, question difficulty levels, stimulus text complexities (Reading and Writing only), and question formats consistent with the tests' designs. All questions used in the study, like all those of the SAT Suite, are discrete, meaning that no set-based questions were used and that each question could be answered independently of all others.

Table 1 summarizes the most salient characteristics of the Reading and Writing (RW) and Math test questions presented to participants in this study. An explanation of the table's columns immediately follows.

Table 1. Characteristics of Reading and Writing (RW) and Math Questions Presented to Study Participants.

Test Section	Q#	Content Domain	Skill/Knowledge Testing Point	Subject Area	TC (*RW only*)	PSB	Question Format
Reading and Writing	1	Craft and Structure	Words in Context	SCI	PSR	7	MC
	2		Text Structure and Purpose	LIT	MID	3	MC
	3		Text Structure and Purpose	HSS	PSR	7	MC
	4	Information and Ideas	Command of Evidence: Quantitative	SCI	SCO	4	MC
	5		Command of Evidence: Textual	LIT	SCO	4	MC
	6	Expression of Ideas	Transitions	HSS	SCO	5	MC
	7		Rhetorical Synthesis	HUM	MID	4	MC
	8		Rhetorical Synthesis	SCI	PSR	5	MC
	9	Craft and Structure	Words in Context	SCI	PSR	4	MC
	10		Cross-Text Connections	HUM	SCO	4	MC
	11	Information and Ideas	Central Ideas and Details	LIT	SCO	3	MC
	12		Central Ideas and Details	HUM	PSR	6	MC
	13		Command of Evidence: Textual	SCI	SCO	4	MC
	14		Command of Evidence: Quantitative	SCI	PSR	7	MC
	15		Inferences	HSS	MID	4	MC
Math	1	Algebra	Linear Inequalities: Identify	SCI		4	MC
	2	Problem-Solving and Data Analysis	Ratios	RWT		5	MC
	3	Geometry and Trigonometry	Circles	None		6	MC
	4	Advanced Math	Nonlinear Functions: Rewrite	None		7	MC
	5	Problem-Solving and Data Analysis	Percentages	None		7	MC
	6	Advanced Math	Nonlinear Functions: Make Connections	None		7	MC
	7	Algebra	Linear Functions: Identify	SCI		2	MC
	8	Geometry and Trigonometry	Measures of Angles in a Triangle	None		3	MC
	9	Advanced Math	Nonlinear Functions: Interpret	SCI		4	MC
	10	Problem-Solving and Data Analysis	Scatterplot	None		4	MC
	11	Problem-Solving and Data Analysis	Probability	RWT		4	MC
	12	Advanced Math	Nonlinear Equations: Solve	None		5	SPR
	13	Algebra	Linear Equations in Two Variables: Make Connections	None		5	SPR
	14	Geometry and Trigonometry	Scale Factor and Area	None		6	MC
	15	Algebra	Systems of Two Linear Equations in Two Variables: Solve	None		6	SPR

Table 1 displays key traits of each of the SAT test questions used in this study.

- **Test section.** Reading and Writing or Math
- **Q#.** Question number (1–15), representing the order in which the questions were presented to participants
- **Content domain.** One of the major conceptual divisions within each of the two test sections: Information and Ideas, Craft and Structure, and Expression of Ideas in Reading and Writing; Algebra, Advanced Math, Problem-Solving and Data Analysis, and Geometry and Trigonometry in Math
- **Skill/knowledge testing point.** The skill/knowledge element targeted by the question (e.g., Words in Context in Reading and Writing; Probability in Math)
- **Subject area.** The content area, if any, sampled by the question: literature (LIT), history/social studies (HSS), the humanities (HUM), or science (SCI) in Reading and Writing; science (SCI) or real-world topics (RWT) in Math. (Social studies, a third content area sampled by SAT Suite Math questions, was not represented.) Math questions with a subject area of "None" test aspects of "pure" mathematics outside of context.
- **TC.** Stimulus text complexity. Reading and Writing test passages (only) are formally rated for text complexity by College Board subject matter experts using both quantitative and qualitative means. Passages developed for the section fall into one of three categories:
  - MID: Middle school/junior high school level (equivalent to grades 6–8)
  - SCO: Upper secondary level (grades 9–11)
  - PSR: Postsecondary readiness level (grades 12–14)
- **PSB.** Performance score band, a numerical rating of a question's statistical difficulty aligned to the test sections' scales. In SAT Suite terms, questions in PSBs 1 to 3 are considered easy and are associated with Reading and Writing section scores from 200 (the lowest possible) to 480 and with Math section scores from 200 to 460 (out of 800, in ten-point intervals). Questions in PSBs 4 and 5 are considered medium difficulty and are associated with Reading and Writing section scores from 490 to 600 and with Math section scores from 470 to 600. Questions in PSBs 6 and 7 are considered hard and are associated with Reading and Writing and Math section scores from 610 to 800. Each test section's question sample included questions typically ranging in PSB from 3 to 7; with one exception in Math, questions in PSBs 1 and 2 were excluded from selection, as the research literature (e.g., Bainbridge and Sanderson 1995) suggests that such relatively cognitively simple tasks are unlikely to elicit much conscious thought from test takers.
- **Question format.** All Reading and Writing questions, both in the study and on the actual SAT Suite tests, are in the four-option multiple-choice (MC) format, with each question having a single best answer (*key*). Math questions are either in this same MC format or in the student-produced response (SPR) format, for which students must generate and enter their own answers without the structure and support of provided answer choices.

As a group, the fifteen sampled Reading and Writing questions represented three of the section's four content domains (with Standard English Conventions being excluded, as previously noted), all major skill/knowledge testing points within those three domains, all four sampled subject areas, all three sampled stimulus text complexity levels, and all levels of difficulty from 3 (easy) to 7 (hard). As a group, the Math questions represented all four of the section's content domains, many skill/knowledge testing points within those domains, in-context questions representing two of three sampled subject areas as well as questions set outside of context, all levels of difficulty from 2 to 7, and both multiple-choice and student-produced response formats.

In addition to the fifteen Reading and Writing and fifteen Math questions formally presented to participants, three questions from each section were incorporated into participant training. Before a given participant did their own thinking aloud on the fifteen study questions in either Reading and Writing or Math, the session moderator, following a script, exemplified thinking aloud through a sample question from the same section, after which the participant would have one or (if deemed necessary by the moderator) two opportunities to practice thinking aloud themselves before beginning the actual question set. These training questions were drawn from the same practice test form from which all other questions were taken and can be found in the appendix. The practice portions of sessions were neither recorded nor analyzed.

Question Type–Level Construct Definition

The College Board subject matter experts who selected the questions for the study also identified constructs for the questions by skill/knowledge testing point. These *constructs*, in the form of lists of behaviors demonstrable by test takers, describe the kinds of cognitively complex thinking students are expected to exhibit if they approach answering the questions as intended by the test developers.

For each Reading and Writing testing point (e.g., Words in Context), staff developed a list of behaviors test takers were required to exhibit in order to answer each question as intended. Because many Math questions include, by design, multiple and often mutually exclusive pathways test takers may pursue in answering, these behaviors were defined as expected rather than required, and participants needed only to exhibit at least one of them to be considered as having enacted the construct. Answering correctly was always a required Reading and Writing behavior; for Math, participants' correct and incorrect answers for each question were tracked separately from the behavior list. Additionally, both Reading and Writing and Math staff identified generic sets of common behaviors that skillful test takers may or may not exhibit while answering questions; these optional behaviors were coded for but not analyzed in this report.

These construct definitions (lists of behaviors) can be found with their associated test questions in Section 4: Results.

The constructs (required/expected behaviors) used for this study are highly similar to the ones used in previous research (College Board 2024a), with some refinements made to better reflect learnings from the prior study.

Protocol Development

The lead author of this study, in collaboration with other College Board researchers and vendor Vidlet, Inc., developed closely parallel Reading and Writing and Math protocols for conducting the cognitive interviews in which students would participate. These protocols were designed as guides for the moderators conducting sessions with participants. The guides included general instructions for conducting the sessions, scripts for moderators to follow, and suggested probes and prompts that moderators could use during sessions should participants lapse into extended silence while working through the test questions. Consistent with best practices (as discussed in Section 2: Literature Review), moderators were directed to limit probes and prompts as much as possible and to make them as nondirective as possible (e.g., "Please keep thinking aloud") so as not to unduly influence participants' responses. Moderators were also advised against asking participants to clarify or explain their responses, as such would divert participants from direct, concurrent reporting of their thinking and actions in the moment to less reliable retrospective inferences. Vidlet moderators were briefed and trained on the protocol and given multiple opportunities to provide feedback and suggest refinements.

Test Question Delivery Method

SAT Suite test questions presented to participants during the think-aloud activity (including its training portion) were administered via Bluebook, the custom-built test application College Board uses to give the SAT Suite tests in their standard digital-adaptive form. The use of Bluebook, which most students use to take SAT Suite tests and engage in full-form practice, enhanced the study's verisimilitude, gave participants ready and standardized access to the universal tools available in Bluebook (including a built-in version of the Desmos® graphing calculator for the Math section), and overall represented a methodological improvement relative to the prior cognitive lab study investigating students' interactions with the digital-adaptive SAT Suite (College Board 2024a), but it did come with its own limitation. In contrast to the prior study, in which only the focal test questions (and training questions) were presented via a third-party digital survey tool, participants in this study had to "skip around" to the specific focal questions, as directed by a moderator following the protocol script. On very rare occasions, this resulted in participants being misdirected to an "incorrect" question (i.e., one in the test form being used but not one of the focal questions); these few instances, as well as a small number of additional cases in which participants ran out of time to answer particular questions, are effectively discounted by the methodology, as the metrics calculated consider only numbers and proportions of correctly answered questions. To account for the fact that the digital-adaptive test sections are divided into two separately timed modules and that test takers can't return to the first module once they've moved on to the second, moderators were directed to inform participants they could review their responses (or lack of responses) to the focal questions in the first module before advancing to the second.

Tools Available to Participants

All participants in both the Reading and Writing and Math segments of this study had access to the full range of universal tools available in Bluebook (see College Board 2024b, section 2.2.7.2). This suite of tools includes a graphing calculator built into the app and available for the Math section (only); alternatively, participants could make use of their own handheld calculators, provided those devices conformed to College Board's SAT Suite calculator policy. In addition, participants in either the Reading and Writing or Math segment could use a third-party screen reader. (At the time the cognitive interviews were conducted [2024], Bluebook didn't have a native text-to-speech option as an alternative to the use of screen readers; this feature was added in 2025.) Only one participant across the Reading and Writing and Math activities (RW14) used a screen reader during the session. (For context, six participants—three each in Reading and Writing and Math—indicated via the screener that they'd already received or expected to receive an assistive technology accommodation as part of SAT Suite testing, and RW14 was not among them.)

Sample Definition, Selection Criteria, Recruitment, and Characteristics

SAMPLE DEFINITION

For its 2024 cognitive lab studies, College Board sought members of the SAT test-taking population who fit into one (or possibly more) of three categories: students with a specific learning disorder affecting reading (abbreviated here as SLDR; also known as dyslexia), students with attention deficit hyperactivity disorder (ADHD) (College Board 2025a), and students who were English learners (College Board 2025b). The present study reports the results of the study involving students who have SLDR.

The *Diagnostic and Statistical Manual of Mental Disorders*, fifth edition, text revision (DSM-5-TR) (American Psychiatric Association 2022, 77) observes that people diagnosed with a specific learning disorder with impairment in reading have issues with word reading accuracy, reading rate or fluency, and/or reading comprehension. Having any or all of these factors would likely impact test-taking performance on the SAT Suite Reading and Writing section, whose construct is defined as "literacy achievement relative to core college and career readiness requirements in English language arts as well as in the academic disciplines of literature, history/social studies, the humanities, and science" (College Board 2024b, 54). The same would likely be true, albeit probably to a lesser extent, for the Math section, as all its test questions, even those testing aspects of "pure" mathematics, include some linguistic component for framing purposes and roughly 30 percent of questions are set in an English-language context in social studies, science, or a real-world topic, which students must read and analyze to properly answer (College Board 2024b).

As part of the sample selection screener (see appendix), prospective participants were asked to indicate whether they had SLDR (and/or had ADHD or were an English learner). If the answer was "yes," they were further asked to indicate how they were diagnosed with SLDR (formal assessment by a specialist, screening conducted by a teacher or educational professional, or self-diagnosis or diagnosis by a family member) and to describe the impact of their SLDR symptoms in the context of test taking (mild, moderate, or severe, with provided operational definitions discussed subsequently in this report). Students who answered "yes" to the SLDR question and met other selection criteria (see next subsection) were considered eligible for this study, and no further documentation or other evidence of their condition was requested or collected. This approach sidestepped thorny issues of condition definition and minimized medical privacy concerns but did raise the possibility that one or more participants would self-identify as having SLDR when they didn't merely to participate in the study and receive its incentive. To militate against this possibility, the screener didn't specify that students with SLDR were being sought, and the initial query about having SLDR was mingled with other possible conditions and statuses, including ones not expressly sought for this or other studies (e.g., students who are deaf or hard of hearing). As it turned out, observations of students in both the Reading and Writing and Math conditions gave no evidence of self-misidentification.

Prospective participants were also asked whether they had received or expected to receive accommodations as part of SAT Suite testing and, if so, to identify them (extended time on exams, extended breaks, assistive technology [e.g., text-to-speech software], or other). This is important because the provision of appropriate testing accommodations for universally designed exams is the key means by which fairness on the SAT Suite is ensured for students with disabilities (College Board 2024b). Any response to these two questions, including ones indicating that students hadn't received or didn't expect to receive accommodations, was considered acceptable for sample selection.

SAMPLE SELECTION CRITERIA

Prospective participants were deemed eligible for selection if they met the following criteria:

- They were students in either grade 11 or 12.
- They attended school in the United States.
- They answered "yes" when asked whether they had SLDR.
- They provided other required demographic information, including gender, race/ethnicity, and self-reported high school GPA (HSGPA).[1]

  Note: Students were allowed to indicate that they preferred not to respond to the gender and/or race/ethnicity questions without being excluded from consideration.
- They were willing and able to productively participate in a virtual cognitive interview session of up to 120 minutes in length.

Self-reported HSGPA was used as the proxy for student academic achievement in this study. This was necessary because, as discussed immediately below, Vidlet operated as the primary student recruiter, and it was therefore not possible, for logistical and privacy reasons, to link prospective participants to any previous SAT Suite scores they may have had on file with College Board. Students were asked

[1] One Reading and Writing participant (RW13) failed to provide HSGPA and was inadvertently included in the study.

on the screener to report past SAT or PSAT-related test scores, but doing so was not a requirement, and as all students (with one exception in Reading and Writing) provided HSGPA information while not all students provided self-reported SAT/PSAT test scores, the latter weren't considered in this study. This is theoretically a limitation of the study, but evidence (e.g., Sanchez and Buddin 2016) suggests that self-reported HSGPAs are generally sufficiently accurate for research purposes.

SAMPLE RECRUITMENT

In June 2024, College Board contacted vendor Vidlet, Inc., an organization that had successfully aided in a prior cognitive lab study (College Board 2024a), to support a research initiative to learn more about how students from various subpopulations of interest—students with a specific learning disorder affecting reading (SLDR, also known as dyslexia), students with attention deficit hyperactivity disorder (ADHD), and students who are English learners—experience SAT Suite testing.

Prior to recruitment, College Board and Vidlet jointly worked on a sample selection screener (survey) that would be given electronically to prospective participants to complete (see appendix). This screener was designed to collect eligibility information as well as a limited range of demographic detail (e.g., grade in school, gender, race/ethnicity) intended to ensure breadth in sample selection. Demographic survey items deemed potentially sensitive (e.g., gender, race/ethnicity) included a "prefer not to respond" option, and choosing this didn't disqualify the candidate from consideration.

Also prior to recruitment, College Board determined that an incentive of $150 per participant would fairly compensate students for their time and effort. This incentive would come in the form of a gift card, which could be used in a variety of ways.

Vidlet recruited students primarily through its panel and email outreach processes; a small number of additional potential contacts were provided by College Board. The recruitment solicitation (see appendix) highlighted that participants would have an opportunity to provide feedback to influence SAT testing and that they'd receive an incentive of $150 on successful completion of the activity. After initial intake by the Vidlet team, participant information was de-identified and sent to College Board to ensure as diverse a selection as possible (given small sample sizes) by gender, race/ethnicity, geography, and self-reported HSGPA. Recruitment occurred on a rolling basis, meaning that some students were interviewed while others were still being identified.

Once students had confirmed their participation in the study, Vidlet collected consent forms (see appendix). These consent forms, which were either signed by students themselves (if they were age eighteen or over) or a parent/guardian (if not), described the nature of the activity, explained what participants would be asked to do, and made participants aware that they could opt out of some or all of the activity for any reason if they so chose (although successful completion of the activity was required to receive the incentive).

Participants were then assigned randomly by Vidlet to either the Reading and Writing or Math activity. Each activity consisted of two main elements: (1) a think-aloud portion, in which participants shared their thoughts concurrently as they worked through a set of SAT Suite test questions and (2) a postexperience interview using a standardized set of questions focused on participants' impression of the think-aloud activity as well as self-identified sources of challenge in answering particular questions or categories of questions. Collectively, these components were scheduled to take no more than 120 minutes.

Recruitment and interviewing for this SLDR-focused study took place concurrently with recruitment and interviewing for two other cognitive lab studies: those involving students with ADHD (College Board 2025a) and students who were English learners (College Board 2025b). Over the course of approximately four months, the Vidlet research team led a total of about 120 students, divided roughly equally across the three subgroups of interest, through cognitive interview sessions structured according to protocol documents developed by College Board and vetted by Vidlet. No student was allowed to participate in more than one study.

SAMPLE CHARACTERISTICS

Reading and Writing

Table 2 displays the roster of Reading and Writing participants. For each participant, the table includes the participant identifier (a unique code used in place of a student's name); demographic information, including the participant's gender, race, ethnicity, home state, grade in school, and self-reported HSGPA; and information about the participant's SLDR condition, including whether they have SLDR (always "yes," with the exception of one "no response"), which SAT Suite testing accommodations they have already received or expected to receive, and a rating of the impact of their SLDR symptoms on test taking (mild, moderate, severe; definitions discussed below).

Table 2. Reading and Writing Participant Roster by Demographics, SAT Suite Accommodations Status, and Self-Reported SLDR Impact on Test-Taking Ability.

	Demographics						SAT Suite Accommodations		
Part. ID	**Gender**	**Race**	**Ethnicity**	**Home State**	**Grade in School**	**Self-Reported HSGPA**	**Received/ Expected?**	**Type(s)**	**Self-Reported SLDR Impact**
RW4	Female	White	Mexican	IL	11	B (83–86)	Yes	ET, EB	Moderate
RW8	Female	Black or African American	Not of Hispanic, Latino, or Spanish origin	LA	11	B+ (87–89)	Yes	ET, EB	Moderate
RW11	Female	White	Not of Hispanic, Latino, or Spanish origin	NY	12	A (93–96)	*NR*	*NR*	*NR*
RW13	Male	White	Not of Hispanic, Latino, or Spanish origin	IL	12	*NR*	Yes	ET, EB	Moderate
RW14	Female	White	Not of Hispanic, Latino, or Spanish origin	OH	11	B (83–86)	Yes	ET, EB	Moderate
RW22	Female	White	Not of Hispanic, Latino, or Spanish origin	VA	11	B (83–86)	Yes	ET	Moderate
RW23	Male	Black or African American	Not of Hispanic, Latino, or Spanish origin	FL	11	C+ (77–79)	Yes	ET	Mild
RW30	Male	*NR*	*NR*	TX	11	B– (80–82)	Yes	ET, EB, AT	Moderate
RW32	Male	Black or African American	Not of Hispanic, Latino, or Spanish origin	LA	12	B (83–86)	Yes	ET, EB	Moderate
RW34	Male	White	Hispanic, Latino, or Spanish origin other than Cuban, Mexican, or Puerto Rican	TX	11	B (83–86)	Yes	ET, EB	Mild
RW36	Male	Black or African American	Not of Hispanic, Latino, or Spanish origin	GA	11	A (93–96)	Yes	ET, EB	Moderate
RW39	Female	Other (White and Black)	Not of Hispanic, Latino, or Spanish origin	VA	11	B (83–86)	Yes	ET, AT	Moderate
RW40	Female	White	Mexican	NM	11	D (65–66)	Yes	ET, EB, AT	Moderate
RW41	Male	White	Not of Hispanic, Latino, or Spanish origin	NY	11	D+ (67–69)	Yes	ET	Moderate
RW43	Female	Black or African American	Not of Hispanic, Latino, or Spanish origin	TX	12	A– (90–92)	Yes	ET	Severe

NR: No response

SAT Suite Accommodations Types:

ET = Extended time
EB = Extended breaks
AT = Assistive technology (e.g., text-to-speech)

Definitions for Self-Reported SLDR Impact on Test-Taking Ability:

Mild = Symptoms are manageable and have minimal impact on test performance
Moderate = Symptoms interfere with test taking but can be managed with accommodations
Severe = Symptoms significantly impair test-taking ability even with accommodations

Table 2 suggests that within the strictures of the small sample size (*n* = 15) and self-selection methodology used for this study, Reading and Writing participants represented a relatively diverse sample in terms of gender, race/ethnicity, grade in school, self-reported HSGPA, and, to a lesser extent, SLDR symptom impact on test-taking ability. Specifically:

- **Gender.** An approximately equal proportion of female (eight) and male students (seven) participated.
- **Race and ethnicity.** Most numerous were White participants not of Hispanic, Latino, or Spanish origin (five) and African American or Black participants not of Hispanic, Latino, or Spanish origin (five), which together accounted for ten of the fifteen participants. Two White students of Mexican ethnicity; a White student of Hispanic, Latino, or Spanish origin other than Cuban, Mexican, or Puerto Rican; a student identifying as White and Black; and a student who declined to respond composed the rest of the sample. In total, nine participants identified as having a race/ethnicity combination other than White/Not of Hispanic, Latino, or Spanish origin, although several racial categories (Asian; Native Hawaiian or other Pacific Islander; Native American or Alaska Native) weren't represented, which constitutes a limitation on the study (see Section 5: Discussion).
- **Grade in school.** Students from both grade 11 (eleven) and grade 12 (four) were represented.
- **Self-reported HSGPA.** Three participants indicated an "A" HSGPA, eight indicated a "B" HSGPA, one indicated a "C" HSGPA, two reported a "D" HSGPA, and one didn't provide a response. This represents a fairly broad range of high school achievement in line with the study's goal to be as representative as possible, within small sample size and self-selection limitations, of the SLDR subpopulation. While the sample is seemingly biased toward higher HSGPAs, this outcome should be considered in the context of grade inflation generally (e.g., Sanchez 2024), which suggests that we should expect to see fewer students overall with average-and-below HSGPAs.
- **SLDR symptom impact.** Most participants (eleven) indicated that their SLDR symptoms had moderate impact on their test-taking ability; two participants indicated a mild impact, one participant indicated a severe impact, and one participant declined to respond.

Math

Following the same approach as for the Reading and Writing participant roster, table 3 displays the roster of Math participants.

Table 3. Math Participant Roster by Demographics, SAT Suite Accommodations Status, and Self-Reported SLDR Impact on Test-Taking Ability.

	Demographics						SAT Suite Accommodations		
Part. ID	Gender	Race	Ethnicity	Home State	Grade in School	Self-Reported HSGPA	Received/ Expected?	Type(s)	Self-Reported SLDR Impact
M10	Male	White	Not of Hispanic, Latino, or Spanish origin	OH	11	B+ (87–89)	Yes	ET, EB	Moderate
M11	Male	White	Hispanic, Latino, or Spanish origin other than Cuban, Mexican, or Puerto Rican	GA	11	B− (80–82)	Yes	ET	Moderate
M13	Female	White	Not of Hispanic, Latino, or Spanish origin	NC	12	B (83–86)	Yes	ET, EB	Severe
M16	Female	White	Puerto Rican	CT	11	B+ (87–89)	Yes	ET	Mild
M20	Male	Other (Caribbean)	Not of Hispanic, Latino, or Spanish origin	NV	11	C− (70–72)	No	–	Mild
M21	Female	Asian	Not of Hispanic, Latino, or Spanish origin	NC	11	C+ (77–79)	Yes	ET, EB	Moderate
M22	Female	White	Not of Hispanic, Latino, or Spanish origin	NC	11	B− (80–82)	Yes	ET, EB	Mild
M24	Male	White	Not of Hispanic, Latino, or Spanish origin	MO	12	A− (90–92)	Yes	ET, EB	Moderate
M26	Female	White	Hispanic, Latino, or Spanish origin other than Cuban, Mexican, or Puerto Rican	NM	11	B+ (87–89)	Yes	ET	Mild
M27	Male	White	Not of Hispanic, Latino, or Spanish origin	TX	12	A (93–96)	Yes	ET	Moderate
M28	Female	White	Not of Hispanic, Latino, or Spanish origin	FL	12	B+ (87–89)	Yes	ET	Mild
M32	Male	White	Not of Hispanic, Latino, or Spanish origin	OH	11	C (73–76)	Yes	ET, EB	Moderate
M33	Male	White	Not of Hispanic, Latino, or Spanish origin	NC	11	A+ (97–100)	Yes	ET	Moderate
M34	Female	White	Not of Hispanic, Latino, or Spanish origin	TX	11	A− (90–92)	Yes	ET	Moderate
M38	Male	White	Not of Hispanic, Latino, or Spanish origin	TX	12	B (83–86)	Yes	ET	Moderate
M39	Male	White	Not of Hispanic, Latino, or Spanish origin	OH	12	B (83–86)	Yes	ET, AT	Moderate
M44	Male	Asian	Not of Hispanic, Latino, or Spanish origin	NC	11	A− (90–92)	Yes*	ET, EB	Moderate
M46	Male	White	Not of Hispanic, Latino, or Spanish origin	SD	12	B (83–86)	Yes	ET, AT	Severe
M56	Male	White	Not of Hispanic, Latino, or Spanish origin	PA	11	B (83–86)	Yes	ET	Moderate
M58	Male	Black or African American	Not of Hispanic, Latino, or Spanish origin	TN	12	C− (70–72)	Yes	ET, AT	Severe
M60	Female	Black or African American	Not of Hispanic, Latino, or Spanish origin	IN	11	B+ (87–89)	Yes	ET	Moderate

* Participant M44 replied “no” to the “Received/expected SAT Suite accommodations?” question but listed expected/received accommodations.

SAT Suite Accommodations Types:

ET = Extended time
EB = Extended breaks
AT = Assistive technology (e.g., text-to-speech)

Definitions for Self-Reported SLDR Impact on Test-Taking Ability:

Mild = Symptoms are manageable and have minimal impact on test performance

Moderate = Symptoms interfere with test taking but can be managed with accommodations

Severe = Symptoms significantly impair test-taking ability even with accommodations

Table 3 suggests that within the strictures of the small sample size (*n* = 21) and self-selection methodology used for this study, Math participants represented a somewhat diverse sample in terms of gender, race/ethnicity, grade in school, self-reported HSGPA, and SLDR symptom impact on test-taking ability. Specifically:

- **Gender.** Male participants (thirteen) were somewhat disproportionately represented relative to female participants (eight).
- **Race and ethnicity.** Thirteen participants identified as White and not of Hispanic, Latino, or Spanish origin. Three participants identified as White and of Hispanic, Latino, or Spanish origin. Two participants identified as Black or African American and not of Hispanic, Latino, or Spanish origin. Two participants identified as Asian and not of Hispanic, Latino, or Spanish origin. One participant identified as Caribbean and not of Hispanic, Latino, or Spanish origin. Two racial categories (Native Hawaiian or other Pacific Islander; Native American or Alaska Native) weren't represented, which constitutes a limitation on the study (see Section 5: Discussion).
- **Grade in school.** Students from both grade 11 (thirteen) and grade 12 (eight) were represented.
- **Self-reported HSGPA.** Five participants indicated an "A" HSGPA, twelve indicated a "B" HSGPA, and four reported a "C" HSGPA. This represents a fairly broad achievement range in line with the study's design and the limitations of both small sample size and the self-selection methodology and in an environment of grade inflation (e.g., Sanchez 2024).
- **SLDR symptom impact.** Most participants (thirteen) indicated that their SLDR symptoms had moderate impact on their test-taking ability; five participants indicated a mild impact, and three participants indicated a severe impact.

Coding and Analysis

CODING

The lead College Board researcher uploaded the interview transcripts generated by Vidlet into MAXQDA, a qualitative/mixed-methods research software package. Reading and Writing and Math teams, using MAXQDA's cloud service, then coded each transcript against the previously defined required (Reading and Writing) / expected (Math) and optional behaviors associated with the question types' constructs. In cases in which transcripts were vague or ambiguous (e.g., the participant didn't verbalize the answer they selected or entered but had answered in Bluebook), the research team consulted the video recordings to confirm participant behaviors and answer choices.

Team members were also directed to code as "vignette candidates" any participant response that exhibited all required behaviors (Reading and Writing) / at least one expected behavior (Math) and that served to illustrate well-reasoned responses without significant errors, omissions, or uncorrected missteps. We elected to adopt a "case study" approach for the presentation of such vignettes in Section 4: Results, sharing transcript excerpts from a single participant in Reading and Writing and in Math and supplementing those excerpts with those from other participants when the case study participant failed to demonstrate adequate behaviors and/or failed to answer a given question correctly. In the few cases in which no participant answered a given question correctly and exhibited appropriate behaviors, no supplementary vignette was incorporated.

As a supplement to MAXQDA, the team concurrently recorded, in Microsoft Excel, whether each participant had answered each question correctly and exhibited each of the required/expected behaviors for the questions; these Excel spreadsheets served as the basis for tabulating the statistics presented in Section 4: Results. The coding process resulted in approximately nineteen hundred codes being assigned to forty-six participants' interactions with the thirty studied questions across Reading and Writing and Math.

ANALYSIS

The College Board researchers then analyzed the coded data to assess in various ways both participant and test question performance, as elicited from the think-aloud activity, as well as participant perceptions of their simulated test-taking experience, as elicited from postexperience interview questions.

1. **Participant performance** was analyzed in terms of the number and proportion of correctly answered questions for which participants demonstrated appropriate cognitive behaviors. Vignettes (transcript excerpts) from select participants were used when available to illustrate demonstrations of the cognitively complex thinking elicited by the test questions.
2. **Question performance** was analyzed in terms of the number and proportion of correctly answering participants who also demonstrated appropriate cognitive behaviors.
3. **Participant perceptions** of the question-answering activity, in the form of responses to postexperience interview questions, were analyzed for both general themes and for any cases in which participants identified potential construct-irrelevant barriers to their success in the activity and to SAT Suite test taking more broadly.

Each of these approaches is discussed in turn below.

Participant Performance

Participant performance on each Reading and Writing or Math question was assigned a *performance level* (PL) from 1 to 5 based on two intersecting considerations: whether the participant answered the question correctly and whether appropriate behaviors were demonstrated.

Table 4 displays the definitions of the five performance levels in Reading and Writing and in Math.

Table 4. Participant Performance Level (PL) Definitions.

Performance Level	Definition	
	Reading and Writing	Math
1 (highest)	Answered correctly; demonstrated all required behaviors	Answered correctly; demonstrated at least one expected behavior
2	Answered correctly; demonstrated fewer than all required behaviors	*Not applicable; see below*
3	Answered correctly; demonstrated no other required behaviors	Answered correctly; demonstrated no expected behaviors
4	Answered incorrectly; demonstrated at least one other required behavior	Answered incorrectly; demonstrated at least one expected behavior
5 (lowest)	Answered incorrectly; demonstrated no other required behaviors	Answered incorrectly; demonstrated no expected behaviors

PL 2 is present in Reading and Writing and unobtainable in Math given the previously discussed differences between required (Reading and Writing) and expected (Math) behaviors, as Math participants received a PL of 1 if they demonstrated at least one expected behavior. PL 2 was also unobtainable in Reading and Writing when a given question type had only two required behaviors, one of which was (always) answering correctly.

In Section 4: Results, performance levels are displayed in figures, with each cell representing a participant-by-question interaction. PLs are indicated by number (1–5) and by supplementary color shading, with shades of blue indicating PLs 1 through 3 and shades of orange indicating PLs 4 and 5. Unobtainable PL 2s are indicated by a dash ("–").

Using these performance level findings, the research team calculated what this study refers to as the *participant differential*, or D_p, for each participant. Mathematically, D_p is represented by the following formulas:

$$\text{Reading and Writing: } D_p = \#AC - \#RB$$

$$\text{Math: } D_p = \#AC - \#EB$$

In these formulas, D_p is the participant differential, *#AC* is the total number of questions a given participant answered correctly, and *#RB* and *#EB* are, respectively, the number of correctly answered questions for which the participant also demonstrated all required behaviors (Reading and Writing) or at least one expected behavior (Math). D_p is always either zero or a positive integer except in the rare circumstance (not encountered in this particular study) in which a participant answered no questions correctly, in which case no "true" differential exists. In performance level terms, *#RB* and *#EB* represent PL 1.

Conceptually, D_p represents the "difference" between simply answering questions correctly and doing so while also exhibiting the cognitive behaviors intended by the test developers. D_p is thus a more appropriate and robust measure of participant performance than the raw number of questions answered correctly because D_p, in essence, removes from consideration those questions that participants may have answered correctly by means other than those intended by the test makers (e.g., by random guessing or by finding a "shortcut" past the intended intellectual activity).

Zero or low participant differentials are desirable, as ideally each participant answered questions correctly only by enacting the question types' constructs. Owing to the sometimes variable number of participants who answered each Reading and Writing or Math question, the threshold for a "good" differential is set at 70 percent or greater—meaning, for example, that if a participant answered all fifteen Reading and Writing or Math questions correctly, they would also have needed to have demonstrated all required behaviors on at least eleven of these questions (73 percent) to yield a "good" differential (in this example, 4 or lower). The "70 percent or greater" threshold is somewhat arbitrary, but it does represent a significant majority of correctly answered questions being responded to in ways that enact the question type–level constructs while at least partially accounting for the possibility that a given participant may well have understood how to "properly" answer a particular question but may simply have not verbalized one or more elements of doing so (essentially "underreporting" their skills and knowledge owing to the artificiality of the simulated testing experience and/or their lack of familiarity and comfort with thinking aloud).

To illustrate and concretize the cognitively complex thinking required to answer each of the studied test questions, the research team identified during coding cases in which participants exhibited exemplary (if not necessarily "perfect") reasoning in accordance with the question type's construct. These "vignettes" (transcript excerpts) are presented primarily in the form of a case study of a single participant as they answered each of the Reading and Writing or Math questions. For questions for which the case-study participant failed to demonstrate appropriate behavior(s), supplementary vignettes from other participants are provided when available.

Question Performance

The performance of the test questions themselves in the study can also be subjected to an analysis similar to that used for participant performance. To assess question performance, the research team calculated what this study refers to as the *question differential* (D_q), which can be represented by the following formulas:

$$\text{Reading and Writing: } D_q = \#AC - \#RB$$

$$\text{Math: } D_q = \#AC - \#EB$$

In these formulas, D_q is the question differential, *#AC* is the total number of participants answering a given question correctly, and *#RB* and *#EB* are, respectively, the number of correctly answering participants who also demonstrated all required behaviors (Reading and Writing) or at least one expected behavior (Math). In performance level terms, *#RB* and *#EB* again represent PL 1.

Conceptually, D_q is closely analogous to D_p in that the former "discounts" from consideration instances in which participants correctly answered a given question without demonstrating appropriate cognitive behaviors. Zero to low differentials are again considered desirable, a result of no "true" differential could occur (as it did in one case in Math in this study) when no participant answered a given question correctly, and the same 70 percent-or-greater threshold for “good” differentials applies here, this time meaning that for each question, 70 percent or more of correctly answering participants also demonstrated all required behaviors/at least one expected behavior. Like D_p, D_q is concerned with the number of answered questions only, thus mitigating the effect of omitted responses.

Participant Perceptions

All participants were asked the following postexperience interview questions immediately after completing the think-aloud activity in Reading and Writing or Math:

1. Please tell me a bit about the experience you just had. What was it like to answer those questions?
2. How would you describe your general approach, in terms of strategies, for answering the questions?
3. Was there a particular type of question that you found especially easy to answer? If so, which one and why?
4. Was there a particular type of question that you found especially hard to answer? If so, which one and why?
5. Did you encounter anything in the questions that you had difficulty with given that you have a specific learning disorder affecting reading? If so, what was it, and why was it difficult for you?
6. Is there anything about your test-taking experience today or about the test-taking strategies you used today that we haven’t talked about yet but that you’d like us to know?

Questions 1 and 6 were designed as open-ended prompts for participants to share anything on their minds about the think-aloud experience. Question 2 concerned general test-taking strategies used in the think-aloud activity. Questions 3 and especially 4 and 5 were more precisely targeted to elicit participant perceptions of potential construct-relevant and construct-irrelevant impediments to their successful performance in the activity.

Participants’ responses to these postexperience interview questions are summarized in Section 4: Results.

Section 4: Results

Reading and Writing

PARTICIPANT AND QUESTION PERFORMANCE

Participant and Question Performance Levels and Differentials

Figure 1 displays, as a single matrix, the Reading and Writing participant and question performance data derived from this study. The intended method of reading the figure is discussed immediately following.

Figure 1. Reading and Writing Participant and Question Performance Summary Matrix.

Part. ID	Question #															Performance by Level, by Participant						Participant Performance Summary		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	1	2	3	4	5	*NR*	*#AC*	*#RB*	D_p
RW4	5	1	4	4	1	4	2	4	1	1	1	1	1	–	–	7	1	0	4	1	2	8	7	1 ✔
RW8	3	1	4	4	1	5	1	1	1	1	1	4	1	–	1	9	0	1	3	1	1	10	9	1 ✔
RW11	4	1	4	4	1	1	4	1	1	1	1	1	–	–	–	8	0	0	4	0	3	8	8	0 ✔
RW13	3	1	5	2	5	5	4	5	4	5	1	5	5	5	5	2	1	1	2	9	0	4	2	2 ✘
RW14	4	1	4	1	1	3	4	4	1	5	1	4	1	4	1	7	0	1	6	1	0	8	7	1 ✔
RW22	5	1	4	1	1	5	5	1	1	1	1	4	4	2	1	8	1	0	3	3	0	9	8	1 ✔
RW23	3	1	5	1	5	5	1	2	5	1	1	4	5	2	1	6	2	1	1	5	0	9	6	3 ✘
RW30	5	1	4	1	1	5	1	1	1	1	1	4	1	4	3	9	0	1	3	2	0	10	9	1 ✔
RW32	3	1	3	1	1	–	5	1	5	2	1	1	4	–	4	6	1	2	2	2	2	9	6	3 ✘
RW34	4	1	4	1	1	5	5	2	1	5	1	5	1	5	1	7	1	0	2	5	0	8	7	1 ✔
RW36	5	1	5	1	1	4	1	1	1	2	4	5	1	5	1	8	1	0	2	4	0	9	8	1 ✔
RW39	5	1	5	2	4	5	5	2	1	5	1	5	1	5	4	4	2	0	2	7	0	6	4	2 ✘
RW40	5	1	5	5	1	5	–	1	3	3	3	5	–	5	5	3	0	3	0	7	2	6	3	3 ✘
RW41	4	1	1	4	1	5	4	4	1	1	1	4	1	1	4	8	0	0	6	1	0	8	8	0 ✔
RW43	5	1	5	5	1	5	2	5	5	4	3	5	4	4	4	2	1	1	4	7	0	4	2	2 ✘

Performance by Level, by Question

1	0	15	1	7	12	1	4	7	10	7	12	3	8	1	6
2	–	–	–	2	0	–	2	3	–	2	–	–	0	2	–
3	4	0	1	0	0	1	0	0	1	1	2	0	0	0	1
4	4	0	7	4	1	2	4	3	1	1	1	6	3	3	4
5	7	0	6	2	2	10	4	2	2	4	0	6	2	5	2
NR	0	0	0	0	0	1	1	0	0	0	0	0	2	4	2

Performance Level Legend

1 (highest): Answered correctly; exhibited all behaviors

2: Answered correctly; exhibited fewer than all other behaviors

3: Answered correctly; exhibited no other behaviors

4: Answered incorrectly; exhibited other behaviors

5 (lowest): Answered incorrectly; exhibited no other behaviors

Question Performance Summary

#AC	4	15	2	9	12	2	6	10	11	10	14	3	8	3	7
#RB	0	15	1	7	12	1	4	7	10	7	12	3	8	1	6
D_q	4 ✘	0 ✔	1 ✘	2 ✔	0 ✔	1 ✘	2 ✘	3 ✔	1 ✔	3 ✔	2 ✔	0 ✔	0 ✔	2 ✘	1 ✔

Summary Legend

#AC = # answered correctly

#RB = # answered correctly; demonstrated all other behaviors

D_p, D_q = Differentials (*#AC* – *#RB*); ✔ = criterion-passing differential (70%+), ✘ = criterion-failing differential (<70%)

In the top-left portion of the figure, participants are listed in the far-left column ("Part. ID") and questions in the topmost row ("Question #"). Each cell created by the intersection of a row and column represents the performance of a single participant on a given test question (i.e., a participant-by-question interaction). Five performance levels, numbered 1 through 5 (further identified with color shading) and defined in Section 3: Methodology, are used to indicate how a particular participant did on a particular question. A "1" in a cell represents the most successful outcome (a participant answering correctly and demonstrating all required behaviors), while a "5" represents the least successful outcome (a participant answering incorrectly and demonstrating no other required behaviors). A dash ("–") in one of these cells indicates that the participant didn't answer the question. (This could be because they ran out of time, attempted the question but didn't complete it, or, in rare cases, were misdirected from the question by the moderator.)

From this participant-by-question portion, the matrix can be read either horizontally for a summary of participant performance or vertically for a summary of question performance.

Participant Performance

The "Performance by Level, by Participant" sub-table (top center) shows the number of questions answered by each participant in terms of performance levels attained (including *NR* / no response). The "Participant Performance Summary" sub-table (top right) indicates the total number of questions each participant answered correctly (*#AC*), the number of questions each participant answered correctly while also demonstrating all required behaviors (*#RB*), and the participant differential (D_p), or the arithmetic difference between *#AC* and *#RB*. Cells in the "D_p" column include a symbol and are shaded to indicate whether a given participant differential met or exceeded (✔; blue) or fell below (✘; orange) the threshold for a "good" differential. Recall that Section 3: Methodology defines a good D_p as one indicating that at least 70 percent of a participant's correctly answered questions were responded to using all required behaviors, a statistic derived by dividing *#RB* by *#AC*.

Example: Participant Performance

Performance by Level, by Participant

1	2	3	4	5	*NR*
7	1	0	4	1	2

Participant Performance Summary

#AC	*#RB*	D_p
8	7	1 ✔

Participant RW4, per the top row in “Performance by Level, by Participant” sub-table, attained PL 1 (the most successful outcome) on seven questions, PL 2 on one question, PL 4 on four questions, and PL 5 (the least successful outcome) on one question, while not answering two other questions (as indicated by the “2” in the “NR / no response” cell). Turning to the “Participant Performance Summary” sub-table, we find that RW4 answered a total of eight questions correctly (*#AC*, calculated by adding together the number of question responses attaining PLs 1, 2, and 3) and answered seven of those questions correctly while demonstrating all required behaviors (*#RB*, which is the same as the number in the “PL 1” cell in the “Performance by Level, by Participant” sub-table). This results in a participant differential (D_p) of 1, as 8 (*#AC*) minus 7 (*#RB*) equals 1. This D_p exceeds the threshold for a “good” differential, hence the checkmark and blue shading, as *#RB* divided by *#AC* equals .875, or 87.5 percent, which is above the 70 percent cutoff.

Question Performance

The “Performance by Level, by Question” sub-table (center left) shows for each test question the number of participants whose responses attained each of the five performance levels (plus *NR* / no response). A dash (“–”) in cells for PL 2 indicates cases in which that level is unobtainable due to there being only two potential behaviors being evaluated for that question type. The “Question Performance Summary” sub-table (bottom left) indicates the total number of participants answering each question correctly (*#AC*), the number of correctly answering participants who also exhibited all required behaviors (*#RB*), and the question differential (D_q), or the arithmetic difference between *#AC* and *#RB*. Cells in the “D_q” row include a symbol and are shaded to indicate whether a given question differential met or exceeded (✔; blue) or fell below (✘; orange) the threshold for a “good” differential.

Findings

Participant Performance

As shown in the “Participant Performance Summary” sub-table of figure 1, nine of fifteen participants (60 percent) met or exceeded the criterion for a good D_p, which provides evidence that these participants were able to adequately demonstrate cognitively complex thinking in line with the question types’ constructs. These participants were also among the most successful in terms of raw question-answering performance, responding correctly to eight, nine, or ten questions out of a potential fifteen.

The performance of the remaining six participants failed to meet the criterion for a good differential. For example, participant RW13 answered four questions correctly and demonstrated all required behaviors for two of those questions, resulting in a D_p of 2, representing 50 percent of the correctly answered questions. That these criterion-failing participants were also generally lower achieving in the activity than their criterion-meeting peers—answering as many as nine but as few as four questions correctly—suggests that the former had lower levels of appropriate subject matter knowledge, a clearly test construct–relevant consideration. Reinforcing a subject matter–based conclusion is the fact that even these participants were able to demonstrate all required behaviors on half to two-thirds of the questions they did answer correctly. While this performance fell below the criterion, it nonetheless indicates that these participants were able to demonstrate cognitively complex thinking in line with the question types' constructs at least some of the time and suggests that their differentials may have been at least partially a product of a relative lack of ability to verbalize their thinking processes in consistently clear and effective ways.

Question Performance

As shown in the "Question Performance Summary" sub-table of figure 1, ten of the fifteen studied Reading and Writing questions (67 percent) met or exceeded the criterion for a good D_q, which provides evidence that these questions are capable of eliciting cognitively complex thinking from students with SLDR. Of the remaining five questions (questions 1, 3, 6, 7, and 14), all but question 1 were still answered correctly by at least one participant who also demonstrated all required behaviors (*#RB*, PL 1), suggesting that these questions, too, are capable of eliciting cognitively complex thinking from students with SLDR, even if they didn't always in the study. These higher-than-desirable D_qs may be attributable in part to some participants' relative lack of think-aloud verbalization skill or experience. That the criterion-failing questions were also the study's least successfully answered Reading and Writing questions—with only two to six participants getting the right answers—further suggests that participants, by and large, simply struggled with the content in construct-relevant ways related to their English language arts/literacy achievement. Indeed, three of the five criterion-failing questions (questions 1, 3, and 14) were PL 7, the highest difficulty on the scale. (The other two—questions 6 and 7—were of medium difficulty, with PLs of 5 and 4, respectively.)

PARTICIPANT PERFORMANCE VIGNETTES

Vignettes from participant performance on the examined Reading and Writing questions provide further evidence that participants with SLDR were able to exhibit cognitively complex thinking in line with the questions types' constructs.

This section relies primarily on a case study approach, in which we follow a single participant, RW36, as he works through all fifteen Reading and Writing questions, succeeding on some and struggling with others. In the latter cases, RW36's vignettes are supplemented with those from

Example: Question Performance

7
2
0
4
2
0

9
7
2 ✔

Performance statistics for Reading and Writing question 4 are pictured above. The responses from seven participants attained PL 1 (the most successful outcome), those from two participants attained PL 2, those from four participants attained PL 4, and those from two participants attained PL 5 (the least successful outcome). All participants answered the question, so the value in the "*NR* / no response" cell is 0. Adding together the counts in PLs 1–3, we find that a total of nine participants answered the question correctly (*#AC*). Seven of these participants also demonstrated all required behaviors (*#RB*, which is the same as the number in the "PL 1" cell in the "Performance by Level, by Question" sub-table). Subtracting *#RB* (7) from *#AC* (9) yields a question differential (D_q) of 2. This D_q exceeds the threshold for a "good" differential, hence the checkmark and blue shading, as *#RB* divided by *#AC* equals approximately .778, or 77.8 percent, which is above the 70 percent cutoff.

participants who were more successful, with the exception of question 1, for which no participant attained PL 1. These supplements serve to show that even when the case study participant encountered difficulties with particular questions, other participants were typically able to answer correctly and demonstrate all required behaviors, suggesting that participants with SLDR were able to demonstrate cognitively complex thinking in accordance with the question types' constructs at least some of the time on these questions.

Case Study: Participant RW36

Participants were considered good candidates for the case study approach when they met the following criteria:

- They indicated that their SLDR condition had at least a moderate impact on their test-taking ability, suggesting that they've observed at least some negative impact from their condition on their ability to take tests.
- They'd received or they expected to receive accommodations as part of SAT Suite testing.
- They answered all fifteen Reading and Writing questions (e.g., didn't run out of time).
- They exhibited a good participant differential (D_p).

Participant RW36, a male eleventh grader from Georgia, met these conditions. He identified as Black or African American and not of Hispanic, Latino, or Spanish origin. He self-reported a high school GPA (HSGPA) of A, indicated that he'd received or he expected to receive extra time and extra breaks accommodations in SAT Suite testing, and characterized his SLDR symptoms as moderate, meaning (per the definition provided in the screener) that his SLDR condition had some impact on his test-taking ability but that this impact could be managed with appropriate testing accommodations. RW36 answered nine of the fifteen Reading and Writing questions correctly and demonstrated all required behaviors for eight of those nine, resulting in a participant differential of 1 (89 percent), which exceeded the criterion for a good D_p.

Reading and Writing Question 1

Skill/Knowledge Testing Point	Words in Context
Performance Score Band	7
Stimulus Subject Area	Science
Stimulus Text Complexity	PSR (postsecondary readiness, grades 12–14)
Required Behaviors	1. Read and demonstrate comprehension of the passage. 2. Select the answer choice that completes the passage with the most logical and precise word or phrase.
RW36 Performance Level	5

To demonstrate that the integrity of underground metal pipes can be assessed without unearthing the pipes, engineer Aroba Saleem and colleagues ________ the tendency of some metals' internal magnetic fields to alter under stress: the team showed that such alterations can be measured from a distance and can reveal concentrations of stress in the pipes.

Which choice completes the text with the most logical and precise word or phrase?

A) hypothesized

B) discounted

C) redefined

D) exploited

Question 1, a hard (PSB 7) Words in Context question set in a highly challenging (PSR) science context, requires test takers to determine the word or phrase that completes the text (i.e., fills in the blank) in the most logical and precise way. The best answer (*key*) is choice D, as the engineers "exploited," or made use of, "the tendency of some metals' internal magnetic fields to alter under stress" to assess the "integrity of underground metal pipes" without digging up those pipes.

> So even though I'm struggling with comprehending it, but I think the best choice might be "hypothesized" [choice A] since they're trying to test something. And normally, you do, you make a hypothesis. In the sentence, "Engineer Aroba Saleem and colleagues hypothesized the tendency of some metals . . ." So it fits and makes sense. Like, C ["redefined"] is also a possible answer, but ["hypothesized"; *participant says "hypothesis"*] just makes the most sense.
>
> *Participant RW36*

Participant RW36 answered the question incorrectly and didn't demonstrate the other required behavior, resulting in a PL of 5. RW36 expresses difficulty with reading the passage content ("I'm struggling with comprehending it") and seems to rely on the general sense that the passage involves an experiment ("they're trying to test something") when selecting choice A, "hypothesized." This answer choice is incorrect because the engineering team led by Aroba Saleem didn't merely hypothesize, or assume, that a property of some metals' internal magnetic fields—their tendency to alter under stress—would allow the team to assess the integrity of underground pipes without unearthing them; rather, the team used, or "exploited," this already known property and "showed" that "such alterations can be measured from a distance and can reveal concentrations of stress in the pipes."

Participant RW36's struggle with this question was a common one: Only four participants answered the question correctly, and none did so while also demonstrating appropriate passage comprehension. RW36's incorrect answer selection was, by his own acknowledgment, based

Vignette and Transcription Notes

Broadly speaking, the vignettes throughout this report are verbatim representations of participants' verbalizations, and the College Board researchers verified the transcripts' accuracy and completeness against the sessions' video recordings. However, for the sake of readability, some minor alterations were made in the vignettes' presentation. Repetitions (e.g., "It's like, it's like") were cleaned up, "ums" and similar verbal hesitations were removed, and [bracketed] text was sometimes added by the researchers to clarify participants' points or to complete or correct quotations from test passages.

on a lack of understanding of this highly challenging passage. Instead of a full understanding of the context, RW36 relied on a broad sense of the passage as describing something seemingly uncertain in science, which led him to choose “hypothesized.” This performance is suggestive of difficulties in close reading and comprehension of the specifics commensurate with the challenge of the passage text complexity (PSR) and difficulty of the question itself (PSB 7).

Reading and Writing Question 2

Skill/Knowledge Testing Point	Text Structure and Purpose (Passage main purpose subtype)
Performance Score Band	3
Stimulus Subject Area	Literature
Stimulus Text Complexity	MID (middle school/junior high, grades 6–8)
Required Behaviors	1. Read and demonstrate comprehension of the passage. 2. Select the answer choice that best states the main purpose of the passage.
RW36 Performance Level	1

The following text is adapted from Jean Webster’s 1912 novel *Daddy-Long-Legs*. The narrator is a young college student writing letters detailing her weekly experiences.

[The college is] organizing the Freshman basket-ball team and there’s just a chance that I shall make it. I’m little of course, but terribly quick and wiry and tough. While the others are hopping about in the air, I can dodge under their feet and grab the ball.

Which choice best states the main purpose of the text?

A) To compare basketball with other sports

B) To provide details of how to play basketball

C) To state how players will be chosen for the basketball team

D) To explain why the narrator thinks she might make the basketball team

Question 2, an easy (PSB 3) Text Structure and Purpose question set in a moderately challenging (MID) literature context, requires test takers to determine which answer choice best states the main purpose of the passage. The best answer is choice D, as the focus of the text is on the reasons the narrator thinks she’ll make the basketball team: she’s “terribly quick and wiry and tough” and, because of her small stature, can “dodge under [other players’] feet and grab the ball.”

> So it can’t be A because it’s not talking about any other sports but basketball. It can’t be B since it’s not really telling you how to play basketball, not giving instructions. It’s also being, it’s also in the first person, I think? He’s talking about, he hasn’t even talked about it in the first person. So it’s not really talking to the reader. It’s possible to be C, but even still it’s talking about, it’s in person, the person that is speaking is talking about what makes them special and why they’re good at basketball. So it has to be D.
>
> *Participant RW36*

Participant RW36 answered the question correctly and demonstrated both required behaviors, resulting in a PL of 1. RW36 reaches the keyed response, choice D (behavior 2), partially through a process of eliminating the other answer choices as incorrect, in so doing demonstrating appropriate passage comprehension (behavior 1). RW36 reasonably blocks choice A because the passage isn't "talking about any other sports but basketball." He correctly rules out choice B because the passage isn't "really telling you how to play basketball, not giving instructions." RW36 is somewhat tempted by choice C, incorrectly asserting that the passage isn't in "first person," but properly selects choice D after determining that the passage emphasizes the narrator "talking about what makes them special and why they're good at basketball."

Reading and Writing Question 3

Skill/Knowledge Testing Point	Text Structure and Purpose (Part-whole relationships subtype)
Performance Score Band	7
Stimulus Subject Area	History/social studies
Stimulus Text Complexity	PSR (postsecondary readiness, grades 12–14)
Required Behaviors	1. Read and demonstrate comprehension of the passage. 2. Select the answer choice that best describes the main function of the underlined portion in the passage as a whole.
RW36 Performance Level	5

More than 60% of journeys in Mexico City occur via public transit, but simply reproducing a feature of the city's transit system—e.g., its low fares—is unlikely to induce a significant increase in another city's transit ridership. As Erick Guerra et al. have shown, transportation mode choice in urban areas of Mexico is the product of a complex mix of factors, including population density, the spatial distribution of jobs, and demographic characteristics of individuals. System features do affect ridership, of course, but there is an irreducibly contextual dimension of transportation mode choice.

Which choice best describes the function of the underlined portion in the text as a whole?

A) It presents an objection to the argument of Guerra et al. about transportation mode choice in urban areas of Mexico.

B) It explains why it is challenging to influence transit ridership solely by altering characteristics of a transit system.

C) It illustrates the claim that a characteristic associated with high transit ridership in Mexico City is not associated with high transit ridership elsewhere.

D) It substantiates the assertion that population density, the spatial distribution of jobs, and demographic characteristics are important factors in transportation mode choice.

Question 3, a hard (PSB 7) Text Structure and Purpose question set in a highly challenging (PSR) history/social studies context, requires test takers to determine the main function of the underlined portion of the passage in terms of the passage as a whole. The best answer is choice B. The underlined portion—"there is an irreducibly contextual dimension of transportation mode choice"—restates the passage's claim that "simply reproducing" an aspect of Mexico City's transit system, such as its low fares, is "unlikely to induce a significant increase in another city's transit ridership," a claim supported in the passage by findings from Erick Guerra et al., who determined that "transportation mode choice in urban areas of Mexico is the product of a complex mix of factors."

> So I think it's D because it talks about all of those different aspects and also how it is integrated in public transportation, or transportation mode choice. And all the other don't really make that much sense to me. So I'm gonna pick D.
>
> *Participant RW36*

Participant RW36 answered the question incorrectly and didn't demonstrate the other required behavior, resulting in a PL of 5. RW36 seems to have been drawn to the fact that choice D repeats specific factors cited in the passage as influencing transit ridership in urban areas of Mexico: "population density, the spatial distribution of jobs, and demographic characteristics of individuals." While it's true, per the passage, that these factors have important roles to play in transportation mode choice, the main function of the underlined portion is to remind readers that, in the wider passage, these considerations only suggest the kinds of "contextual" factors that either support or inhibit transit ridership in various places. Choice D is further incorrect because the underlined portion in no way "substantiates," or offers additional support for, the assumption that the listed elements are "important factors in transportation mode choice"; rather, it merely reiterates the passage's central claim without providing new evidence.

Supplementary Vignette: Participant RW41

Participant RW41 answered question 3 correctly and also demonstrated both required behaviors, resulting in a PL of 1. RW41 was one of only two participants to have answered the question correctly and the only participant to have done so while also demonstrating both required behaviors.

> So I'm gonna go back to the text and try to find the main claim of the passage. "More than 60% of journeys in Mexico [City] occur via public transit, but simply reproducing a feature of the city's transit system—[e.g.,] its low fares—is unlikely to induce a significant increase in another city's [transit] ridership." So this is telling me that it would be unlikely to induce a significant increase. So I'm gonna go back and try to maybe eliminate or select the choice.
>
> A, "It presents an objection to the argument [of Guerra et al.] about transportation mode choice in urban areas of Mexico." It's a possible choice.
>
> B, "It explains why it is challenging to influence transit ridership solely by altering characteristics of [a] transit system." Possible.

> C, “It illustrates [the claim] that a characteristic associated with high transit ridership in Mexico City is not associated with high [transit] ridership elsewhere.” I don’t know about that one.
>
> D, “It substantiates the assertion that population density, the spatial distribution of jobs, and demographic characteristics are important factors in transportation mode choice.”
>
> So it’s between A or B for me. I would have to say between these two, [it’s] B because in the passage, it says that [Mexico City’s] low fares are unlikely to induce a significant increase in another city’s transit ridership. So this is going back to B. The answer choice is pretty much saying that it’s challenging to influence transit ridership solely by altering characteristics of the transit system. So I’m gonna go with B.
>
> *Participant RW41*

Participant RW41’s first task is to “find the main claim of the passage,” which he accomplishes by rereading the passage’s first sentence and slightly paraphrasing the claim found there, in the process demonstrating adequate passage comprehension (behavior 1): “So this is telling me that [merely reproducing a feature of Mexico City’s transit system] would be unlikely to induce a significant increase” in transit ridership in another city. He then “go[es] back” to try to “eliminate or select the choice.” During his initial survey of the answer options, he settles on choices A and B as the most likely candidates (“so it’s between A and B for me”). RW41 then selects choice B, the best answer (behavior 2). In doing so, he first restates the passage’s claim: “. . . in the passage, it says that [Mexico City’s] low fares are unlikely to induce a significant increase in another city’s transit ridership.” He then notes that choice B lines up closely with that claim: “The answer choice is pretty much saying that it’s challenging to influence transit ridership solely by altering characteristics of the transit system.”

Reading and Writing Question 4

Skill/Knowledge Testing Point	Command of Evidence: Quantitative (Table subtype)
Performance Score Band	4
Stimulus Subject Area	Science
Stimulus Text Complexity	SCO (upper secondary, grades 9–11)
Required Behaviors	1. Read and demonstrate comprehension of the passage. 2. Demonstrate an understanding of the table, including what the table as a whole as well as its various rows and columns represent. 3. Demonstrate an understanding of the relationship among the passage, the table, and the criterion set forth in the question’s stem. 4. Select the answer choice that best meets the criterion set forth in the question’s stem.
RW36 Performance Level	1

Partial List of Candidate Species for De-extinction

Common name	Scientific name	Became extinct
Huia	*Heteralocha acutirostris*	1907
Caribbean monk seal	*Monachus tropicalis*	1952
Passenger pigeon	*Ectopistes migratorius*	1914
Saber-toothed cat	*Smilodon*	11,000 years before present
Woolly mammoth	*Mammuthus primigenius*	6,400 years before present

The passage of time is among the many obstacles faced by scientists who are pursuing de-extinction efforts—that is, efforts to use breeding or a mixture of cloning and genetic engineering to bring back extinct species. Specifically, researchers are concerned that the longer a species has been extinct, the less likely it is that a suitable habitat still exists for that species. Among candidate species for de-extinction, this problem would be especially concerning for the ________

Which choice most effectively uses data from the table to complete the statement?

A) passenger pigeon (*Ectopistes migratorius*), which became extinct only a few years after the huia (*Heteralocha acutirostris*).

B) saber-toothed cat (*Smilodon*), which became extinct 11,000 years ago.

C) woolly mammoth (*Mammuthus primigenius*), which became extinct several thousand years before the saber-toothed cat (*Smilodon*).

D) Caribbean monk seal (*Monachus tropicalis*), which became extinct in 1952.

Question 4, a medium-difficulty (PSB 4) Command of Evidence: Quantitative question set in a challenging (SCO) science context, requires test takers to draw on both passage and table to complete the statement containing the blank with the most effective data from the table. The passage establishes that "the longer a species has been extinct, the less likely it is that a suitable habitat still exists for that species," thus making longer-extinct species progressively worse candidates for de-extinction efforts. Per the table, the saber-toothed cat (*Smilodon*) went extinct "11,000 years before present," making it the longest-extinct candidate in the table and making choice B the best answer.

> "Among candidate species for de-extinction, this problem would be especially concerning for the [blank]." So just by looking at the passage, I can already get an idea what the question is. "Which choice most effectively uses data from the table to complete the statement?" So just kind of connecting both of them, it's asking which species would be harder to bring back, or de-extinct. Just already looking at it, I would say it's the saber-tooth mainly because it went extinct 11,000 years ago; compared to the other ones, there's a huge difference.

> So, looking at the answer choices, with A being "Passenger pigeon [(*Ectopistes migratorius*)], which became extinct only a few years ago." Can automatically eliminate that. B, "Saber-toothed cat [(*Smilodon*)], which became extinct 11,000 years ago," which is most likely the right choice. C, "Woolly mammoth [(*Mammuthus primigenius*)], which became extinct several thousand years [before the saber-toothed cat (*Smilodon*); *participant says* "*ago*"]," which is a choice. And "[D, Caribbean; *participant says "Canadian"*] monk seal, which became extinct in 1952." So I can automatically get rid of A and D. So it has to be between B and C. And looking back at the chart, it looks like saber-tooth[ed] tiger [*sic*] has, I guess, a lot more years on the woolly mammoth, so I'm going to pick B.
>
> *Participant RW36*

Participant RW36 answered the question correctly and demonstrated all required behaviors, resulting in a PL of 1. Early in his response, RW36 indicates a clear understanding of the passage (behavior 1) and the intended criterion linking passage, table, and question stem (behavior 3): "So just kind of connecting both of them, it's asking which species would be harder to bring back, or de-extinct." He then demonstrates table comprehension (behavior 2) by zeroing in on the intended answer: "Just already looking at [the table], I would say it's the saber-tooth mainly because it went extinct 11,000 years ago; compared to the other ones, there's a huge difference." Reading through the answer choices, RW36 "automatically eliminate[s]" choice A, the passenger pigeon, for having gone extinct too recently and rules out choice D, the Caribbean monk seal, for the same reason. He ultimately and correctly selects choice B, the saber-toothed cat (behavior 4), on the grounds that the table indicates the cat went extinct 11,000 years ago, whereas the woolly mammoth died out more recently. RW36 doesn't attend to the fact that choice C itself is factually incorrect, as it wrongly asserts that the woolly mammoth went extinct before the saber-toothed cat, but this doesn't impede his ability to draw on a clear understanding of the passage and table to reach the correct conclusion.

Reading and Writing Question 5

Skill/Knowledge Testing Point	Command of Evidence: Textual
Performance Score Band	4
Stimulus Subject Area	Literature
Stimulus Text Complexity	SCO (upper secondary, grades 9–11)
Required Behaviors	1. Read and demonstrate comprehension of the passage. 2. Demonstrate an understanding of the relationship between the criterion set forth in the question's stem and the passage. 3. Select the answer choice that best meets the criterion set forth in the question's stem.
RW36 Performance Level	1

"The Yellow Wallpaper" is an 1892 short story by Charlotte Perkins Gilman. In the story, the narrator expresses mixed feelings about her surroundings: ________

Which quotation from "The Yellow Wallpaper" most effectively illustrates the claim?

A) "This wallpaper has a kind of sub-pattern in a different shade, a particularly irritating one, for you can only see it in certain lights, and not clearly then."

B) "By moonlight—the moon shines in all night when there is a moon—I wouldn't know it was the same paper."

C) "I'm really getting quite fond of the big room, all but that horrid [wall]paper."

D) "The color is repellant, almost revolting; a smouldering, unclean yellow, strangely faded by the slow-turning sunlight."

Question 5, a medium-difficulty (PSB 4) Command of Evidence: Textual question set in a challenging (SCO) literature context, requires test takers to determine which of the provided quotations from the short story "The Yellow Wallpaper" most clearly expresses the narrator's mixed feelings about her surroundings. The best answer is choice C, as it illustrates both the narrator's general appreciation for the room ("I'm really getting quite fond of the big room") and specific dislike of its "horrid" wallpaper.

> "'The Yellow Wallpaper' is an 1892 short story by Charlotte Perkins Gilman. In the story, the narrator expresses mixed feelings about her surroundings." So just from that, I already know I'm looking for mixed feelings about her surroundings. . . . [*Reads answer choices*] So I would say a lot of them already have, like, one senior feeling; nothing is really mixed. The only one that really stands out to me is C since it says "I'm [really] getting quite fond of the big room"—which is, like, one emotion—"all but [that] horrid [wall]paper." It's talking about her surroundings and mixed feelings.
>
> *Participant RW36*

Participant RW36 answered the question correctly and demonstrated all required behaviors, resulting in a PL of 1. RW36 exhibits an understanding of the connection between the passage and criterion (behavior 2), noting that the appropriate choice would exemplify "mixed feelings" and contrasting that with other choices having "one senior feeling" only. He also shows proper passage (here, answer choice) comprehension (behavior 1) while selecting choice C, the keyed response (behavior 3), observing that choice C "really stands out" as "mixed."

Reading and Writing Question 6

Skill/Knowledge Testing Point	Transitions
Performance Score Band	5
Stimulus Subject Area	History/social studies
Stimulus Text Complexity	SCO (upper secondary, grades 9–11)
Required Behaviors	1. Read and demonstrate comprehension of the passage. 2. Select the answer choice that completes the passage with the most logical transition.
RW36 Performance Level	4

According to Duverger’s law, countries with single-ballot majoritarian elections for single-member districts tend to polarize into two-party systems, wherein dueling political parties consistently dominate the political system. ________ countries with proportional-representation electoral systems tend to support multi-partyism, under which power gets distributed among many political parties.

Which choice completes the text with the most logical transition?

A) Subsequently,

B) Conversely,

C) For instance,

D) In other words,

Question 6, a medium-difficulty (PSB 5) Transitions question set in a challenging (SCO) history/social studies context, requires test takers to determine the most logical transition word or phrase to complete the sentence in the passage with the blank. The best answer is choice B, “conversely,” as the passage’s last sentence (the one containing the blank) contrasts proportional-representation electoral systems and multi-partyism with the single-ballot majoritarian elections for single-member districts and two-party systems mentioned in the passage’s first sentence.

> “Which choice completes the text with the most logical transition?” So just using my prior knowledge with, majoritarian elections in single-member districts in two-party systems are, I think, about the U.S., with Democrats and Republicans. It was, the multipartisan, multiparticism [*sic*] part is Europe, because I know they have multiple different parties. [*Reads answer choices*] So I don’t know about A. B is a possibility. C, I don’t think makes sense. But D, “In other words,” it flows and makes sense and is talking about another side, or the opposite. And I think “subsequently” can be used as, like, a, as, like, a synonym for “in other words.” But I think I’m gonna go with, stick with D.
>
> *Participant RW36*

Participant RW36 answered the question incorrectly but demonstrated a single required behavior, resulting in a PL of 4. RW36 begins by making a reasonable connection between the passage and his prior knowledge of political systems:

". . . majoritarian elections in single-member districts in two-party systems are, I think, about the U.S., with Democrats and Republicans," whereas "the [multi-party] part is Europe, because I know they have multiple different parties." This connection demonstrates adequate passage comprehension (behavior 1) because it indicates that RW36 has recognized the central contrast presented by the passage's discussion of Duverger's law. However, he elects to opt for choice D, "in other words," on the grounds that "it flows and makes sense and is talking about another side, or the opposite." This choice both simultaneously shows an understanding of the basic contrast presented in the passage and a misunderstanding of the meaning of "in other words," which signals a paraphrase or restatement rather than a contrast. Such misunderstanding is further exemplified when RW36 erroneously asserts that "subsequently" (choice A) and "in other words" (choice D) are essentially synonymous. Nonetheless, RW36 shows partial enactment of the question type's construct.

Supplementary Vignette: Participant RW11

Participant RW11 answered question 6 correctly and demonstrated both required behaviors, resulting in a PL of 1. RW11 was one of only two participants to have answered the question correctly and the only participant to have done so while also demonstrating both required behaviors.

> So, well, obviously, you need to complete the text with, well, a suitable transition word. OK. ". . . wherein dueling political parties consistently dominate the political . . ." Let's see if maybe they have, like, a dictionary or a thesaurus. Find out. That kind of helps me sometimes. [*Checks Bluebook's options; this sort of tool isn't present.*] I need to review what I was doing all over again. OK, so, fill in the blank. ". . . wherein dueling political parties consistently dominate the political system. Blank, countries with proportional-representation electoral systems tend to support multi-partyism, under which power gets distributed among many political parties."
>
> Huh. ". . . dueling . . . distributed among many political parties." OK. "Subsequently" means "following this." [*Gestures with pointer to the blank*] And I'm going to, I'll read this. "Subsequently, countries with proportional-representation electoral systems tend to support multi-partyism . . ." Oh, no. No, no, I don't think it's option A. B, "conversely." [*Long pause*] Huh. Could be. This could be a good—I'll mark this as, like, "maybe." C, "for instance." Huh. I highly doubt it's C. It just doesn't feel right because it doesn't make sense there. "In other words." Huh. "In other words, countries with proportional-representation electoral systems tend to support multi-partyism, under which power gets distributed among many political parties." "Many political parties," not dueling two-party systems. I don't know. It seems like not a good transition word. "Conversely" feels like it could—I think it's "conversely." I feel like they're talking about something else now. I think it could be this.
>
> *Participant RW11*

Participant RW11's response begins by exhibiting awareness of the question's task, which is to "complete the text with, well, a suitable transition word," which shows conceptual understanding. RW11 briefly looks around the Bluebook testing interface for a dictionary or thesaurus to help answer the question; not finding either, she finishes reading through the passage. She reasonably defines “subsequently” as “following this” and then figuratively plugs the word into the blank to get a sense of how the option “sounds” in context—a common solving strategy employed by students. She seems to rely considerably on the general sense of a given answer option contextualized by the passage, deciding that “for instance” (choice C) “just doesn’t feel right” and that “in other words” (choice D) “seems like not a good transition word.” At the same time, RW11 demonstrates adequate passage comprehension (behavior 1) when selecting choice B, the best answer (behavior 2), on the grounds that the passage is “talking about something else now” beginning with the sentence containing the blank, an adequate working definition of “conversely” in context.

Reading and Writing Question 7

Skill/Knowledge Testing Point	Rhetorical Synthesis
Performance Score Band	4
Stimulus Subject Area	Humanities
Stimulus Text Complexity	MID (middle school/junior high, grades 6–8)
Required Behaviors	1. Read and demonstrate comprehension of the student-produced notes. 2. Demonstrate an understanding of the relationship between the notes and the criterion set forth in the question’s stem. 3. Select the answer choice that best meets the criterion set forth in the question’s stem.
RW36 Performance Level	1

While researching a topic, a student has taken the following notes:

- In 1859, the novel *Adam Bede* was published in England.
- According to the novel’s title page, the author’s name was George Eliot.
- George Eliot was widely assumed to be a pseudonym.
- A pseudonym is a fake name used to conceal an author’s identity.
- A woman named Mary Ann Evans later revealed herself as the novel’s real author.

The student wants to identify the real author of *Adam Bede*. Which choice most effectively uses relevant information from the notes to accomplish this goal?

A) The real author of *Adam Bede* was Mary Ann Evans, who published the novel using the pseudonym George Eliot.

B) George Eliot, which *Adam Bede*’s title page indicated was the name of the novel’s author, was widely assumed to be a pseudonym.

C) The title page of the novel *Adam Bede* indicated that the author’s name was George Eliot.

D) A woman who had used a pseudonym to conceal her identity later revealed herself as the real author of *Adam Bede*.

Question 7, a medium-difficulty (PSB 4) Rhetorical Synthesis question set in a moderately challenging (MID) humanities context, requires test takers to select the answer choice that best uses relevant information from the student-produced "notes" (bulleted list of informational points, ostensibly gathered from research) to meet the question's criterion, which, in this case, is to identify the real author of *Adam Bede*. The best answer is choice A, as it clearly indicates that *Adam Bede*'s author was Mary Ann Evans, who used the pseudonym George Eliot when publishing.

> So it can't be D since . . . [*Long pause*] I actually think it might be A because it's talking about the book *Adam Bede*, whose author was Mary Ann, who published a novel using the name George Eliot. And that kind of, then it adds up because according to the novel's title [*sic*], the author's name was George Eliot, and the novel was *Adam Bede*. And then at the end, it says a woman named Mary Ann later revealed herself as the novel's real author. So I'm gonna pick A.
>
> *Participant RW36*

Participant RW36 answered the question correctly and demonstrated all required behaviors, resulting in a PL of 1. Prior to the response vignette quoted above, RW36 struggles considerably to pronounce the word "pseudonym" when encountering it in the notes. The notes, however, gloss "pseudonym" as "a fake name used to conceal an author's identity," and, from this, RW36 realizes that it "was a word I know, just [one that I] don't know how to say." Overall, however, his response indicates adequate knowledge of both the notes (behavior 1) and the intended relationship between the notes and the criterion set forth in the question's stem (behavior 2). He observes that the question is about the book *Adam Bede*, recognizes that the author's real name was "Mary Ann," and notes that the novel was published under the name George Eliot. Based on this understanding, RW36 correctly picks choice A as the answer (behavior 3).

Reading and Writing Question 8

Skill/Knowledge Testing Point	Rhetorical Synthesis
Performance Score Band	5
Stimulus Subject Area	Science
Stimulus Text Complexity	PSR (postsecondary readiness, grades 12–14)
Required Behaviors	1. Read and demonstrate comprehension of the student-produced notes. 2. Demonstrate an understanding of the relationship between the notes and the criterion set forth in the question's stem. 3. Select the answer choice that best meets the criterion set forth in the question's stem.
RW36 Performance Level	1

While researching a topic, a student has taken the following notes:

- Scientists have developed a "freeze-thaw" battery that can retain 92% of its charge after twelve weeks.
- The battery contains molten salt (a type of salt that liquifies when heated and solidifies at room temperature).
- When the salt is in a liquid state, energy flows through the battery.
- When the salt is in a solid state, energy stops flowing and is stored in the battery.
- The stored (frozen) energy can be used by reheating (thawing) the battery.

The student wants to specify how the salt enables energy storage. Which choice most effectively uses relevant information from the notes to accomplish this goal?

A) Scientists have developed a freeze-thaw battery that contains molten salt, which liquifies when heated and solidifies at room temperature.

B) The stored energy in a freeze-thaw battery, which contains molten salt, can be used by reheating the battery.

C) When the molten salt in a freeze-thaw battery solidifies at room temperature, energy stops flowing and can be stored in the battery.

D) Molten salt allows a freeze-thaw battery to retain 92% of its charge after twelve weeks.

Question 8, a medium-difficulty (PSB 5) Rhetorical Synthesis question set in a highly challenging (PSR) science context, requires test takers to, again, select the answer choice that best uses relevant information from the notes to accomplish the writer's goal, which, in this case, is to specify how the salt in the freeze-thaw battery described in the notes enables energy storage. The best answer is choice C, as this option addresses how solidifying the battery's molten salt, which occurs at room temperature, stops energy flow and thereby permits energy storage.

> [*Reads notes and question stem*] So using just the information that gives me, I would say he would want to, for relevant information, to say we'll talk about the liquid and solid states and how it stops flowing energy or lets energy through. OK. [*Reads answer choices*] So A, "Scientists have developed a freeze-thaw battery that contains molten salt that [*sic*] liquefies when heated and solidifies at room temperature." Just off the bat, no, can't be A. B, "The stored energy in a frozen-thaw battery [*sic*], which contains molten salt, can be used to reheat [*sic*] the battery." Just going from my personal—well, not personal experience, but my past thoughts: When you heat a battery, it explodes, so that can't be [the] one. C, "When the molten salt in the frozen-thaw battery solidifies at room temperature, the [*sic*] energy stops flowing and can be stored in the battery." D, "Molten salt allows a frozen-thaw battery to retain 92% of its

> charge after 12 weeks." So knowing all of this, I think C because it talks about the battery, it talks about the molten salt, it talks about how at room temperature, when it solidifies, the energy stops flowing and is stored in the battery.
>
> *Participant RW36*

Participant RW36 answered the question correctly and demonstrated all required behaviors, resulting in a PL of 1. From the notes and question stem, RW36 correctly concludes that the best answer to the question will focus on "the liquid and solid states [of the freeze-thaw battery] and how it stops flowing energy or lets energy through" (behavior 1). After considering the answer options, he selects choice C, the best answer (behavior 3), on the grounds that "it talks about how at room temperature, when [the salt] solidifies, the energy stops flowing and is stored in the battery." Since the question's criterion asks for the choice that best shows "how the salt enables energy storage," RW36 also demonstrates behavior 2, which concerns the relationship among the notes, question stem, and answer choices. It should be noted that while RW36 does answer correctly and demonstrate all required behaviors, he exhibits some prior knowledge interference, as he concludes that choice B can't be right because "when you heat a battery, it explodes"—an overly broad conclusion not supported by the passage. This serves as a reminder that even successful question performance isn't necessarily perfect performance.

Reading and Writing Question 9

Skill/Knowledge Testing Point	Words in Context
Performance Score Band	4
Stimulus Subject Area	Science
Stimulus Text Complexity	PSR (postsecondary readiness, grades 12–14)
Required Behaviors	1. Read and demonstrate comprehension of the passage. 2. Select the answer choice that completes the passage with the most logical and precise word or phrase.
RW36 Performance Level	1

According to a team of neuroeconomists from the University of Zurich, ease of decision making may be linked to communication between two brain regions, the prefrontal cortex and the parietal cortex. Individuals tend to be more decisive if the information flow between the regions is intensified, whereas they make choices more slowly when information flow is ________.

Which choice completes the text with the most logical and precise word or phrase?

A) reduced

B) evaluated

C) determined

D) acquired

Question 9, a medium-difficulty (PSB 4) Words in Context question set in a highly challenging (PSR) science context, requires test takers to select the most logical and precise word or phrase to fill in the blank in the passage. The best answer is choice A. "Reduced" most effectively completes the blank, as what's called for here is a word or phrase that logically concludes the passage's contrast between increased decisiveness when information flow between the prefrontal cortex and parietal cortex is intensified and decreased decisiveness when information flow between these two brain regions is lowered.

> [*Reads the passage and question stem*] Just from that, I can already know some has, it has to do with something slowing down. So just read the choices. Question: "Which choice completes the text with the most logical and precise word or phrase?" So A, "reduced." I already think that makes sense. But B, "evaluated," C, "determined," D is "acquired." I think it has to be A since it's making a comparison between slow and fast, or how getting information faster or— [*Rereads sentence with the blank*] "Intensify" means fast or strong; the one that really relates to it in opposite sense is "reduced."
>
> *Participant RW36*

Participant RW36 answered the question correctly and demonstrated both required behaviors, resulting in a PL of 1. RW36 exhibits adequate passage comprehension (behavior 1) when noting that the passage "has to do with something slowing down" and is "making a comparison between slow and fast." Based on this correct interpretation, he then picks choice A, the best answer, as it's the option "that really relates to ["intensify" in the passage] in opposite sense" (behavior 2).

Reading and Writing Question 10

Skill/Knowledge Testing Point	Cross-Text Connections
Performance Score Band	4
Stimulus Subject Area	Humanities
Stimulus Text Complexity	SCO (upper postsecondary, grades 9–11)
Required Behaviors	1. Read and demonstrate comprehension of Text 1, including its point of view on the topic. 2. Read and demonstrate comprehension of Text 2, including its point of view on the topic. 3. Demonstrate an understanding of the fundamental relationship between the two passages in terms of topic, content, and/or point of view. 4. Select the answer choice that best meets the criterion set forth in the question's stem.
RW36 Performance Level	2

Text 1
Graphic novels are increasingly popular in bookstores and libraries, but they shouldn't be classified as literature. By definition, literature tells a story or conveys meaning through language only; graphic novels tell stories through illustrations and use language only sparingly, in captions and dialogue. Graphic novels are experienced as series of images and not as language, making them more similar to film than to literature.

Text 2
Graphic novels present their stories through both language and images. Without captions and dialogue, readers would be unable to understand what is depicted in the illustrations: the story results from the interaction of text and image. Moreover, Alison Bechdel's *Fun Home* and many other graphic novels feature text that is as beautifully written as the prose found in many standard novels. Therefore, graphic novels qualify as literary texts.

Based on the texts, how would the author of Text 2 most likely respond to the overall argument presented in Text 1?

A) By asserting that language plays a more important role in graphic novels than the author of Text 1 recognizes

B) By acknowledging that the author of Text 1 has identified a flaw that is common to all graphic novels

C) By suggesting that the story lines of certain graphic novels are more difficult to understand than the author of Text 1 claims

D) By agreeing with the author of Text 1 that most graphic novels aren't as well crafted as most literary works are

Question 10, a medium-difficulty (PSB 4) Cross-Text Connections question set in a challenging (SCO) humanities context, requires test takers to draw the most reasonable conclusion connecting the content of the two topically related passages presented. This involves comprehension of each passage separately as well as making the appropriate synthetic connection "bridging" the two passages. In this case, test takers are asked to determine how the author of Text 2 would most likely respond to the argument presented in Text 1. Text 1's premise is that graphic novels don't qualify as and therefore "shouldn't be classified" as literature because the words in a graphic novel are subordinate to the visuals in meaning making; Text 2, on the other hand, argues that "graphic novels qualify as literary texts" because the words are just as important as the visuals to comprehension and because the language used in some graphic novels is as beautiful as that in some traditional prose works. Given this, the author of Text 2 would most likely respond to the author of Text 1 as choice A, the best answer, does, by "asserting that language plays a more important role in graphic novels than the author of Text 1 recognizes." Note that, commensurate with its relative level of challenge, the question doesn't simply ask for a statement of each author's point of view but rather calls on test takers to focus on a specific part of the comparison the two authors implicitly draw between each other's views.

> So just thinking, I always think that Text 2's author would probably be fighting for their point of view. So D is really trying to say that graphic novels are less complex and aren't as well made as most literature works. So it can't be D. [*Long pause*] I don't think it's C because it kind of falls under the same kind of situation, but it's possible. [*Long pause*] I think A is a possible choice because when you're reading graphic novels and looking at pictures, the text needs to describe what is happening more detailed in less words. So I think it's A.
>
> *Participant RW36*

Participant RW36 answered the question correctly but demonstrated only two required behaviors, resulting in a PL of 2. RW36 exhibits at least a general understanding of Text 2 (behavior 2) by ruling out choice D on the grounds that the author of that text wouldn't agree that "graphic novels are less complex and aren't as well made as most literature works." At the same time, RW36 doesn't fully explain in his verbalization the main point of Text 1, nor does he precisely specify the relationship between the two texts. Indeed, RW36 presents the intended link between the two texts as a simple matter of pro and con ("I always think that Text 2's author would probably be fighting for their point of view"). While the two texts in this question are, in fact, broadly of the pro-con nature, choice A focuses on a specific element of the contrast—graphic novels' use of language—that RW36 largely sidesteps or actually misstates to some degree, as Text 1 doesn't really argue that graphic novels "describe what is happening more detailed in less words." As RW36 ends up choosing the best answer to the question (behavior 4), his response and its coding may simply suggest a limitation of the think-aloud methodology: that not all things thought are stated, or stated clearly. Alternatively, RW36 may have used his general (and in this case correct) sense of the passages' adversarial relationship to find the best answer without a full understanding of Text 1 and/or the precise nature of the contrast identified in choice A. In any event, RW36 shows partial enactment of the question type's construct.

Supplemental Vignette: Participant RW30

Participant RW30 answered question 10 correctly and demonstrated all required behaviors, resulting in a PL of 1. RW30 was one of ten participants to have answered the question correctly and one of seven participants to have done so while also demonstrating all required behaviors.

> OK. Looking at the prompt, I already know I have two different texts, meaning that I'm probably using or combining them in a way that they'll be working together. So I'll read Text 1. [*Reads Text 1*] So in Text 1, it's basically making the argument that graphic novels are not actual literature.
>
> So let's read Text 2. [*Reads Text 2*] Hearing this, I know the prompt, the first prompt [Text 1], is talking about how graphic novels should be not considered like literature due to the fact they use many images to explain the story, while Text 2 is using it as a, more of a literature piece because there is [*sic*] words and images, and the words help elevate those images to different extents.

> So reading, let's read the question: "Based on [the] texts, how would the author of Text 2 most likely respond to the overall argument [presented in; *participant says "of"*] Text 1?"
>
> So we're using Text 2's evidence of—basically, their whole point is that the text evaluates the images in a way that the images can't do by themselves. So we're comparing Text 2's author to Text 1['s].
>
> So answer A, "By asserting that language plays a more important role in graphic novels than the author of Text 1 recognizes." Overall, I feel like that's a great answer for that, for this prompt, but let's keep reading.
>
> Answer B, "[By acknowledging; *participant says "It's acknowledged"*] that the author of Text 1 [has] identified a flaw that is in [*sic*], that is common to all graphic novels." I feel like that wouldn't be a really good way to argue his piece, for the author of Text 2, so I would x out answer B.
>
> So I would read answer C, "By suggesting that the story lines of certain graphic novels are more difficult to understand than the author of Text 1 claims." In Text 2, it doesn't claim that anywhere. It says that the text and the images of the graphic novel make it what it is and make it, help you understand it, so it won't be answer C.
>
> Answer D, "By agreeing with the author of Text 1 that most graphic novels aren't as well crafted as most [literary; *participant says "literacy"*] works are." So, clearly, that doesn't help the guiding question, argument, likely responds to the overall argument presented in Text 1.
>
> So I would go with answer A, "By asserting that language plays a more important role in graphic novels than the author of Text 1 recognizes."
>
> *Participant RW30*

Participant RW30's response to question 10 is, in many respects, exemplary, and it's worth looking closely at his process. At the outset, RW30 exhibits a clear conceptual understanding of the task posed by the question: "Looking at the prompt, I already know I have two different texts, meaning that I'm probably using or combining them in a way that they'll be working together." After reading Text 1, RW30 summarizes it (behavior 1): Text 1 is "basically making the argument that graphic novels are not actual literature." After reading Text 2, he both captures the gist of that passage (behavior 2) and expresses the fundamental relationship conveyed between the texts (behavior 3): "Hearing this, I know the prompt, the first prompt [Text 1], is talking about how graphic novels should be not considered like literature due to the fact they use many images to explain the story, while Text 2 is using it as a, more of a literature piece because there is [*sic*] words and images, and the words help elevate those images to different extents." RW30 then evaluates the answer choices in turn, quickly settling on and ultimately selecting the best answer, choice A (behavior 4).

Reading and Writing Question 11

Skill/Knowledge Testing Point	Central Ideas and Details
Performance Score Band	3
Stimulus Subject Area	Literature
Stimulus Text Complexity	SCO (upper secondary, grades 9–11)
Required Behaviors	1. Read and demonstrate comprehension of the passage. 2. Select the answer choice that best states the main idea of the passage or accurately states a detail from the passage.
RW36 Performance Level	4

The following text is adapted from Ann Petry's 1946 novel *The Street*. Lutie lives in an apartment in Harlem, New York.

The glow from the sunset was making the street radiant. The street is nice in this light, [Lutie] thought. It was swarming with children who were playing ball and darting back and forth across the sidewalk in complicated games of tag. Girls were skipping double dutch rope, going tirelessly through the exact center of a pair of ropes, jumping first on one foot and then the other.

©1946 by Ann Petry

Which choice best describes what is happening in the text?

A) Lutie is observing the appearance of the street at a particular time of day and the events occurring on it.

B) Lutie is annoyed by the noise of children playing games on her street.

C) Lutie is puzzled by the rules of certain children's games.

D) Lutie is spending time alone in her apartment because she doesn't want to interact with her neighbors.

Question 11, an easy (PSB 3) Central Ideas and Details question set in a challenging (SCO) literature context, requires test takers to generalize about the content presented in the passage. Choice A is the best answer. The background information presented in the question informs readers that Lutie, the passage's narrator, lives in a Harlem apartment. The passage itself suggests that Lutie is observing activities on the street from her apartment window at a particular time of day: the "sunset was making the street radiant," and the street was "swarming with children" playing various games, such as rope jumping.

> So it can't be D since it doesn't seem like she's going—unless she's inside—it says "going tirelessly through the exact center of [a pair; *participant says "the par"*] of ropes, jumping [first; *participant says "the front"*] on one foot and then the other." Wait, hold on. [*Long pause*] Actually, it's possible it could be D since it's talking about, she's talking about how other girls and other children are playing with, playing games and playing with jump ropes. There aren't really any rules that are being

> mentioned, so it can't be C. It doesn't seem like she's annoyed, and she's not really observing the appearance of the street. So, actually, I think it's D.
>
> *Participant RW36*

Participant RW36 answered the question incorrectly but did demonstrate one required behavior, resulting in a PL of 4. RW36 exhibits some initial confusion about where the narrator, Lutie, is positioned ("unless she's inside"), but he also demonstrates passage comprehension (behavior 1) by blocking choice C on the grounds that no rules for the children's games are being explained and choice B on the grounds that Lutie doesn't seem annoyed. RW36, however, incorrectly surmises that Lutie isn't "really observing the appearance of the street," perhaps assuming that "appearance" here would need to be addressed by more than a passing reference to the sunset "making the street radiant." He also either reaches choice D's unsubstantiated conclusion that Lutie is trying to avoid her neighbors or simply selects choice D by elimination of choices A, B, and C. Nevertheless, RW36 shows partial enactment of the question type's construct.

Supplementary Vignette: Participant RW39

Participant RW39 answered question 11 correctly and demonstrated both required behaviors, resulting in a PL of 1. RW39 was one of fourteen participants who answered the question correctly and one of twelve participants to have done so while also demonstrating both required behaviors. (Participant RW36 was, in fact, the only participant to have incorrectly answered the question.)

> So I think the text is talking about a street that is just, like, that's really nice, has a good community, and it's just a good place to be at. And yeah.
>
> So A, "[Lutie] is observing the appearance of the street at a particular time of the [*sic*] day and [the] events occurring [on; *participant says "at"*] it." It could be this one because it, but it doesn't really talk about how she's observing it at a particular time of day. I'm going to say it's probably, like, probably around, like, sunset because it's talking about how, like, [there's] a glow from the sunset. So it does talk about a certain time of the day. And, yeah, that's A.
>
> And then B, "[Lutie] is annoyed by the noises [*sic*] of children playing games on [her; *participant says "the"*] street." I don't think this text really has, like, a negative feel to it, so, and like that, like, B is kind of negative. So I don't think it has any negative [connotations; *participant may have said "ambitions"*] to it, so it's probably not B.
>
> And C, "[Lutie] is puzzled by the rules of certain children's games." Again, this one, like, kind of has, like, negative [connotation; *participant may have said "annotation"*] to it, and I don't think the story has, like, a negative tone to it. And that's, like, more negative, so I don't think it's C.
>
> And D, "[Lutie] is spending time alone in her apartment because she doesn't want to interact with [her; *participant says "the"*] neighbors." No, because, obviously, like, she's, like, observing what's happening right now, and she's seeing what it is, so she's not in her apartment.
>
> So, probably A.
>
> *Participant RW39*

Participant RW39 demonstrates adequate passage comprehension (behavior 1) at several points. She reasonably describes the setting as “a street that is just, like, that’s really nice, has a good community, and it’s just a good place to be at.” She also correctly rules out choices B and C on the grounds that the negative suggestions of these options (“annoyed,” “puzzled”) doesn’t fit with the passage’s positive tone. In the process of selecting the best answer, choice A (behavior 2), RW39 does briefly block choice A because she initially doesn’t recognize how the passage is talking about “how she’s observing [the street] at a particular time of day” but then figures out that the time of day being referred to is sunset (“so it does talk about a certain time of the day”).

Reading and Writing Question 12

Skill/Knowledge Testing Point	Central Ideas and Details
Performance Score Band	6
Stimulus Subject Area	Humanities
Stimulus Text Complexity	PSR (postsecondary readiness, grades 12–14)
Required Behaviors	1. Read and demonstrate comprehension of the passage. 2. Select the answer choice that best states the main idea of the passage or accurately states a detail from the passage.
RW36 Performance Level	5

Many literary theorists distinguish between *fabula*, a narrative’s content, and *syuzhet*, a narrative’s arrangement and presentation of events. In the film *The Godfather Part II*, the *fabula* is the story of the Corleone family, and the *syuzhet* is the presentation of the story as it alternates between two timelines in 1901 and 1958. But literary theorist Mikhail Bakhtin maintained that *fabula* and *syuzhet* are insufficient to completely describe a narrative—he held that systematic categorizations of artistic phenomena discount the subtle way in which meaning is created by interactions between the artist, the work, and the audience.

Which choice best states the main idea of the text?

A) Literary theorist Mikhail Bakhtin argued that there are important characteristics of narratives that are not fully encompassed by two concepts that other theorists have used to analyze narratives.

B) Literary theorist Mikhail Bakhtin claimed that meaning is not inherent in a narrative but is created when an audience encounters a narrative so that narratives are interpreted differently by different people.

C) The storytelling methods used in *The Godfather Part II* may seem unusually complicated, but they can be easily understood when two concepts from literary theory are utilized.

D) Narratives that are told out of chronological order are more difficult for audiences to understand than are narratives presented chronologically.

Question 12, a hard (PSB 6) Central Ideas and Details question set in a highly challenging (PSR) humanities context, requires test takers to determine the passage's main idea. The best answer is choice A, as the main focus of the question is Mikhail Bakhtin's view that *fabula* and *syuzhet* are "insufficient to completely describe a narrative." The passage defines the concepts of *fabula* and *syuzhet* and illustrates them with the example of *The Godfather Part II* but then questions these concepts' adequacy by citing Bakhtin's belief that "meaning [in art] is created by interactions between the artist, the work, and the audience."

> So I don't think it's D. I'm so iffy about B because I was just confused on the question because it said "narrative," like, three times in a row. [*Long pause*] I think it's possible for it to be C. I think it's most likely C because it doesn't really talk about its order but more about its complexity.
>
> *Participant RW36*

Participant RW36 answered the question incorrectly and didn't demonstrate either required behavior, resulting in a PL of 5. In his verbalization, RW36 offers no clear indication of having understood this highly challenging passage and further suggests that he doesn't understand choice B ("I was just confused on the question"). Moreover, choice C, the answer option he selects, both mistakes the passage's example of *The Godfather Part II* for its main point and misinterprets the passage's claim about the usefulness of concepts such as *fabula* and *syuzhet* in analyzing works of art.

Supplementary Vignette: Participant RW4

Participant RW4 answered question 12 correctly and demonstrated both required behaviors, resulting in a PL of 1. RW4 was one of only three participants who answered question 12 correctly, and all three did so while demonstrating both required behaviors. RW4's lengthy engagement with the question is worth attending to because while it exhibits some missteps and hesitations, it also shows a participant using careful reasoning to push through a difficult passage and question despite serious struggle.

> [*Reads passage*] Huh? OK. So it's, like, talking about two different things. I'll have to come back to that. The question is, "Which choice best states the main idea of the text?" Yeah, it's, like, talking about two things.
>
> [*Reads answer choices*] Question: "Which choice—" What was the idea? Let me look at that. The theorists are distinguishing two different things, and one is narrative content and another is, like, an event kind of thing. OK. "But literary theorist [Mikhail Bakhtin] maintained—" OK. So it's, like, comparing two things to these, like, stories. "Discounted [*sic*]." OK. "Maintained." "Which choice best—" What was the main idea, though? It was, like, other than just the two stories? What other information— "He held that systematic [categorization; *participant says "characterization"*] . . . subtle ways [*sic*] [in; *participant says "of"*] which meaning is created by interactions . . ." So it's, like, artwork and audience, their interaction makes the account of the story. That's what seriously makes the story, like, a story, I guess?
>
> [*Reads choice A*] ". . . argue[d] that there are—Literary theorist [Mikhail Bakhtin] argue[d] that there are important characteristics of narratives that are not fully [encompassed; *participant says "uncomprised"*] by two

concepts that other theorists have used—" OK, what was [Bakhtin's] idea? Wasn't that what I just read? "Maintained that *fabula*—" OK, no, he's saying that there's more than just, like, two things.

What about B? "Literary theorist [Mikhail Bakhtin] claimed that meaning is not inherent in a narrative but is created when an audience encounters"— Yeah, so it's saying— ". . . interpreted differently . . ." OK, yeah, it could be B too, because it does mention "audience" in the story and in the answer, about the narrative, can interpret the story differently, I guess, by different people too.

OK. C, "The storytelling method[s] used in *The Godfather Part II*—" Doesn't, like—I wonder why B is the main idea if it's talking about two different things? It probably isn't C.

D, "Narratives that are told out of chronological order—" Well, it's not really mentioning that [*inaudible*] there. "[In the film] *The Godfather Part II*, the *fabula* is the story—" OK. And what if I go down to, like, "two timelines . . . But literary theorist . . ." Phew. [*Returns to passage*] [*Fabula,*] the narrative content, and *syuzhet*, the narrative arrangement and presentation of events. See, so it kind of does matter on the order, I guess, or—OK, wait, no. It's, like, it's saying a lot of theorists kind of base, like, two things, like, two ideas, while these other two people—or unless it's just one—are, like, talking about three different components, so. [*inaudible*]

[*Returns to choice C*] It "can be easily understood when two concepts from literary theory are utilized." I mean, it's talking—I don't even know.

[*Returns to passage*] The events in the film *The Godfather Part II* . . . "the *fabula* is the story of the [Corleone] family"—so we're mentioning two stories. Presentation of story form, and it goes "between two timelines and [*sic*] in 1901 and 1958." So it's kind of going in between. "But literary theorist . . ."—OK—"maintained that *fabula* and *syuzhet* are insufficient to completely—" OK, yeah, so he's saying that, like, not—these two things can't describe it, so that means— OK, so it's going to be C then. What is—no, because he's saying [*inaudible*], I guess. OK.

[*Returns to choice B*] "Literary theorist Mikhail [Bakhtin claimed] that meaning is not inherent, is not [*sic*] a narrative . . ." But it's, OK, so it could be B, then, because it's saying, like, it's really determined on who—

[*Returns to choice A*] "Literary theorist [Mikhail Bakhtin] argument [*sic*] that there are important characteristics of narratives that are not fully [encompassed; *participant struggles to pronounce*] by two concepts . . ." That makes me think A; it's either A or B.

OK, A is saying there are important characteristics of narrative that are not fully [encompassed; *participant says "incompromised"*] by two—so it's, like, more than just two things, which, kind of from the beginning of it, because he's, like, kind of not like the other, he's not like other narratives, or theorists. So maybe it's not A then. I mean, no, it could be A then actually. OK, so, yeah, it would be A, because he's kind of contradicting what they're saying. OK. I'm going to do answer A because it seems the most logical to, it's kind of like he's not like the other theorists and stuff.

Participant RW4

After reading the passage, participant RW4 demonstrates a basic grasp of its content and structure: “So it’s, like, talking about two different things.” She later clarifies that the “two things” she refers to are *fabula* (“narrative content”) and *syuzhet* (“an event kind of thing”). Before reading the answer choices, RW4 provisionally concludes that the main idea of the passage concerns “artwork and audience, their interaction,” which “makes the account of the story.” This belief explains the appeal that incorrect answer choice B (“. . . meaning is not inherent in a narrative but is created when an audience encounters a narrative so that narratives are interpreted differently by different people”) has for her throughout her work on the question. After reading choice A (the best answer), she narrows in on the true main idea of the passage: “OK, no, he’s saying that there are more than just, like, two things.” Predictably, RW4 is drawn to answer choice B (“OK, yeah, it could be B too, because it does mention ‘audience’ in the story and in the answer, about the narrative, can interpret the story differently, I guess, by different people too”). After quickly ruling out choice C and considering choice D, she gets closer to a more precise interpretation of the passage’s main idea: “It’s, like, it’s saying a lot of theorists kind of base, like, two things, like two ideas, while these other two people—or unless it’s just one [Bakhtin]—are, like, talking about three different components.” It’s not clear what this third component might be—the most likely answer is that it’s simply a placeholder for a factor other than *fabula* and *syuzhet*—but, in any event, at this point RW4 has demonstrated adequate passage comprehension (behavior 1) by figuring out that the gist of the passage is a contrast of the view of many literary theorists with the more expansive view of Bakhtin, a summary she refines a bit later: “OK, yeah, so he’s saying that, like, not—these two things can’t describe [a narrative adequately].” After ruminating on choice B, RW4 ultimately selects the best answer, choice A (behavior 2): “OK, so, yeah, it would be A, because [Bakhtin is] kind of contradicting what [other literary theorists are] saying.”

Reading and Writing Question 13

Skill/Knowledge Testing Point	Command of Evidence: Textual
Performance Score Band	4
Stimulus Subject Area	Science
Stimulus Text Complexity	SCO (upper secondary, grades 9–11)
Required Behaviors	1. Read and demonstrate comprehension of the passage. 2. Demonstrate an understanding of the relationship between the criterion set forth in the question’s stem and the passage. 3. Select the answer choice that best meets the criterion set forth in the question’s stem.
RW36 Performance Level	1

Fish whose DNA has been modified to include genetic material from other species are known as transgenic. Some transgenic fish have genes from jellyfish that result in fluorescence (that is, they glow in the dark). Although these fish were initially engineered for research purposes in the 1990s, they were sold as pets in the 2000s and can now be found in the wild in creeks in Brazil.

A student in a biology seminar who is writing a paper on these fish asserts that their escape from Brazilian fish farms into the wild may have significant negative long-term ecological effects. Which quotation from a researcher would best support the student's assertion?

A) "In one site in the wild where transgenic fish were observed, females outnumbered males, while in another the numbers of females and males were equivalent."

B) "Though some presence of transgenic fish in the wild has been recorded, there are insufficient studies of the impact of those fish on the ecosystems into which they are introduced."

C) "The ecosystems into which transgenic fish are known to have been introduced may represent a subset of the ecosystems into which the fish have actually been introduced."

D) "Through interbreeding, transgenic fish might introduce the trait of fluorescence into wild fish populations, making those populations more vulnerable to predators."

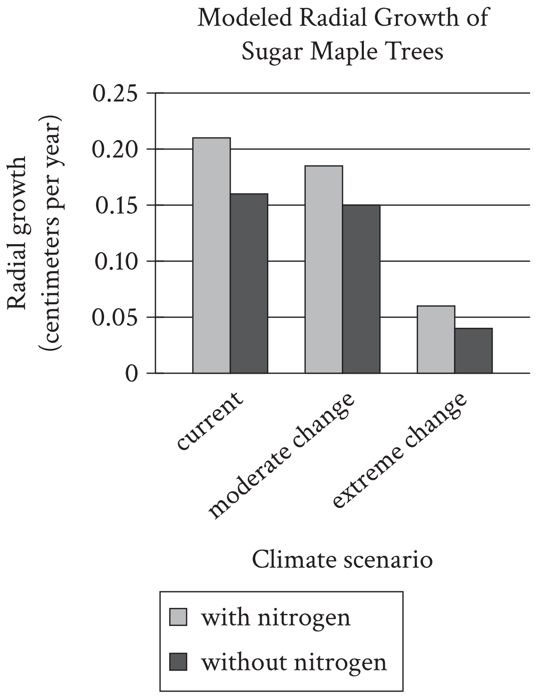
Question 13, a medium-difficulty (PSB 4) Command of Evidence: Textual question set in a challenging (SCO) science context, requires test takers to select the quotation from among the answer choices that best supports the student's claim that the escape from containment of transgenic fish "may have significant negative long-term ecological effects." The best answer is choice D. The passage defines the term *transgenic* as it relates to fish and brings up the example of fluorescent fish found in the wild in Brazilian creeks. Given this, choice D makes the most sense here, as it describes a tangible negative consequence of such fish escaping into the wild: By passing on their trait of fluorescence via breeding, these fish may make their populations more vulnerable to predators.

> I think it's, it has to be D since we're talking about the effect that, the negative effect that fluorescent fish have on the environment, and making [them] the more, or easier to see, against predators is one of them.
>
> *Participant RW36*

Participant RW36 answered the question correctly and demonstrated all required behaviors, resulting in a PL of 1. Though terse, RW36's response is sufficient to indicate both passage comprehension (behavior 1) and understanding of the relationship between passage and question stem (behavior 2). He correctly observes that the student's claim is about "the negative effect that fluorescent fish have on the environment" and then draws the proper inference that fluorescence would be a problem for fish in Brazilian creeks because it would make them "easier to [be] see[n]" by potential predators. With this, he selects the best answer, choice D, as his response (behavior 3).

Reading and Writing Question 14

Skill/Knowledge Testing Point	Command of Evidence: Quantitative (Graph subtype)
Performance Score Band	7
Stimulus Subject Area	Science
Stimulus Text Complexity	PSR (postsecondary readiness, grades 12–14)
Required Behaviors	1. Read and demonstrate comprehension of the passage. 2. Demonstrate an understanding of the graph, including what the graph as a whole as well as its various components (e.g., bars) represent. 3. Demonstrate an understanding of the relationship among the passage, the graph, and the criterion set forth in the question's stem. 4. Select the answer choice that best meets the criterion set forth in the question's stem.
RW36 Performance Level	5



Inés Ibáñez and colleagues studied a forest site in which some sugar maple trees receive periodic fertilization with nitrogen to mimic the broader trend of increasing anthropogenic nitrogen deposition in soil. Ibáñez and colleagues modeled the radial growth of the trees with and without nitrogen fertilization under three different climate scenarios (the current climate, moderate change, and extreme change). Although they found that climate change would negatively affect growth, they concluded that anthropogenic nitrogen deposition could more than offset that effect provided that change is moderate rather than extreme.

Which choice best describes data from the graph that support Ibáñez and colleagues' conclusion?

A) Growth with nitrogen under the current climate exceeded growth with nitrogen under moderate change, but the latter exceeded growth without nitrogen under extreme change.

B) Growth without nitrogen under the current climate exceeded growth without nitrogen under moderate change, but the latter exceeded growth with nitrogen under extreme change.

C) Growth with nitrogen under moderate change exceeded growth without nitrogen under moderate change, but the latter exceeded growth without nitrogen under extreme change.

D) Growth with nitrogen under moderate change exceeded growth without nitrogen under the current climate, but the latter exceeded growth with nitrogen under extreme change.

Question 14, a hard (PSB 7) Command of Evidence: Quantitative question set in a highly challenging (PSR) science context, requires test takers to use data from the graph to best support the conclusion of Ibáñez and colleagues that "anthropogenic nitrogen deposition could more than offset" the negative impact of climate change "provided that change is moderate rather than extreme." Choice D is the best answer, as it accurately and appropriately compares growth without nitrogen (i.e., without "anthropogenic nitrogen deposition," or artificial fertilization) under the current climate to both growth with nitrogen under moderate climate change and growth with nitrogen under extreme climate change. These comparisons are relevant to supporting the researchers' claim because the researchers assert that using nitrogen fertilizer will "more than offset" the effects of moderate climate change but not those of extreme climate change. This claim is supported by data in the graph drawn from two comparisons: first, that growth without nitrogen under the current climate (dark gray bar above the heading "current") is exceeded by growth with nitrogen under moderate climate change (light gray bar above the heading "moderate change"), which indicates an offsetting of the effects of moderate climate change via the use of artificial fertilizer, and, second, that growth without nitrogen under the current climate exceeds growth with nitrogen under extreme climate change (light gray bar above the heading "extreme change"), which indicates that the effects of extreme climate change can't be offset by adding nitrogen.

> I'm honestly just having a really hard comprehending the question since [the answer choices] always look the same with very slight differences. So just to make an educated guess, let's just start with C. So "the growth of nitrogen [*sic*] under moderate change." So we look at the chart, look at moderate change with nitrogen, is here. [*Puts pointer on "moderate change, with nitrogen" bar of graph*] Change "exceeded growth without nitrogen under moderate change." OK? "But the latter exceeded growth without nitrogen under extreme change." So that's partly right, but the end is wrong. [*Turns to choice B*] "Growth without nitrogen under [the] current climate exceeded growth without nitrogen under moderate

> [change; *participant says "climate"*]." OK? "But the latter exceeded growth with nitrogen under extreme change." So it seems like just all of them are the same, at the end, because it's "growth without nitrogen," "growth without nitrogen"— Hmm, the last one [choice D] is different. Oh, same with B. [*Turns to choice A*] "Growth with nitrogen under the current climate change [*sic*] exceeded growth with nitrogen under moderate change." OK. "But the latter exceeded growth without nitrogen under extreme change." I think, just to make an educated guess, I think it's C.
>
> *Participant RW36*

Participant RW36 answered the question incorrectly and failed to demonstrate any required behaviors, resulting in a PL of 5. RW36 exhibits a lack of understanding of the question itself ("I'm honestly just having a really hard time comprehending the question since [the answer choices] always look the same with very slight differences") and elects to make an "educated guess," which seems ultimately more like a random guess, of choice C.

Supplementary Vignette: Participant RW41

Participant RW41 answered question 14 correctly and demonstrated all required behaviors, resulting in a PL of 1. RW41 was one of only three participants to have answered the question correctly and the only participant to have done so while also demonstrating all required behaviors.

> [*Reads passage*] So I'm gonna go back. They studied a forest site in which some of the sugar maple trees received a periodic fertilization with nitrogen to mimic the broader trend of increasing anthropogenic nitrogen deposition in soil. So they modeled the growth of the trees under three different climate scenarios. They found that the climate change would negatively affect the growth.
>
> So I'm gonna go back to the question choices.
>
> [A,] "Growth with nitrogen under the [current] climate exceeded growth with nitrogen under moderate change, but the latter exceeded the growth [*sic*] without nitrogen under extreme change." I don't think it's A.
>
> [B,] "Growth without nitrogen under the current climate exceeded growth without nitrogen under moderate change, but the latter exceeded growth with nitrogen under extreme change." I think that's definitely a choice.
>
> C, "Growth with nitrogen under moderate change exceeded growth without nitrogen under moderate change, but the latter exceeded growth without nitrogen under extreme change." I think B would be a better choice than C.
>
> And then [D,] "Growth with nitrogen under moderate change exceeded growth without nitrogen under the current climate, but the latter exceeded growth with nitrogen under extreme change."
>
> So I'm gonna go back to the chart, and I can see that the current [climate, with nitrogen] is at approximately 0.21. OK. And then the moderate change [with nitrogen], it's around [0.]17, without nitrogen is at [0.]15. So I believe that B would be the best choice because in the text, it states that—I'm just going to start from here—"Although they found that climate

> change would negatively affect the growth [*sic*], they concluded that anthropogenic nitrogen deposition could more than offset that effect provided that change is moderate rather than extreme."
>
> Which matches with choice B, which says, "Growth without nitrogen under the current climate exceeded growth without—" Actually, it wouldn't be B because it says that, it states it exceeded growth, and it negatively affected growth. So, actually, I think it would be D because it negatively affected the growth; it didn't really exceed. So I think it's D.
>
> *Participant RW41*

Participant RW41 demonstrates adequate passage comprehension (behavior 1) most clearly in his summary of the passage: "They studied a forest site in which some of the sugar maple trees received a periodic fertilization with nitrogen to mimic the broader trend of increasing anthropogenic nitrogen deposition in soil. So they modeled the growth of the trees under three different climate scenarios. They found that the climate change would negatively affect the growth." RW41 shows comprehension of the graph (behavior 2) when accurately citing the (sometimes approximate) value of several of the graph's bars: "I can see that the current [climate, with nitrogen] is at approximately 0.21. OK. And then the moderate change [with nitrogen], it's around [0.]17, without nitrogen is at [0.]15." He exhibits a grasp of the relationship among the passage, graph, and question (behavior 3) by the way in which he rules out choice B: "Actually, it wouldn't be B because it says that, it states it exceeded growth, and it negatively affected growth." Though somewhat oblique, this comment suggests attainment of behavior 3 in that RW41 determines that choice B can't be the best answer because its terms of comparison don't align with those of the passage's central claim—specifically, that it's true but irrelevant that growth without nitrogen under moderate change exceeded growth with nitrogen under extreme change, as the passage asserts that artificial fertilization can offset the effects of climate change under moderate but not extreme conditions. After this dalliance with choice B, RW41 ultimately picks the best answer, choice D (behavior 4): "So, actually, I think it would be D because it negatively affected the growth; it didn't really exceed. So I think it's D."

Reading and Writing Question 15

Skill/Knowledge Testing Point	Inferences
Performance Score Band	4
Stimulus Subject Area	History/social studies
Stimulus Text Complexity	MID (middle school/junior high, grades 6–8)
Required Behaviors	1. Read and demonstrate comprehension of the passage. 2. Select the answer choice that most logically completes the passage.
RW36 Performance Level	1

In dialects of English spoken in Scotland, the "r" sound is strongly emphasized when it appears at the end of syllables (as in "car") or before other consonant sounds (as in "bird"). English dialects of the Upland South, a region stretching from Oklahoma to western Virginia, place similar emphasis on "r" at the ends of syllables and before other consonant sounds. Historical records show that the Upland South was colonized largely by people whose ancestors came from Scotland. Thus, linguists have concluded that ________

Which choice most logically completes the text?

A) the English dialects spoken in the Upland South acquired their emphasis on the "r" sound from dialects spoken in Scotland.

B) emphasis on the "r" sound will eventually spread from English dialects spoken in the Upland South to dialects spoken elsewhere.

C) the English dialects spoken in Scotland were influenced by dialects spoken in the Upland South.

D) people from Scotland abandoned their emphasis on the "r" sound after relocating to the Upland South.

Question 15, a medium-difficulty (PSB 4) Inferences question set in a moderately challenging (MID) history/social studies context, requires test takers to complete the text (i.e., fill in the blank) with the most logical text-based inference. Choice A is the best answer. The passage establishes, first, that the "r" sound is sometimes strongly emphasized in English dialects spoken in Scotland; second, that English dialects in the Upland South of the United States carry the same emphasis; and, third, that the Upland South region was largely colonized by Scots. The most logical inference from this information is that the English dialects spoken in the Upland South gained their emphasis on the "r" sound from English dialects spoken in Scotland.

> The one that really makes sense is A because the other ones just make zero sense. Because with B, most dialects just don't spread like that. C, the dialect [*sic*] spoken in Scotland wasn't influenced by the Upland South. And D, they wouldn't just abandon their dialect just like that after moving to the Upland South.
>
> *Participant RW36*

Participant RW36 answered the question correctly and demonstrated both required behaviors, resulting in a PL of 1. RW36's response exhibits adequate passage comprehension (behavior 1), most clearly in the way in which RW36 rules out choice C. By correctly observing that "the dialect [*sic*] spoken in Scotland wasn't influenced by the Upland South," RW36 draws on the passage's claim about the directionality of influence from Scotland to the Upland South. In partial contrast, he seems to have ruled out choices B ("most dialects just don't spread like that") and D ("they wouldn't just abandon their dialect just like that after moving to the Upland South") more on prior knowledge and a general sense of how the world works ("the other ones just make zero sense"). In any event, RW36 properly selects choice A as the best answer (behavior 2).

PARTICIPANT PERCEPTIONS

Following the think-aloud activity, Reading and Writing participants were asked a standardized set of six follow-up questions. An analysis of participants' responses to each of the questions follows.

General Impressions

1. Please tell me a bit about the experience you just had. What was it like to answer those questions?

Participant responses to postexperience question 1, which concerned their general impressions of the think-aloud experience, expressed sentiments that were typically positive or neutral toward the experience.

> It was all, it was pretty all right. *RW11*
>
> Yeah. I mean, there were a couple of confusing questions. Other than that, it seemed decently OK. *RW13*
>
> But, yeah, other than [there being no dark mode], I'd have to say it was a pretty OK—or rather pleasant—experience. It wasn't too shabby. *RW14*
>
> So reading it, reading something and giving it to me, and then asking a question based on that . . . simple, easy. *RW23*
>
> Some of the questions were really easy compared to others. For example, the fill-in-the-blanks with simple one-word answers were easier for me personally. But doing all the reading with the long answer questions is a little bit harder for me personally. *RW30*
>
> Answering them out loud felt, for some reason, oddly good because I never get to do that. Usually, when I take a test, I have to be all quiet and think in my mind. But now, whenever I speak it out loud, it feels so natural for some reason. I just talk a lot in my life, and it feels really nice to do that while taking a test. *RW34*
>
> I think some of them were a lot more challenging than others because some of the questions were worded a little weird, and it was kind of hard for me to understand what it was, like, meaning. So I think some of them were harder, but then some of them were really vague and, like, the answer was answered in the question or in one of the answers. *RW39*
>
> It was difficult for me, but it wasn't as hard as I thought it would be. The biggest challenge for me was reading some of the passages out loud, especially the ones with longer words or complex sentences. But, overall, I think I did OK. *RW40*
>
> I mean, I feel like I always do on every test, but I felt like reading it out loud and then having to think about it out loud helped a little bit. But I feel like they were a little more difficult than the regular ones we get. *RW43*

RW36, the case study participant, was the only one with a strongly negative perspective: "Personally, I honestly hated it. It was like, so it wasn't difficult, but reading them out loud, I struggled a lot. And there's a lot of words I didn't know that are hard to understand. And, as I said before, there are some words that repeated several times that just kind of threw me off."

Several participants remarked, in one way or another, on the artificiality of the think-aloud process used in the study. Most of these comments were broadly neutral. "Hearing yourself talk—the talk-out-loud part—was different," RW8 noted, "but answering the questions wasn't hard." RW14 mentioned having "a little bit of anxiety knowing that I [had] to read it out loud, but other than that, it wasn't too bad." By contrast and as noted above, RW34 thought the thinking-aloud experience was "oddly good," and RW43 felt reading aloud "helped a little bit" with answering the questions. RW36 was, as indicated above, the only one to strongly signal a negative experience.

Strategies

2. How would you describe your general approach, in terms of strategies, for answering the questions?

Postexperience question 2, which concerned participants' strategy use while answering the think-aloud questions, elicited information about the specific strategies used during the activity as well as general test-taking approaches. As this variegated response suggests, this question proved somewhat ambiguous, as it's not always possible from participants' answers to distinguish their think-aloud approach from their more typical silent test-taking practices.

Rereading was the most commonly cited strategy.

> So repeating it multiple to, like, fully understand, because the first couple of times, I was kind of just reading it and I'm like, "Wait, what am I reading?" So, yeah, I just had to, like, think about it multiple times to fully, like, understand and get it in my head. *RW4*
>
> Well, I like to take my time and kind of read very fast. But if I catch myself doing that, I have my arrow over the word [on the screen], and sometimes I reread what I just read if I need to clarify something. *RW11*
>
> I think over the years, I've figured out ways and tactics to help me answer questions better. So I think with reading it—reading the passage once—and then if I don't understand it, I read it again, and then going to the questions, just kind of being really repetitive, which is exhausting after a while. It's something, you know, no kid should have to do, but if it works, it works. And I think just being repetitive and having to read things over and over and over again helps. *RW22*

A number of participants also mentioned using incorrect-answer elimination, sometimes in combination with rereading.

> I like to do the elimination, just eliminate it as I go, and just like I said, just keep reading the text over and over and just trying to make [context; *participant says "contact"*] clues on what answer to choose from. *RW8*

> Normally whenever I'm answering questions, I always eliminate the [choices] I know are not correct. And I always, before I pick an answer, I go back and scan the text. *RW39*

Participants also frequently mentioned various approaches that could be classified under the heading of "fit," or the sense that the selected answer is the best match for the question. This "fit" could be based on an overall sense of the passage, as participant RW23 observes:

> My strategy to answer the question[s]? Well, I don't really have a strategy. Oh, well, what I would do is, I would go to the question and ask me—and I would see, I would read, obviously, what it's asking me. I would look at the multiple-choice [options], and I would see if it either fits for the question, and then I would go back into the paragraph or sentence that it was given to me and see if it's, like, correct and fits, and then I would just answer that.

Keyword matching, or efforts to line up key concepts in the stimulus with identical or similar language in the multiple-choice answer options, was a frequently mentioned variation on the idea of general "fit."

> I'd say when you read the question, look for, like, keywords because, like, you know how I specifically found, like, that underline[d] part [of the passage] [Reading and Writing question 3]? I knew that was gonna be useful inside the question sooner or later. So when I read it, I saw that it was the underlined [portion of the passage], and I was like, OK, then that's probably 100 percent a keyword. So I went to go there, and there was some keywords there that I can use inside all the answers. So I use those keywords to find out the answer. *RW34*

> . . . my main approach was to read [the passage], dissect it, then read the question and dissect that, and then find the similarities between the two and just kind of build off of that. *RW36*

Participant RW30 directly referenced the use of a "plug-in" strategy for certain question types in which a blank must be filled to complete the text: "So if it's a fill-in-the-blank question, I like to emphasize the sentence and then try plugging in the answer and see how it fits there."

Participants mentioning the order in which they approached individual test questions were split between those who would typically read the passage first and then the question and those who would typically begin with the question. (Recall that the protocol for this study required that participants read the passage out loud and then the question.)

> I would normally, I would actually read the question first and the answers and then read the passage and then go back, read again. But just taking different approach, and sometimes I do read the passage first and then the questions and go back to passage . . . *RW36*

> I usually like to read the question first before reading the text, just so I could get an idea of what the question is trying to ask me so I can get important details. *RW41*

> Well, I like to read the passage and get it out of the way because I feel like when I read that, I get to read the passage first. That makes me, like, the most tired. Like, when taking tests, I read the passage first and then read the question. So it makes it easier and I can remember it better. *RW43*

"Easy" Question Types

3. Was there a particular type of question that you found especially easy to answer? If so, which one and why?

Postexperience question 3, which asked about the types of questions in the activity that participants found particularly easy to answer correctly, tended to identify three, sometimes overlapping, characteristics of the Reading and Writing questions that participants associated with ease: (1) short passages, (2) blank-completion passages/tasks, and (3) literature passages.

> . . . one of them was the, like, the wallpaper one [Reading and Writing question 5] because, like, the, it was pretty, it was kind of, like, obvious. Which one was it? Because there's only, like, one [answer choice] that was saying, like, "I like how this room looks, but [for] the wallpaper" while the other ones were saying, like, oh, yeah, no, "The wallpaper just kind of—like, the whole room just looks kind of weird." And the one, the one question talking about . . ., like, the neighborhood, the street watching [Reading and Writing question 11], that one seemed pretty obvious because—I can't think, I don't know—like, the answer was kind of, like, out there, and it really just looked different than all the other answers. *RW4*

> I think the question where you have to, like, fill in the blank with, like, the right word or something—like, so be, like, the best word that's suitable for the sentence—those kind of questions are kind of easy for me. 'Cause after I read them, I feel like I immediately knew what the answer was. I think those ones were the easiest. *RW11*

> Especially easy to answer? Uh, yeah, there's a few. And why? Because it was just simple. It, and some of them was just small stories. It was just something to read quickly, and then it just, it was just kind of obvious, you know? *RW23*

"Hard" Question Types

4. Was there a particular type of question that you found especially hard to answer? If so, which one and why?

Participants' responses to postexperience question 4, which concerned the question types in the think-aloud activity that students found most challenging, repeatedly called out two factors: (1) long passages and (2) passages including informational graphics. Reading and Writing question 14, a very hard (PSB 7) Command of Evidence: Quantitative question set in a highly challenging (PSR) science context, was called out multiple times, as it had a relatively long stimulus, an informational graphic, and answer choices that, due to the nature of the comparisons being drawn, varied only slightly from option to option.

> Those logistical ones? Like, the one I was just working on [Reading and Writing question 14] because I was, because, like, a lot of the words were the same. And so it's, like, repeating it multiple times, was, like, kind of hard because I was like, wait, what am I reading? What am I trying to understand? Because, like, it was just, there's too much of the same thing going on. And so I didn't know what I was supposed to read because, like, I had to think about, like, this growth in population. I had to, like, look at the chart, and then I had to read the passage, and I'm like, what am I reading? *RW4*

> The one I can remember from right now is the one with the scientist—no, the tree and the nitrogen [Reading and Writing question 14]. And I wouldn't say it was hard. It was just more of a carefully reading and going back to the chart and then going back to the answers, are just going back and forth a lot, just trying to make sure you choose the right answer and also getting the right clues as well. *RW8*

> . . . those really long-passage questions, like, I really, like, take my time and reread . . . I could understand it better, but, like, right off the bat, it [was] kind of overwhelming, I guess, to me sometimes, most of the time, I don't know. *RW11*

> Yeah, I think the more longer, like, more paragraph-type questions are more difficult—like, the long words, so much words and stuff. *RW13*

> The nitrogen maple tree [Reading and Writing question 14]. Uh, yeah, I didn't like the repetitive words. And I feel like the passage was unneeded and that you really didn't need it to answer the question, which just created more work. I didn't like that one. *RW22*

> All the ones with charts and graphs were hard. *RW39*

SLDR Symptom Impact

5. Did you encounter anything in the questions that you had difficulty with given that you have a specific learning disorder affecting reading? If so, what was it, and why was it difficult for you?

Postexperience question 5 was intended to elicit from participants their perceptions of any specific impacts that their SLDR symptoms had on their ability to answer the Reading and Writing test questions presented to them. This question was designed to gain more information about SLDR test takers' experiences and identify potential construct-irrelevant barriers to their fair access to test content that College Board should further investigate and, if possible, remediate. Somewhat complicating interpretation here is the fact that the think-aloud protocol used deviates from the far more typical silent test taking that students engage in.

Participants almost exclusively reported issues internal to themselves. These prominently included (1) impacts of their SLDR condition on their executive function abilities, including challenges with maintaining focus and attention, keeping information and ideas in working memory, and having adequate stamina to persist productively in test taking and (2) difficulties with text processing.

> So I have dyslexia and ADHD as well as autism. So it's like, I naturally look at things slightly different, and, like, the whole time while I was reading all, like, fidgeting and stuff because I was trying to, like, get myself to focus. What was I saying? Yeah, I feel like even right now, like, when I was reading, I was kind of, like, getting distracted in my head, and I had to, like, refocus myself. *RW4*

> . . . I think that my disability makes me forget easily. I feel like I would have, I feel like it was kind of hard because I would have like[d] some sort of little dictionary or thesaurus screen so I could have quickly reviewed what [a] word would mean. Other than that, I honestly don't know. Like, I guess I just forget easily. That's why I have to read things all over again. And I, sometimes I read too fast, and I don't understand what I'm doing. So I have to reread all over, and it's a lot easier for me when I have my mouse so I could read along with the mouse. That helps me, but [it's] kind of hard when I don't do that and speak aloud sometimes to myself so quietly. *RW11*

> I mean, I was struggling to read a lot of words, if you couldn't tell, like names and stuff. But it's just—I used a lot of brain power today to read some words I don't usually see. It was really hard for me because I had to sound it out. *RW14*

> For me, it was the longer passages because I have ADHD, and just kind of sitting there and constantly reading a super-long passage is exhausting. And sometimes I will be reading it, and I will zone out while I'm reading it. I'm just gonna be reading it out loud on autopilot. So I've been reading it, but I don't understand what I just read. So I have to read it over again, and it's back to—then I have to read it over again, and over and over and over, until I finally understand it. But if it's something interesting and short, it's easier for me to process and understand. And then also with the repetitive words and the long texts and the long passages—I'm also, I have my IEP for dyslexia too, so I get my words mixed up and, like, swapped around. And, you know, when things are repeated a lot, it makes it look like every single answer choice is the exact same thing. And sometimes I will read the passage wrong, and it could have a totally different meaning because I read it so wrong because of the words switching around due to the, yeah. *RW22*

> Just reading words out loud. Because I'm good at reading, and I mainly read in my head fast, so it's hard to slow down and pronounce syllables out loud. *RW30*

Two participants also or instead mentioned aspects of the Reading and Writing questions that caused them issues.

> It was kind of hard to comprehend, and especially when they were italicized or when the font was different, it was kind of confusing because it was just going from print to different fonts, which confused me. And how there were random hyphens in the middle of the passages. *RW39*

> I remember there was, like, a question with different animals within the Ice Age, like, a woolly mammoth and a saber-toothed tiger [Reading and Writing question 4, which used both common and scientific names for various animals]. I found those—there were different abbreviations for their descriptions, I guess—and I found those kind of challenging to read. *RW41*

Finally, one participant, in response to a different postexperience question, mentioned Bluebook's lack of a dark mode as an impediment.

> I had to turn the brightness of my screen down because looking at it for too long definitely gives me headaches. But, yeah, other than that, I'd have to say it was a pretty OK—or rather say pleasant—experience. It wasn't too shabby. *RW14*

Final Comments

6. Is there anything about your test-taking experience today or about the test-taking strategies you used today that we haven't talked about yet but that you'd like us to know?

Postexperience question 6 was a nondirective query intended to elicit any feedback from participants not otherwise addressed by prior interview questions.

Question 6 typically didn't yield feedback from participants, with a few exceptions. Participant RW4 noted that she uses drawing and doodling during testing as a way to stay focused: "Something about me is that whenever I do, like, something, like, work related, I usually do draw because that's something that keeps me, like, motivated, keeps me, like, in the zone." Participant RW11 felt as though having a dictionary or thesaurus during testing "would be helpful," as did participant RW43. Participant RW22 recommended that answer choices for fill-in-the-blank questions should "just, like, pop up there, and you could see what it would look and sound like when it's in the blank, if that makes sense."

Math

PARTICIPANT AND QUESTION PERFORMANCE

Participant and Question Performance Levels and Differentials

Figure 2 displays, as a single matrix, the Math participant and question performance data derived from this study. An explanation of the intended method of reading the figure is provided in the corresponding subsection of the Reading and Writing results, above, although the following differences should be observed:

- For the Math domain, expected behaviors, rather than required behaviors, were defined to account for the fact that some Math questions are, by design, open to multiple, often mutually exclusive solution paths.
- Because of the above difference, PL 2 was unobtainable by Math participants, as they were only expected to answer each question correctly and demonstrate at least one expected behavior. (For Reading and Writing, by contrast, PL 2 was attainable for questions with more than two required behaviors by participants who answered a given question correctly and demonstrated one or more additional required behaviors but not all such behaviors.)

Figure 2. Math Participant and Question Performance Summary Matrix.

Part. ID	Question #															Performance by Level, by Participant						Participant Performance Summary		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	1	2	3	4	5	*NR*	*#AC*	*#EB*	*D_p*
M10	1	1	5	5	4	5	1	1	1	1	1	4	5	1	5	8	–	0	2	5	0	8	8	0 ✔
M11	1	5	3	5	5	5	1	1	1	1	5	5	5	5	5	5	–	1	0	9	0	6	5	1 ✔
M13	1	1	5	5	5	5	1	1	1	5	1	1	–	–	–	7	–	0	0	5	3	7	7	0 ✔
M16	3	5	5	5	5	5	1	5	5	1	5	5	5	5	5	2	–	1	0	12	0	3	2	1 ✘
M20	5	1	5	5	5	5	1	4	3	5	5	5	5	5	5	2	–	1	1	11	0	3	2	1 ✘
M21	1	5	5	5	5	5	3	5	1	5	1	5	5	5	5	3	–	1	0	11	0	4	3	1 ✔
M22	5	5	5	5	4	5	5	5	1	5	5	5	5	5	5	1	–	0	1	13	0	1	1	0 ✔
M24	1	5	3	5	5	5	5	5	5	5	5	4	4	5	5	1	–	1	2	11	0	2	1	1 ✘
M26	5	5	5	5	5	5	5	5	5	1	5	5	5	5	5	1	–	0	0	14	0	1	1	0 ✔
M27	1	1	5	5	5	5	1	1	1	1	1	1	1	5	5	9	–	0	0	6	0	9	9	0 ✔
M28	1	5	5	5	5	5	1	1	1	5	5	5	5	5	5	4	–	0	0	11	0	4	4	0 ✔
M32	5	5	5	5	4	5	1	1	1	1	4	5	5	1	4	5	–	0	3	7	0	5	5	0 ✔
M33	1	1	1	1	4	5	1	1	1	1	1	1	1	1	1	13	–	0	1	1	0	13	13	0 ✔
M34	1	1	1	5	4	5	1	1	1	1	1	1	4	5	4	9	–	0	3	3	0	9	9	0 ✔
M38	1	1	1	1	4	3	1	1	1	1	1	1	1	5	1	12	–	1	1	1	0	13	12	1 ✔
M39	1	5	1	1	4	3	1	5	5	5	4	5	5	5	5	4	–	1	2	8	0	5	4	1 ✔
M44	1	1	1	1	4	5	1	5	5	5	5	4	1	1	4	7	–	0	3	5	0	7	7	0 ✔
M46	1	1	5	3	5	5	1	1	1	5	5	1	5	5	5	6	–	1	0	8	0	7	6	1 ✔
M56	1	5	5	5	4	5	1	1	1	1	5	1	5	1	5	7	–	0	1	7	0	7	7	0 ✔
M58	5	5	3	5	5	5	1	5	1	5	5	5	5	5	5	2	–	1	0	12	0	3	2	1 ✘
M60	1	1	5	5	4	3	1	4	1	5	1	5	5	5	1	6	–	1	2	6	0	7	6	1 ✔

Performance by Level, by Question

1	15	10	5	4	0	0	17	11	15	10	8	7	4	5	3
2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3	1	0	3	1	0	3	1	0	1	0	0	0	0	0	0
4	0	0	0	0	10	0	0	2	0	0	2	3	2	0	3
5	5	11	13	16	11	18	3	8	5	11	11	11	14	15	14
NR	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1

Performance Legend

1 (highest): Answered correctly; exhibited 1+ expected behaviors

2: *Not applicable to Math*

3: Answered correctly; exhibited no expected behaviors

4: Answered incorrectly; exhibited 1+ expected behaviors

5 (lowest): Answered incorrectly; exhibited no expected behaviors

Question Performance Summary

#AC	16	10	8	5	0	3	18	11	16	10	8	7	4	5	3
#EB	15	10	5	4	0	0	17	11	15	10	8	7	4	5	3
D_q	1 ✔	0 ✔	3 ✘	1 ✔	–	3 ✘	1 ✔	0 ✔	1 ✔	0 ✔	0 ✔	0 ✔	0 ✔	0 ✔	0 ✔

Summary Legend

#AC = # answered correctly

#EB = # answered correctly; demonstrated 1+ expected behaviors

D_p, D_q = Differentials (*#AC* − *#EB*); ✔ = criterion-passing differential (70%+), ✘ = criterion-failing differential (<70%)

Findings

Participant Performance

As shown in the "Participant Performance Summary" sub-table of figure 2, seventeen of twenty-one participants (81 percent) met or exceeded the criterion for a good D_p, which provides evidence that these participants were able to adequately demonstrate cognitively complex thinking in line with the question types' constructs. The four participants who didn't meet the criterion were also among the lowest-performing students on this activity, as judged by raw question-answering success, as they answered either two (one participant) or three (three participants) questions correctly (though it should be noted that two criterion-meeting participants themselves each answered only one question correctly). These criterion-failing participants did, however, attain differentials of 1, meaning that they were still able to demonstrate cognitively complex thinking on half to two-thirds of the (small number of) questions they did answer correctly.

Question Performance

As shown in the "Question Performance Summary" sub-table of figure 2, twelve of the fifteen studied Math questions (80 percent) met or exceeded the criterion for a good D_q, which provides evidence that these questions are capable of eliciting cognitively complex thinking from students with SLDR. Two of the remaining questions had differentials of 3, while the third question lacked a true differential, as no participant answered it correctly; one of the former two questions was answered by five participants who also demonstrated at least one expected behavior, suggesting that this question, too, was capable of eliciting cognitively complex thinking and that this higher-than-desirable differential may have been attributable in part to some participants' relative lack of think-aloud verbalization skill or experience. A common feature of the three criterion-failing questions was their high difficulty: One of the three had a PSB of 6 (out of a possible 7), while the other two—including the question that no participant answered correctly—had PSBs of 7. All three criterion-failing questions were in the multiple-choice format and lacked a context. The latter is notable here, as it suggests that these SLDR participants didn't struggle with text processing so much as the underlying content, a construct-relevant factor.

PARTICIPANT PERFORMANCE VIGNETTES

Case Study: Participant M34

Participant M34 was selected as the Math case study participant using the same criteria as outlined for the Reading and Writing case study. M34, a female eleventh grader from Texas, identified as White and not of Hispanic, Latino, or Spanish origin. She self-reported a HSGPA of A−, indicated that she'd received or she expected to receive an extra time accommodation as part of SAT Suite testing, and described her SLDR symptoms as moderate. M34 answered nine of the fifteen Math questions correctly and demonstrated at least one expected behavior in every case, resulting in a participant differential of 0 (100 percent), which exceeded the criterion for a good D_p.

Math Question 1

Content Domain	Algebra
Skill/Knowledge Testing Point	Linear Inequalities: Identify
Performance Score Band	4
Stimulus Subject Area	Science
Question Format	MC
Expected Behaviors	1. Read and demonstrate comprehension of the context described. 2. Set up/identify a linear equation or inequality as described in the context.
M34 Performance Level	1

For a snowstorm in a certain town, the minimum rate of snowfall recorded was 0.6 inches per hour, and the maximum rate of snowfall recorded was 1.8 inches per hour. Which inequality is true for all values of s, where s represents a rate of snowfall, in inches per hour, recorded for this snowstorm?

A) $s \geq 2.4$

B) $s \geq 1.8$

C) $0 \leq s \leq 0.6$

D) $0.6 \leq s \leq 1.8$

Question 1, a medium-difficulty (PSB 4) multiple-choice Linear Inequalities: Identify question set in a science context, requires test takers to identify a linear inequality that represents the given context. The correct answer (*key*) is choice D. It's given that the minimum and maximum rates of snowfall recorded were 0.6 and 1.8 inches per hour, respectively. Therefore, the rate of snowfall, *s*, ranges from 0.6 to 1.8 inches per hour.

> So, the minimum [rate of snowfall recorded]—I'm going to write this down—the minimum is 0.6 [inches per hour], and the maximum was 1.8. So this is, I feel like this is saying something about the average. So, if it's talking about average, then what is zero? This answer, A, does not make sense. I feel like because it—oh, I should check that. If it's 0.6 plus 1.8 divided by 2, [that] would be the mean, I think, and if that's what it's talking about—but that doesn't really make sense. Oh, actually, OK. The answer is not B because it can't get bigger than 1.8, and that's saying it's bigger than 1.8. And A just does not make sense either because there's no relationship between adding those together. I think adding them together doesn't mean anything because that's minimum and maximum, I think. Yeah. So that leaves C. So that would be *s*, oh, no. Oh, no, I was wrong. I think the correct answer is D because it's, that's saying that it's greater than 0.6 but less than 1.8, which is in the middle of the—like, between the two.
>
> *Participant M34*

Participant M34 answered the question correctly and demonstrated both expected behaviors, resulting in a PL of 1. After reading and demonstrating comprehension of the context ("the minimum [rate of snowfall recorded] is 0.6 [inches per hour], and the maximum was 1.8"; behavior 1), M34 momentarily believes this question is "something about the average" of the two given numbers, 0.6 and 1.8. However, after computing the mean, M34 quickly recognizes her error ("but that doesn't really make sense"). She correctly determines that choice B represents the rate of snowfall, *s*, being "bigger than 1.8" (in actuality, a rate equal to or greater than 1.8), which would exceed the given maximum rate, and that choice A involves the sum of the maximum and minimum values (2.4) ("there's no relationship between adding those together"). She next seems to assume that choice C must be correct but, on further review of choice D, acknowledges this can't be the case. M34 then properly selects the option, choice D, that identifies the linear inequality described in the context (behavior 2), explaining that *s* is "greater than 0.6 but less than 1.8, which is in the middle of the—like, between the two." Although M34 doesn't acknowledge that all the choices use greater-than-or-equal-to and/or less-than-or-equal-to signs, not just greater-than/less-than, she nonetheless demonstrates the fundamental understanding that *s* should be represented as being bounded by the given minimum and maximum snowfall rates.

Math Question 2

Content Domain	Problem-Solving and Data Analysis
Skill/Knowledge Testing Point	Ratios
Performance Score Band	5
Stimulus Subject Area	Real-world topics
Question Format	MC
Expected Behaviors	1. Read and demonstrate comprehension of the context described. 2. Use the ratio and given information to set up and solve a proportion.
M34 Performance Level	1

At a particular track meet, the ratio of coaches to athletes is 1 to 26. If there are *x* coaches at the track meet, which of the following expressions represents the number of athletes at the track meet?

A) $\frac{x}{26}$

B) $26x$

C) $x + 26$

D) $\frac{26}{x}$

Question 2, a medium-difficulty (PSB 5) multiple-choice Ratios question set in a real-world context, requires test takers to identify the expression that best represents the situation by either logically deducing this relationship from the context or through calculation by setting up a proportion. The correct answer is choice B. It's given that at a particular track meet, the ratio of coaches to athletes is 1 to 26. Logically, a test taker could determine from the context that the number of athletes at this track meet, given the provided ratio and *x* number of coaches, must be 26*x* (choice B), as the ratio indicates that there are twenty-six athletes for every coach. By calculation, a test taker could arrive at the same conclusion by setting up and solving the proportion $\frac{1\ coach}{26\ athletes} = \frac{x\ coaches}{y\ athletes}$, resulting in $y = 26x$, where *y* represents the number of athletes at the track meet.

> So there's 1 coach for every 26 athletes. That means it'd be 1 times 26. OK. It's not A because you can't divide—like, it's not A or D because you can't divide people. And that doesn't—1 to 26, there's no division happening. And C, it's not C either because you can't add coaches and athletes, kind of. So it'd be, the answer is B because it's like, if there were 2 coaches, it'd be 26 times 2, but there's 1. So it's 26 times 1.
>
> *Participant M34*

Participant M34 answered the question correctly and demonstrated both expected behaviors, resulting in a PL of 1. After reading the context, M34 correctly restates the given information in a way that better connects it to what's being asked (behavior 1): "So there's 1 coach for every 26 athletes. That means it'd be 1 times 26." Next, M34 dismisses the incorrect answer choices by making less conclusive claims about them. Her ruling out of choices A and D because "you can't divide people" and choice C because "you can't add coaches and athletes, kind of" seems to be an attempt to relate the options to the context, although these interpretations aren't as mathematically precise as one might wish. While M34 doesn't use the variable *x* in her verbalization, as is found in choice B, she uses "26 [athletes] times 2 [coaches]" as a way to express the idea of applying the ratio given a variable number of coaches (behavior 2), allowing her to verify the key: "So it'd be, the answer is B because it's like, if there were 2 coaches, it'd be 26 times 2, but there's 1. So it's 26 times 1."

Math Question 3

Content Domain	Geometry and Trigonometry
Skill/Knowledge Testing Point	Circles
Performance Score Band	6
Stimulus Subject Area	None
Question Format	MC
Expected Behaviors	1. Using the graph of a circle in the *xy*-plane, determine a possible *x*-value on the graph. 2. Identify the center of a circle in the *xy*-plane. 3. Identify the radius of a circle in the *xy*-plane. 4. Using the equation of a circle in the *xy*-plane, identify the domain of the circle.
M34 Performance Level	1

$$(x+4)^2+(y-19)^2=121$$

The graph of the given equation is a circle in the *xy*-plane. The point (a, b) lies on the circle. Which of the following is a possible value for *a*?

A) -16

B) -14

C) 11

D) 19

Question 3, a hard (PSB 6) multiple-choice Circles question outside of context, requires test takers to demonstrate an understanding of where the graph of a circle exists in the *xy*-plane by identifying a possible *x*-coordinate of a point that lies on that circle. The correct answer is choice B. The standard equation for a circle is $(x-h)^2+(y-k)^2=r^2$, where *h* and *k* represent, respectively, the *x*- and *y*-coordinates of the circle's center and where *r* represents the circle's radius. The equation given in the question is written in this standard form, meaning that the described circle's center is (−4, 19) and its radius (the square root of r^2) is 11. The domain of a circle, or set of all possible *x*-values within that circle's boundary, is represented by the inequality $h-r \leq x \leq h+r$, where *x* is the domain, *h* is the *x*-coordinate of the circle's center, and *r* is the circle's radius. For the given equation, the circle's domain is thus $-4-11 \leq x \leq -4+11$, or [−15, 7]. Choice B, −14, is the only offered value that lies within the domain bounded by −15 and 7 and thus the only possible value for *a* among the answer options. Alternatively, students could use a graphing calculator, such as the one built into Bluebook, to graph the equation of the circle, visually inspect where the circle exists in the *xy*-plane, and then identify the only possible value for *a* among the answer choices.

> I'm gonna write this down. OK. I feel like I need to substitute *a* and *b* into the equation. So it'd be $(a+4)^2+(b-19)^2=121$. I'm going to solve this algebraically. So I'm going to factor out the square, which means $(a+4)(a+4)+(b-19)(b-19)$, which is $a^2+4a+16+b^2-38b+361$. And then if we add that together, I guess. All right. No, scratch all of that. I'm gonna substitute each individual answer for *a* into the equation. So that would be $(-16+4)^2+(y-19)^2=121$. Which we put that into the calculator, $(-16+4)^2$ equals 144. Plus, what I did earlier: $b^2-38b+361$. Adding 144 and 361 gives 505. So $b^2-38b=505-121$. So $b^2-38b=384$. Dividing 384 by 38 gives approximately 10. So *b* is 10. Cool. Oh, so I'm just graphing this right now, and then I'm gonna look for a point. [*Graphs circle in Bluebook's built-in graphing calculator*] That's *a*, which is equal to the *x*-value of the answer choices. Now I need to find an *x*-value. So that could be this point. [*Clicks on points on circle*] It can't be 19 because the circle never goes that way that much, and [it] can't be 11 because [the] circle doesn't go that way that much. And so it's either –16 or –14, which—the furthest point on the circle is –15, so [it] can't be –16. So the answer is probably –14, which is B.
>
> *Participant M34*

Participant M34 answered this question correctly and demonstrated a single expected behavior, resulting in a PL of 1. She begins her algebraic solution path by substituting *a* and *b* into the given equation for *x* and *y*, respectively. Simplifying the left side of this equation doesn't lead to any conclusions, so she next tries substituting the answer choices for *x* in the given equation, which also leads to a dead end. Finally, M34 graphs the given equation using Bluebook's built-in graphing calculator and investigates possible values of *x* that lie on the resultant circle. On investigating the graph of the circle and drawing conclusions about not only what she should be looking for but also how this information connects to the answer choices (behavior 1), M34 is quickly able to arrive at the correct answer, choice B.

Math Question 4

Content Domain	Advanced Math
Skill/Knowledge Testing Point	Nonlinear Functions: Rewrite
Performance Score Band	7
Stimulus Subject Area	None
Question Format	MC
Expected Behaviors	1. Use the graph of an exponential function to determine a minimum value. 2. Demonstrate an understanding of key features of the graph of an exponential function. 3. Demonstrate an understanding that exponential functions don't have relative extrema.
M34 Performance Level	5

Which of the following functions has(have) a minimum value at -3?

I. $f(x) = -6(3)^x - 3$

II. $g(x) = -3(6)^x$

A) I only

B) II only

C) I and II

D) Neither I nor II

Question 4, a hard (PSB 7) multiple-choice Nonlinear Functions: Rewrite question outside of context, requires test takers to demonstrate an understanding of minimum value in relation to exponential functions. The correct answer is choice D. Exponential functions continuously increase or decrease and therefore don't have a minimum (or maximum) value. Test takers may simply recall and apply this characteristic, or they could graph both functions to visually make this observation.

> So this would be, like, on a graph typically, like on [function] I. Subtracting 3 is going down on the *y*-axis, which means if it's on graph, that'd be, like, a parabola starting at –3 and just going above that. So I

> say it's A, which is "I only," because the graph starting at –3 means it has a minimum value of –3. And graph number II does not have that.
>
> *Participant M34*

Participant M34 answered this question incorrectly and didn't demonstrate any expected behaviors, resulting in a PL of 5. While M34 does imply a familiarity with the concept of minimum value in her verbalization, she proceeds to incorrectly assess the information that "−3" from exponential function *f* provides by comparing it to what information it would provide were it a quadratic function (parabola)—that is, the *y*-coordinate of its vertex ("subtracting 3 is going down on the *y*-axis"; "starting at −3 and just going above that"). It's unclear whether M34 incorrectly believes the question's functions are quadratic (parabolas) or mistakenly thinks that the vertical shift of exponential functions represents the minimum value instead of the asymptote. Wherever this misconception might have come from, it also leads her to dismiss function *g* as having a minimum value at −3, presumably due to the absence of the vertical shift apparent in function *f*.

Supplementary Vignette: Participant M44

Participant M44 answered question 4 correctly and demonstrated two expected behaviors, resulting in a PL of 1. M44 was one of five participants who answered the question correctly and one of four participants who did so while also demonstrating at least one expected behavior.

> So I think what I'm gonna do is plug this into [Bluebook's built-in graphing calculator]. [*Graphs both functions*] And then it's saying, Which of the functions has a minimum value at $x = -3$? [*Graphs this vertical line*] So –3 is on the $g(x)$ graph, but I know that the minimum is the lowest value, I guess. So I think it's D, "neither I nor II."
>
> *Participant M44*

Participant M44 immediately graphs the given functions using Bluebook's built-in graphing calculator, indirectly demonstrating an understanding of possible key features of graphs of exponential functions (behavior 2). For function *f*, however, the participant graphs $f(x) = -6(x)^x - 3$ instead of $f(x) = -6(3)^x - 3$, which is an apparent typographical error. To gain a better perspective, M44 also graphs the vertical line $x = -3$, showcasing what's happening at this location on the graphs. His assertion "so −3 is on the *g*(*x*) graph" is only a product of his entry error for the graph of function *f*; by stating "I know that the minimum is the lowest value," he still demonstrates an understanding of the notion that a graph can be used to determine the minimum value of an exponential function (behavior 1). This understanding leads M44 to select the correct answer of "neither I nor II," choice D.

Math Question 5

Content Domain	Problem-Solving and Data Analysis
Skill/Knowledge Testing Point	Percentages
Performance Score Band	7
Stimulus Subject Area	None
Question Format	MC
Expected Behaviors	1. Convert percentages greater than 100 to decimals. 2. Write an equation to compute an increase to a quantity by a percentage greater than 100. 3. Solve a linear equation. 4. Logically eliminate multiple-choice distractors (incorrect answers) by size of numbers relative to given information and the question asked.
M34 Performance Level	4

The result of increasing the quantity *x* by 400% is 60. What is the value of *x*?

A) 12

B) 15

C) 240

D) 340

Question 5, a hard (PSB 7) multiple-choice Percentages question outside of context, requires test takers to demonstrate an understanding of a percentage increase greater than 100. The correct answer is choice A. Four hundred percent is equivalent to $\frac{400}{100}$, or 4. Therefore, increasing quantity *x* by 400% can be represented by the expression $x + 4x$, or $5x$. It's given that the result of increasing a certain quantity, *x*, by 400% is 60. Therefore, $5x = 60$, which when solved yields $x = 12$.

> So the result of increasing the quantity *x* by 400% is 60. This is something—what's that called? I can't remember what it's called. I might just put this in a calculator and troubleshoot it. The quantity *x*, that's the fixed amount. So increasing that 400% equals—oh, I'm gonna write that down actually: *x* times 400%. That would be, turning that into a decimal would be 4. So $x \times 4 = 60$. OK. $4x = 60$. So what is *x*? That 60 divided by 4 is 15, and 15 is an answer. So I say B.
>
> *Participant M34*

Participant M34 answered the question incorrectly but exhibited one expected behavior, resulting in a PL of 4. She demonstrates an ability to work with a percentage greater than 100 by successfully turning 400% into the "decimal" of 4 (behavior 1), but she makes the fundamental error of concluding that a 400% increase in the quantity *x* is 4, rather than 5, times the original value of *x* ("so $x \times 4 = 60$") and then proceeds to divide the result of increasing *x* by 400% (60) by 4, ending up with 15 as the value of *x* and the incorrect answer of choice B.

Question 5 is arguably the most difficult Math question in the study, which is not only reflected by its PSB of 7 (the scale's highest) but also by the fact that no participant answered it correctly. Choice B was extremely attractive, as seventeen of twenty-one participants ultimately selected this option. This is likely due to applying an intuitively appealing but incorrect set of assumptions: If a quantity increases by 400%, then that quantity is four times larger than its original value; therefore, because $4x = 60$, the original quantity, x, is 15.

Math Question 6

Content Domain	Advanced Math
Skill/Knowledge Testing Point	Nonlinear Functions: Make Connections
Performance Score Band	7
Stimulus Subject Area	None
Question Format	MC
Expected Behaviors	1. Make connections between the equation of a quadratic function and its *x*-intercepts. 2. Rewrite a quadratic equation in a form that facilitates identifying unknown values. 3. Given certain pieces of information, recognize characteristics of the unknown values of a quadratic function.
M34 Performance Level	5

The function *f* is defined by $f(x) = ax^2 + bx + c$, where a, b, and c are constants. The graph of $y = f(x)$ in the xy-plane passes through the points $(7, 0)$ and $(-3, 0)$. If a is an integer greater than 1, which of the following could be the value of $a + b$?

A) -6

B) -3

C) 4

D) 5

Question 6, a hard (PSB 7) multiple-choice Nonlinear Functions: Make Connections question outside of context, requires test takers to draw connections between a quadratic function with unknown constants and its two given *x*-intercepts. Test takers must also be capable of handling a fair amount of algebraic computation as well as understand the significance of an unknown constant being called out as a specific type of number. The correct answer is choice A. It's given that function *f* passes through the points (7, 0) and (−3, 0). Substituting 7 for *x* and 0 for $f(x)$ and also −3 for *x* and 0 for $f(x)$ in the function $f(x) = ax^2 + bx + c$ yields the equations $49a + 7b + c = 0$ and $9a - 3b + c = 0$. It follows that $49a + 7b = 9a - 3b$. Combining like terms in this equation gives $40a = -10b$, or $-4a = b$. To find $a + b$, substituting $-4a$ for b gives $a - 4a$, or $-3a$. So $a + b$ is equivalent to $-3a$, which is a multiple of −3. Since it's given that a is an integer greater than 1, when a is 2, then $a + b = -3a = -3(2) = -6$.

> The function is $ax^2 + bx + c$. OK. So it passes through the point (7, 0). [*Identifies points displayed by Bluebook's built-in graphing calculator*] So

> that's right here. (7, 0) and (–3, 0) over here. I need to substitute 7 into the equation where the *x*'s are. So that'd be $a(7)^2 + b(7) + c$, which is $49a + 7b + c$. Oh, I can't do that because of *a* and *b*. OK. I'm trying to imagine what the graph looks like right now. There's a curve somehow. There's a way to find the slope, but that doesn't matter. OK, so *c* doesn't matter. I'm just going to go with $ax^2 + bx$ and if *a* is greater than 1. Say, *a* is equal to 2. So $2(7)^2 + b(7)$ is $2(49)$ is $98 + 7b$. Let's just divide it: $\frac{98}{7} = 14$. That ain't right. But I think I'm on the right track, which means if *a* is greater than 1, it can't be rotated over the *x*-axis, which means it could still be like this. I just don't know what to do. I feel like I need to graph this somehow. [*Graphs function in calculator, with* x *equal to 7*] It's being stretched by something because it's bigger than 1. I don't think it's A or B. So I'm going to make an educated guess and say it would be C.
>
> *Participant M34*

Participant M34 answered the question incorrectly and didn't exhibit any expected behaviors, resulting in a PL of 5. It seems that the amount of information provided in the question didn't allow M34 to find a clear entry point. M34 makes multiple attempts to solve the problem, some leading to additional misinterpretations and eventually the conclusion that the correct answer depends on the "need to graph this somehow." Ultimately, she makes an "educated guess," albeit an incorrect one, but still shows perseverance given the question's level of difficulty (PSB 7).

This question proved challenging for other participants in the study as well, as only three answered correctly and none exhibited any expected behaviors. Indeed, comments from other study participants echoed those of participants M32 and M10, who emphasized that the amount of both given and unknown information in the question proved overwhelming.

> I don't really understand this question at all. So that's just my best guess, which is A.
>
> *Participant M32*

> I'm not 100 percent sure about [my answer] 'cause I can't use all of the given information at the moment. And I don't really see anything else that I can use in this problem that helps me find anything more about what I can do.
>
> *Participant M10*

Math Question 7

Content Domain	Algebra
Skill/Knowledge Testing Point	Linear Functions: Identify
Performance Score Band	2
Stimulus Subject Area	Science
Question Format	MC
Expected Behaviors	1. Read and demonstrate comprehension of the context described. 2. Set up/identify a linear equation or inequality as described in the context.
M34 Performance Level	1

A veterinarian recommends that each day a certain rabbit should eat 25 calories per pound of the rabbit's weight, plus an additional 11 calories. Which equation represents this situation, where c is the total number of calories the veterinarian recommends the rabbit should eat each day if the rabbit's weight is x pounds?

A) $c = 25x$

B) $c = 36x$

C) $c = 11x + 25$

D) $c = 25x + 11$

Question 7, an easy (PSB 2) multiple-choice Linear Functions: Identify question set in a science context, requires test takers to identify a linear equation in two variables that represents the given context. The correct answer is choice D. It's given that a veterinarian recommends that each day a certain rabbit eat 25 calories per pound of the rabbit's weight, plus an additional 11 calories. If the rabbit's weight is x pounds, then the total number of calories, c, can be written as $c = 25x + 11$.

> OK. Each day a certain rabbit eats 25 calories per pound of the rabbit's weight plus 11 calories already. So already we know it's gotta be plus-11 no matter what; that's a constant. And then equals *c*, but all the answers equal *c*, and the rabbit's weight is *x*. So that's 25 per pound. So 25 times 1. I think the answer is D because 25*x* is 25 times the rabbit's weight, which is [in] pounds. So it would be 25 [calories; *participant says "pounds"*] times each pound plus 11.
>
> *Participant M34*

Participant M34 answered the question correctly and demonstrated both expected behaviors, resulting in a PL of 1. After reading the context, M34 demonstrates comprehension (behavior 1) by observing that "already we know it's gotta be plus-11 no matter what" given the fixed number of additional calories per day mentioned in the question. Then M34 discusses what the variables in the linear equation describing the context represent—a step often overlooked by students. She concludes that choice D is correct because the answer "would be 25 [calories] times each pound plus 11" (behavior 2).

Question 7 was the easiest Math question in the study, which is not only reflected by its PSB of 2 but also by the fact that eighteen of twenty-one students answered this question correctly, with seventeen of these demonstrating at least one expected behavior.

Math Question 8

Content Domain	Geometry and Trigonometry
Skill/Knowledge Testing Point	Measure of Angles in a Triangle
Performance Score Band	3
Stimulus Subject Area	None
Question Format	MC
Expected Behaviors	1. Demonstrate an understanding of the triangle sum theorem. 2. Use logic to determine the maximum value of an angle in a triangle given the measure of one of the other angles.
M34 Performance Level	1

In $\triangle RST$, the measure of $\angle R$ is 63°. Which of the following could be the measure, in degrees, of $\angle S$?

A) 116

B) 118

C) 126

D) 180

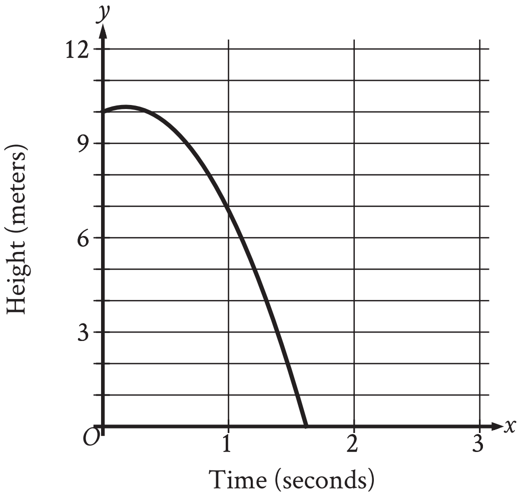
Question 8, an easy (PSB 3) multiple-choice Measure of Angles in a Triangle question outside of context, requires test takers to demonstrate an understanding of the triangle sum theorem, the concept that the sum of all interior angles of a triangle is 180°. The correct answer is choice A. For $\triangle RST$, it's given that the measure of $\angle R$ is 63°. Therefore, by the triangle sum theorem, the sum of the measures of $\angle S$ and $\angle T$ is $(180 - 63)^\circ$, or 117°. This means that the measure of $\angle S$ must be less than 117°. Of the given answer options, only choice A, 116, is less than 117 and therefore could be the measure, in degrees, of $\angle S$.

> I'll draw this. So $\triangle RST$, the measure of $\angle R$ is 63 [°]. What could be the measure, in degrees, of $\angle S$? We need to know. Except I don't know what type of triangle this is, but it also equals 180, which means it can't be answer D because that's already 180 [°]. So in an equation, this would look like $63 + S + T = 180$. I'll go ahead and subtract 63 from 180, which is 117. Drawing this out, that would look like 63, and then somehow the other two angles have to make up to 117, which means 63. But wait. OK, I'm now going to actually troubleshoot and just do 63 + 116 [*choice A*]. 179 could be. It can't be 100. OK. It is 116 because 63 + 118 [*choice B*] is 181, and 63 + 126 [*choice C*] is 189. So all those would be bigger than the amount of degrees a triangle can be. So the answer is A.
>
> *Participant M34*

Participant M34 answered the question correctly and demonstrated both expected behaviors, resulting in a PL of 1. After drawing a triangle, M34 states she "[doesn't] know what type of triangle this is," an irrelevant point she justifiably dismisses as she transitions to applying the triangle sum theorem (behavior 1), which describes triangles universally. Using this theorem allows M34 to rule out choice D, an impossible answer because it represents the sum of the measures of all interior angles of a triangle rather than a possible measure of ∠*S*. M34 then sets up an equation showing that, per the triangle sum theorem, the sum of the measures of ∠*R*, ∠*S*, and ∠*T* is 180, leading her to conclude that "the other two angles have to make up to 117" since the given measure of ∠*R* is 63°. From here, M34 "troubleshoot[s]" the remaining choices—inserts the answers into the equation she created—and logically determines that only choice A, 116, could be a valid measure of ∠*S*, in degrees, given that choices B (118) and C (126) result in sums greater than 180° when added to 63° (behavior 2).

Math Question 9

Content Domain	Advanced Math
Skill/Knowledge Testing Point	Nonlinear Functions: Interpret
Performance Score Band	4
Stimulus Subject Area	Science
Question Format	MC
Expected Behaviors	1. Read and demonstrate comprehension of the context described. 2. Identify the *x*-intercept of a graph of a quadratic function. 3. Interpret the context of an *x*-intercept of the graph of a quadratic function.
M34 Performance Level	1



A competitive diver dives from a platform into the water. The graph shown gives the height above the water *y*, in meters, of the diver *x* seconds after diving from the platform. What is the best interpretation of the *x*-intercept of the graph?

A) The diver reaches a maximum height above the water at 1.6 seconds.

B) The diver hits the water at 1.6 seconds.

C) The diver reaches a maximum height above the water at 0.2 seconds.

D) The diver hits the water at 0.2 seconds.

Question 9, a medium-difficulty (PSB 4) multiple-choice Nonlinear Functions: Interpret question set in a science context, requires test takers to interpret a key feature of the graph of a quadratic function in terms of the context. The correct answer is choice B. The *x*-intercept of a graph is the point at which a graph intersects the *x*-axis, which, in the given graph, represents time, in seconds. The given graph intersects the *x*-axis between $x = 1$ and $x = 2$. In context, this means that the diver hits the water (reaches 0 on the *y*-axis, which represents height, in meters, above the water) between 1 and 2 seconds after diving from the platform, making choice B the best interpretation of the graph's *x*-intercept.

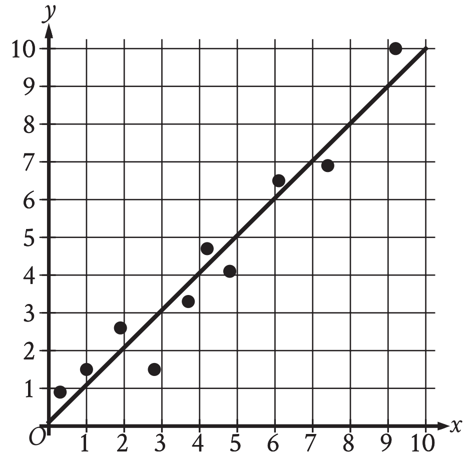
> So the height shows how high the diver got *x* seconds after diving from a platform. I say B because the timer stops, which is OK. I know it's this answer, but it's because it can't be the height; the *x*-value is at 1.6, and it can't be height because that's where the biggest height is behind 1.6 seconds, according to the *y*-value. So the answer is B.
>
> *Participant M34*

Participant M34 answered this question correctly and demonstrated all expected behaviors, resulting in a PL of 1. M34 demonstrates understanding of the context (behavior 1) by describing the *x*-axis as representing "seconds after diving from a platform" and implying that the *y*-axis represents height, in meters, above the water. She seems to immediately rule out choices A and C on the grounds that each misinterprets the meaning of the *x*-axis and *x*-intercept ("it can't be the height"). She correctly identifies the value of the graph's *x*-intercept (behavior 2) as 1.6 and the *x*-intercept's meaning in context (behavior 3) as representing 1.6 seconds after the diver jumped from the platform.

Math Question 10

Content Domain	Problem-Solving and Data Analysis
Skill/Knowledge Testing Point	Scatterplot
Performance Score Band	4
Stimulus Subject Area	None
Question Format	MC
Expected Behaviors	1. Understand that the data points in a scatterplot represent actual values and that the line of best fit represents predicted values. 2. Understand that for the actual *y*-values in a scatterplot to be greater than the predicted *y*-values, the data points will have to be above the line of best fit.
M34 Performance Level	1

The scatterplot shows the relationship between two variables, *x* and *y*. A line of best fit for the data is also shown.



For how many of the 10 data points is the actual *y*-value greater than the *y*-value predicted by the line of best fit?

A) 3

B) 4

C) 6

D) 7

Question 10, a medium-difficulty (PSB 4) multiple-choice Scatterplot question outside of context, requires test takers to understand what a line of best fit represents in a scatterplot. The correct answer is choice C. In conceptual terms, any data point located above a scatterplot's line of best fit has a *y*-value greater than that predicted by the line. For the given scatterplot, six of the data points are positioned above the line of best fit.

> So we need to find [for] how many of the 10 data points is the actual *y*-value greater than the *y*-value predicted. So we need to compare *y*-values to how many points of their *y* are greater than the line, which would be on the left side of the graph, the left side of the line. So now I'm going to count those points. That's 1, 2, 3, 4, 5, 6. So the answer is C.
>
> *Participant M34*

Participant M34 answered the question correctly and demonstrated both expected behaviors, resulting in a PL of 1. She first exhibits an understanding of the relationship between the data points and the line of best fit (behavior 1): "So we need to compare *y*-values to how many points of their *y* are greater than the line." She then correctly observes that the data points in question will be those above the line of best fit (behavior 2): ". . . which would be on the left side of the graph, the left side of the line." In context, it's clear that M34's reference to the "left" side of the line is to data points above the line of best fit. From there, it's a simple matter of M34 counting the number of points above that line to determine the correct answer.

Math Question 11

Content Domain	Problem-Solving and Data Analysis
Skill/Knowledge Testing Point	Probability
Performance Score Band	4
Stimulus Subject Area	Real-world topics
Question Format	MC
Expected Behaviors	1. Calculate, express, or interpret the probability of an event. 2. Apply the understanding that the sum of probabilities of all possible outcomes of an event is 1. 3. Determine an unknown number using probability and the context described.
M34 Performance Level	1

At a movie theater, there are a total of 350 customers. Each customer is located in either theater A, theater B, or theater C. If one of these customers is selected at random, the probability of selecting a customer who is located in theater A is 0.48, and the probability of selecting a customer who is located in theater B is 0.24. How many customers are located in theater C?

A) 28

B) 40

C) 84

D) 98

Question 11, a medium-difficulty (PSB 4) multiple-choice Probability question set in a real-world context, requires test takers to determine an unknown quantity using probability and given information. The correct answer is choice D. Per the context, each of 350 customers is located in one of three theaters, A, B, or C. It's further given that the probability of randomly selecting a customer located in theater A is 0.48 and that the probability of randomly selecting a customer located in theater B is 0.24. Therefore, the probability of randomly selecting a customer located in either theater A or theater B is $0.48 + 0.24$, or 0.72. As the sum of probabilities of all possible outcomes of an event is 1, it follows that the probability of randomly selecting a customer located in theater C is $1 - 0.72$, or 0.28. This means there are $(0.28)(350)$, or 98, customers located in theater C.

> So there's a total of 350. I'm going to write this out. If there's a total of 350, I'm also going to add the probabilities, which is $0.46 + 0.24 = 0.70$. So $1 - 0.70 = 0.30$. So that means theater C has 30 percent of the total number of customers. So $350 \times 0.30 = 105$. So the answer is D because it's in the middle.
>
> *Participant M34*

Participant M34 answered the question correctly and demonstrated two expected behaviors, resulting in a PL of 1. M34 appears to have a solid understanding of how to apply probability in a context but makes a simple reading mistake, leading her to select the correct answer by questionable logic. When M34 attempts to add the given probabilities, 0.48 and 0.24, she incorrectly writes 0.46 instead of 0.48, resulting in a sum of 0.70, and uses this number for the remainder of her solution path. By appropriately making use of the principle of complementary events (behavior 2), she determines that 0.30 (the result of $1-0.70$) times the total number of customers should lead to the key. Since $(0.30)(350)=105$ (behavior 3) and given that 105 isn't an answer choice, she takes a leap and concludes that choice D is the answer "because it's in the middle." It's unclear what she means by that statement, but given that her incorrect calculation isn't off by very much, she still is able to choose the closest option to the answer she comes up with, which happens to be correct. In other respects, M34 demonstrates great proficiency in the tested skill.

Math Question 12

Content Domain	Advanced Math
Skill/Knowledge Testing Point	Nonlinear Equations: Solve
Performance Score Band	5
Stimulus Subject Area	None
Question Format	SPR
Expected Behaviors	1. Set a quadratic equation equal to zero. 2. Apply an understanding of the zero-product property. 3. Solve a quadratic equation algebraically. 4. Solve a quadratic equation graphically.
M34 Performance Level	1

$$(d-30)(d+30)-7=-7$$

What is a solution to the given equation?

Question 12, a medium-difficulty (PSB 5) student-produced response Nonlinear Equations: Solve question outside of context, requires test takers to solve a quadratic equation, which in this case yields two distinct solutions. Correct answers are −30 and 30, though (as indicated by "a solution" as well as the overall test section directions) test takers are expected (and allowed) only to supply one such correct answer. To solve this equation algebraically, students could add 7 to both sides of the given equation. This gives $(d-30)(d+30)=0$. The zero-product property states that a product of two factors is equal to 0 if and only if at least one of the factors is 0. Therefore, $d-30=0$ or $d+30=0$. It follows that $d=30$ or $d=-30$. Another reasonable algebraic approach would be to multiply the binomials and combine like terms, resulting in the equation $d^2=900$. Applying the square root property, which states that if $x^2=c$, then $x=\pm\sqrt{c}$, to this equation gives $d=30$ or $d=-30$. This quadratic equation could also be solved graphically by entering the given equation into a graphing calculator (using *x* instead of *d*) and applying the understanding that the two vertical lines produced represent the distinct solutions to the equation.

> So what is the solution to the given equation? The equation is $(d-30)(d+30)-7=-7$. So I'm gonna write this out and solve this algebraically. So I need to factor the first two parentheses, which means $d^2+30d-30d+30^2-7=-7$. Then, $+30d$ and $-30d$ cancel out. So it's $d^2-900-7=-7$. So $d^2-907=-7$. Then $+907$; $d^2=900$ because $-7+907$. Then I'm gonna square-root both sides, which means $d=\sqrt{900}$, which is 30. So $d=30$.
>
> *Participant M34*

Participant M34 answered the question correctly and demonstrated a single expected behavior, resulting in a PL of 1. After writing out the given equation, M34 decides to "solve this algebraically." Her approach might not be the most efficient but proves to be effective. Taking an already factored quadratic equation and expanding it doesn't allow for demonstration of the zero-product property, but M34 confidently solves this equation algebraically (behavior 3) by using the square root property. After getting $d^2=900$, she states she's "gonna square-root both sides" but doesn't in the process acknowledge that this action would generate both a positive and negative solution—in this case, $d=\pm 30$. However, as this question only requires a single solution, this omission doesn't impede M34 from correctly answering with 30.

Math Question 13

Content Domain	Algebra
Skill/Knowledge Testing Point	Linear Equations in Two Variables: Make Connections
Performance Score Band	5
Stimulus Subject Area	None
Question Format	SPR
Expected Behaviors	1. Rewrite a linear equation into an appropriate form to identify the slope of a graph. 2. Perform numerical calculations involving fractions and/or decimals. 3. Calculate the slope of a graph from two points on the graph.
M34 Performance Level	4

What is the slope of the graph of $y=\frac{1}{3}(29x+10)+5x$ in the *xy*-plane?

Question 13, a medium-difficulty (PSB 5) student-produced response Linear Equations in Two Variables: Make Connections question outside of context, requires test takers to determine the slope of the graph of a line given the equation for that line. The correct answer is $\frac{44}{3}$. A linear equation can be written in the form $y=mx+b$, where m is the slope of the graph of the line. To rewrite the given equation in this form, students could distribute the $\frac{1}{3}$ to the grouped binomial, which gives $y=\frac{29}{3}x+\frac{10}{3}+5x$. Combining like terms gives $y=\frac{44}{3}x+\frac{10}{3}$. Therefore, the slope is $\frac{44}{3}$. In Bluebook, students can validly enter this answer fractionally as

44/3 or as the decimals 14.66 or 14.67. (Either of these decimal answers would be acceptable, as the instructions provided for SPR questions state “If your answer is a **decimal** that doesn’t fit in the provided space, enter it by truncating or rounding at the fourth digit.”)

> The slope is, I could say it’s just $\frac{1}{3}$, but it looks like I should probably solve this algebraically. So I’m gonna write it out, which is $y = \frac{1}{3}(29x + 10) + 5x$. Which means—or I’ll just graph this, see what it looks like. [*Graphs line using Bluebook’s built-in graphing calculator*] Yeah, I need to solve this algebraically. The slope is, like, $y_1 - y_2$ divided by $x_1 - x_2$, which is—To write that out would be—Oh, no. OK. I need to distribute $\frac{1}{3}$ first, which is $29[x] \times \frac{1}{3} = 9.67[x]$, plus $\frac{1}{3} \times 10 = 3.33$, plus $5x$ equals $y = 14.67x + 3.33$. Oh, this isn’t right. OK. I’m just gonna say it’s $\frac{1}{3}$ because, and just looking at the graph, that’s what the slope is currently.
>
> *Participant M34*

Participant M34 answered the question incorrectly but exhibited two expected behaviors, resulting in a PL of 4. M34’s statement “I could say it’s just $\frac{1}{3}$” suggests she has some understanding of the form of an equation of a line in which the slope is the leading coefficient (behavior 1). After initially deciding against $\frac{1}{3}$ as the answer, she proceeds with trying to compute the graph’s slope algebraically using the slope formula $m = \frac{y_2 - y_1}{x_2 - x_1}$ and two points found from graphing the line in Bluebook’s built-in graphing calculator. Given the complexity of the given equation, however, finding suitable points on the graph proves daunting, and she pivots to rewriting the equation. She correctly navigates the fractions/decimals (behavior 2) and successfully writes the equation in the slope-intercept form of $y = mx + b$ but fails to recognize that the slope is the leading coefficient, *m*, perhaps as a result of the uncommon slope value of 14.67. She concludes that “this isn’t right” and then returns to her initial guess of $\frac{1}{3}$. In any event, M34 shows partial enactment of the question type’s construct.

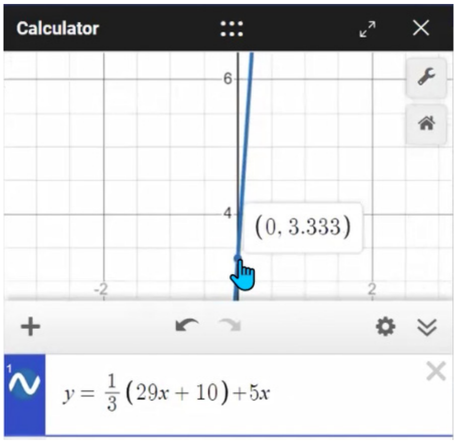
Supplementary Vignette: Participant M38

Successfully using the graphing solution path abandoned by M34, participant M38 answered question 13 correctly and demonstrated a single expected behavior, resulting in a PL of 1. M38 was one of four participants who answered the question correctly, all of whom also exhibited at least one expected behavior.

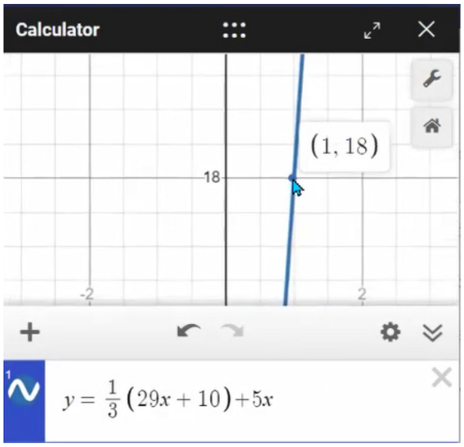
> OK, let’s graph it. [*Graphs line using Bluebook’s built-in graphing calculator*] The slope form would be rise over run. So $\left(0, \frac{1}{3}\right)$. Right now, this is $(1, 18)$, so it went up. $18 - \frac{1}{3}$ is $17\frac{2}{3}$, is what we want up. So the slope is $17\frac{2}{3}$. So, yeah, 17.66. [*Long pause; moderator interjects to ask participant to share thoughts*] So it wants me to find the slope. So I graphed it. So, to find slope, it’s rise over run of two points. So if you find the first, your solid point is 0. Did I do that wrong? Yes, I did. So there is a $3\frac{1}{3}$. So you go up to a second point, which is 118. So I went up 18.67.

> But I, so that's what I rose, and then I ran over 1 to get there. So it's 14.66, not 17.66.
>
> *Participant M38*

Unfazed by the challenging coordinates, M38 successfully approaches this question by calculating the graph's slope from two points on the graph (behavior 3). M38 identifies the following points on the graphed line using Bluebook's built-in graphing calculator:



$y = \frac{1}{3}(29x + 10) + 5x$



$y = \frac{1}{3}(29x + 10) + 5x$

At the outset, M38 mistakenly identifies the first point of this line as $\left(0, \frac{1}{3}\right)$ instead of $\left(0, 3\frac{1}{3}\right)$. Noting that slope can be formulated as "rise over run," M38 gets an answer of 17.66. After entering this as his answer, M38 pauses, seemingly unsure. At this point, the moderator—following the protocol, which allows for nondirective prompts when students lapse into extended silence—interjects and asks the student to share his thoughts. M38 takes the opportunity to retrace his steps and discovers his initial mistake. Once again, he makes mention of how he calculates slope here: "So I went up 18.67. But I, so that's what I rose, and then I ran over 1 to get there." This leads M38 to the correct answer: "So it's 14.66, not 17.66."

Math Question 14

Content Domain	Geometry and Trigonometry
Skill/Knowledge Testing Point	Scale Factor and Area
Performance Score Band	6
Stimulus Subject Area	None
Question Format	MC
Expected Behaviors	1. Apply an understanding of how applying scale factor to side lengths affects the areas of similar rectangles. 2. Calculate the area of similar rectangles using two possible side lengths. 3. Logically eliminate multiple-choice distractors (incorrect answers) by size of numbers relative to given information and the question asked.
M34 Performance Level	5

Rectangles $ABCD$ and $EFGH$ are similar. The length of each side of $EFGH$ is 6 times the length of the corresponding side of $ABCD$. The area of $ABCD$ is 54 square units. What is the area, in square units, of $EFGH$?

A) 9

B) 36

C) 324

D) 1,944

Question 14, a hard (PSB 6) multiple-choice Scale Factor and Area question outside of context, requires test takers to understand how a scale factor applied to one rectangle affects the area of a similar rectangle. The correct answer is choice D. If x represents the length, in units, of the base of rectangle *ABCD* and y represents its height, in units, then the area of rectangle *ABCD* is xy square units. It's given that each side of similar rectangle *EFGH* is 6 times the length of the corresponding side of rectangle *ABCD*. Therefore, $6x$ represents the length, in units, of the base of rectangle *EFGH*, $6y$ represents its height, in units, and $(6x)(6y)$, or $36xy$, square units represents its area. It's also given that the area of rectangle *ABCD* is 54 square units; therefore, $xy = 54$. Substituting 54 for xy in the expression $36xy$ yields $(36)(54)$, or 1,944, square units as the area of rectangle *EFGH*.

> So I'm gonna write this out. Square *ABCD* and—no, rectangle *ABCD* and rectangle *EFGH*. The length of each side [of rectangle *EFGH*] is 6 times the length [of the corresponding side of rectangle *ABCD*]. So each side is six times the length. Um, so if we look at the reference sheet [available in Bluebook]—yeah 54 would just be 54×6, I'm gonna assume. So that'd be $54 \times 6 = 324$. Which, that is an answer. So my final answer is C.
>
> *Participant M34*

Participant M34 answered the question incorrectly and didn't exhibit any expected behaviors, resulting in a PL of 5. M34 succumbs to a very common mistake for this type of question. Even though she cross-checks the reference sheet available in Bluebook during the activity (and actual testing), which gives the formula $A = lw$ for the area of a rectangle, she fails to appropriately apply the scale factor, instead multiplying the area of rectangle *ABCD* (54 square units) by 6. Choice C, 324 square units, is included as an answer option in this question because it represents a very common conceptual error for this sort of problem.

Supplementary Vignettes: Participants M32 and M56

Participant M32 answered question 14 correctly and demonstrated two expected behaviors, resulting in a PL of 1. M32 was one of five participants who answered the question correctly, all of whom demonstrated at least one expected behavior.

> The length of each side of [rectangle] *EFGH* is 6 times the length of the corresponding side of [rectangle] *ABCD*. Since the area of *ABCD* is 54 square units, the area of *EFGH* would be 54 multiplied by the square of 6. This is because the area scales with the square of the side length ratio for

> similar figures. Therefore, it's 54×6^2. 6^2 is 36, so $54 \times 36 = 1{,}944$. The answer is D, 1,944.
>
> *Participant M32*

M32 demonstrates clear command of how scale factor affects the area of similar figures. His approach is clear and concise, making note of the reason the given area of rectangle *ABCD* should be multiplied by 6^2, or 36, to obtain the area of similar rectangle *EFGH* (behaviors 1 and 2): "This is because the area scales with the square of the side length ratio for similar figures."

Participant M56, who also attained PL 1 on this question, is able to show how using both an understanding of the concept of scale factors (behavior 1) and mathematical reasoning (behavior 3) can lead to efficient solving.

> So, thinking about the sides of a rectangle, I imagine *EFGH*. There are really only four combinations you can do. So, 54 times 4 on my calculator is 216. 216 times 6 is 1,296. Now, I know that if I look at it, it can't be 9 [*choice A*], and it can't be 36 [*choice B*]. So I can automatically rule those out because those would be lower than what we had. Thinking about how many sides it would have and how many times I would have to multiply each side, I decide that 324 is probably too low. So I go with D, 1,944.
>
> *Participant M56*

It's given that the area of rectangle *ABCD* is 54 square units and that the side lengths of similar rectangle *EFGH* are longer than the corresponding ones of rectangle *ABCD*. Therefore, the resulting area of rectangle *EFGH* must be greater than 54 square units, which makes choices A (9) and B (36) impossible answer options.

Math Question 15

Content Domain	Algebra
Skill/Knowledge Testing Point	Systems of Two Linear Equations in Two Variables: Solve
Performance Score Band	6
Stimulus Subject Area	None
Question Format	SPR
Expected Behaviors	1. Fluently eliminate a variable in a system of two linear equations. 2. Identify the solution to a linear system from its graph. 3. Solve for a multiple of the value of *x*.
M34 Performance Level	4

$$5y = 10x + 11$$
$$-5y = 5x - 21$$

The solution to the given system of equations is (x, y). What is the value of $30x$?

Question 15, a hard (PSB 6) student-produced response Systems of Two Linear Equations in Two Variables: Solve question outside of context, requires test takers to work with a system of two linear equations in determining a multiple of the value of *x*. The correct answer is 20. Adding the two equations in the system gives $0 = 15x - 10$. Adding 10 to both sides of this equation yields $15x = 10$. The value of $30x$ can be found by multiplying both sides of this equation by 2. Therefore, $30x = 20$.

> So the solution to the given system of equations is (x, y). The equations are $5y = 10x + 11$ and $-5y = 5x - 21$. So I'm gonna write this out. OK. So this means I could substitute or cancel out $5y$ and $-5y$ because they're like terms. $5y - 5y$ leaves nothing, which means 0. No, I need to turn $-5y$ into a positive or—no. $5y = 10x + 11$. And then I'm gonna multiply $-5y = 5x - 21$ by -1. Oh, that doesn't work. OK, substitution. I can't use substitution right now because it wouldn't be equal or I'd get decimals. I'll just suck it up. So, I'm gonna divide by 5 to get *y* on its own, which means $y = 2x + 2.2$. And so now that I can substitute into $-5y$, which is what *y* equals. So $-5(2x + 2.2) = 5x - 21$, which, if I distribute the -5, is $-10x - 11$. So $-10x - 11 = 5x - 21$. If I subtract $5x$ and add 11, that's—oh no, no, no. The other way around. Wait. $-10x - 11 = 5x - 21$. I'm gonna add 21, so that's $21 - 11$, which equals 10. So 10. Then I'm gonna add $10x$ to the other side. So $10 = 15x$. To get *x*, divide by 15. So $\frac{10}{15} = 0.67$. So $x = 0.67$. Then the value of $30 \times x$, and if $x = 0.67$, that's 30×0.67, which is 20.1. So my answer is—which? Oh, we can—no, I'm not gonna do that. OK. My answer is 20.1.
>
> *Participant M34*

Participant M34 answered the question incorrectly but demonstrated two expected behaviors, resulting in a PL of 4. M34 begins by attempting to add the equations in the system but has concerns with $5y - 5y$ resulting in 0. From there, she reluctantly attempts a substitution approach ("I'll just suck it up") that leads to decimals. Although the numbers aren't particularly easy to work with, M34 pushes through and achieves the desired variable elimination, $10 = 15x$ (behavior 1). At this point, M34 chooses to solve for x and in so doing concludes that $x = \frac{10}{15} = 0.67$, an inexact, rounded value. She moves forward with this approximation, multiplying her derived value of *x*, 0.67, by 30 to get 20.1 (behavior 3). Given that this question is in the student-produced response format with a keyed response of 20, M34's approximation of 20.1 would be counted as incorrect. Nevertheless, M34 shows partial enactment of the question type's construct.

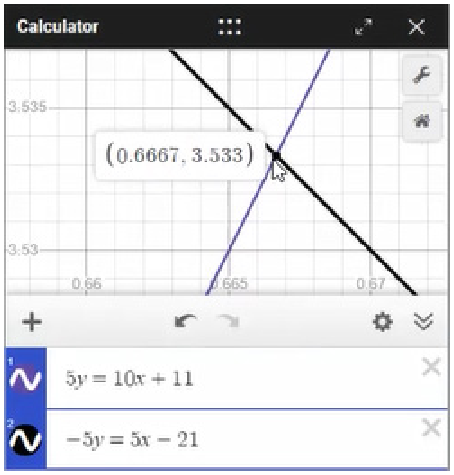
Supplementary Vignette: Participant M33

Taking a graphical approach in contrast to M34's algebraic one, participant M33 answered question 15 correctly and demonstrated two expected behaviors, resulting in a PL of 1. M33 was one of three participants who answered the question correctly, all of whom also demonstrated at least one expected behavior.

> So the [solution to the] given system of equations is (x, y). What is the value of $30x$? OK. I think I know what it's getting at. It's saying that these

> two are lines, I think. So I'm gonna plug these in as lines. [*Graphs lines using Bluebook's built-in graphing calculator*] Thank God for Desmos. Let's see where they intersect. They intersect in a weird spot. Of course they do. Why they have to be up in a weird spot? So what are you gonna get, 0.66 and that? OK. So I'm not entirely sure about this because it says the solution to the given—I'll check [the] reference [sheet available in Bluebook]. No, it's all geometry. OK. Solution to the given system of equations is (x, y). What is the value of $30x$? So I have a couple things are going through my head right now. Number one is that I don't know what, like, I'm not necessarily familiar with what this equation is, like, what do you mean "solution"? So I'm not 100 percent sure what that is, but the other thing is that $30x$. Part of me wants to immediately think that $30x$ is asking, you know, what's the y-value if x is 30? But I don't think it wants that because not only did it not ask me for that, but it would have phrased it as x equals 30, not $30x$. So I can only guess that what it means by $30x$ is it's 30 times the x-value of the coordinate. And it's saying that the solution is a coordinate. I can't think of any coordinate off the top of my head because there's no reason why this line's y-intercept should be more important than this line's y-intercept, or same thing with the x-axis. There's no reason why one should be more important than the other. Which makes me think that the only important thing is where they intersect. OK. Well, so looks like where they intercept is $(0.6667, 3.533)$. I think what it wants, and I'm not 100 percent sure, is 30 times the x-value of this. The x-value being 0.6667. So, 0.6667 times 30 is 20. That's trippy. OK. So 0.6667 is just two-thirds, and two-thirds of 30 is 20. That's what I'm a bit tripped up by is because it says 20.001. I don't know if—I mean, technically, yeah, it would add up. I'm just thinking if they do they want me to just put 20, or do they want me to just put 20.001? My assumption would be that they want me to stick with the formula. Yeah, so, I'm prompted to think that this is 20.001. I can't type that many [digits into the student-produced response answer field]. Ah-ha! Then it must be 20. Yeah, so because I don't really know 100 percent what this kind of problem is. But it's a solution to the given system, and if there's two of them, I would think it would be where they intersect or where they meet. Which was that coordinate of $(0.6667, 3.533)$. It wants to know what's the value of $30x$, which means 30 times that x-value. So if we run 30 times the x-value, which was 0.6667, times 30, it gives us 20.001, and I tried 20.001 and [Bluebook] didn't want to take it. So I guess it just means—oh. [*Consults SPR entry directions*] [SPR answers may be] truncat[ed] at the fourth digit. OK. I'm fine with that. So, yeah, I'm thinking it's gonna be 20 for this.
>
> *Participant M33*

After graphing the system of equations using Bluebook's built-in graphing calculator, M33, like participant M34, calls attention to the less-than-congenial noninteger coordinates for the point of intersection but pushes through with a successful graphical solution approach (behavior 2).



$5y = 10x + 11$

$-5y = 5x - 21$

He seems a bit uncertain about what the question is asking but eventually concludes that "what it means by 30*x* is it's 30 times the *x*-value of the coordinate." After multiplying the approximate *x*-coordinate of the point of intersection, 0.6667, by 30 (behavior 3) and getting 20.001, M33 acknowledges that "0.6667 is just two-thirds, and two-thirds of 30 is 20" but still wavers on which answer to input, 20.001 or 20. When he tries to enter "20.001," he isn't able to because Bluebook allows only five characters (in addition to a negative sign) for responses to SPR questions. As the SPR directions state that test takers should enter decimal answers rounded or truncated at the fourth digit, M33 concludes that the correct answer is 20.

PARTICIPANT PERCEPTIONS

Following the think-aloud activity, Math participants, like their Reading and Writing counterparts, were asked a standardized set of six follow-up questions. An analysis of participants' responses to each of the questions follows.

General Impressions

1. Please tell me a bit about the experience you just had. What was it like to answer those questions?

Postexperience question 1 elicited considerable variation in participant reactions, with many calling out the unfamiliarity of verbalizing their mathematical thinking processes. Participants frequently mentioned challenges with the mathematical content they encountered, attributing difficulties to gaps in recent math coursework or to encountering concepts they hadn't yet learned. A notable observation was the high reading demand of the Math problems, which participants identified as adding complexity. Emotive responses to the experience ranged from negative ("scary") to neutral ("normal") to positive ("fun"), with several participants drawing comparisons to their previous standardized testing experiences.

> It was definitely a fun challenge. I wasn't really 100 percent sure what to expect when I joined, and it was sort of enjoyable, you know? I often struggle a lot with my math, and I'm still just finishing up Algebra I right now. So it was really cool to get kind of, get a preview of what I'm gonna be able to learn how to do and stuff like that and everything. Yeah, it was definitely an interesting experience. *M10*
>
> It was OK. It was kind of awkward trying to explain what was going on. Like, I was trying to figure out, like, how to phrase it, right? And I was also trying to understand it myself. So it was like a lot of, like, processing to figure out everything. *M13*
>
> I feel like, for math questions, it's a lot of reading for math. We did that back in sixth and seventh grade when we first took the practice test. I feel like it's more reading now than it used to be. *M20*
>
> There were a lot of really easy and a lot of really hard ones. So I feel like on the easy ones, it's like, I feel like I know the information, but it feels, like, too easy or, like, too much of a, too good to be true. And then the hard ones were just, I was at, I was like, I knew if I knew how to do it, I could do it. I just didn't know how to do it. *M34*
>
> I mean, it felt normal for the most part. I mean, obviously, it was a little different for me to have to say everything out loud. Especially for me, because when I read things, I comprehend it and process it a lot better when I'm reading it just in my head to myself. When I'm trying to say it out loud, it's like multitasking—reading while doing something—which I struggle with a lot. So just reading it in my head is easier to process. It took a little more time to read it, but other than that, it was normal. *M38*
>
> It was scary as heck. Like, I'm talking about really, really scary. I'm just not good at math. Like, math is not my subject at all. Like, I'm telling you, every semester, math is either one of the ones that everyone is like, "Oh, you gotta get your grade up before you don't pass." And I just sit there and I stress because I don't know. *M58*

Strategies

2. How would you describe your general approach, in terms of strategies, for answering the questions?

In response to postexperience question 2, participants' descriptions of their Math problem-solving strategies revealed consistent patterns, with most participants indicating beginning their approach by carefully reading and attempting to understand the question before proceeding to calculations. Many participants reported using knowledge gained from previous math courses and relying on familiar mathematical properties and procedures. A notable strategy described by several participants was the systematic elimination of answer choices in multiple-choice questions, while others emphasized the importance of breaking down complex problems into manageable parts. Some participants specifically mentioned the value of writing out their work on scratch paper and using calculators—both of which were available during the think-aloud activity and are available in actual testing—though a few admitted to not having clear strategies or simply trying to apply whatever method they could remember.

> I've been taught multiple different approaches for trying to solve math problems, and, depending on the different math problems, it depends on, like, my different approach. But typically I start to, like, look at, first analyze the graph if there is a graph and then I'll read the question, or I'll read the question and then I'll try and figure out how to piece [apart] the question first. And then I go back through and look at the answers, like, once I understand the question, but I've learned different strategies to look at the answers twice and then go back and look at the question. It just depends on the type of question [it] is. *M13*
>
> At first, I was just writing it on paper because I was trying to solve it out, and sometimes I would solve it in my head. But I feel like the best way was the calculator. That helped me more. *M21*
>
> First, I always want to read the question and the answers, but mainly the question, to see if I can pick anything out that I can use, either to plug into a formula or to solve in general. Then I look at the answers to see if any of them correlate with what I did and what I got as my answer. *M28*
>
> I think just going on all my past knowledge of everything. Like, we put a lot of emphasis on or, like, it was always, what can I do to solve this? For example, the one with the [*unspecified, but potentially Math question 3, which gives an equation for a circle in the* xy*-plane*], I should have just immediately graphed it, and I tried to solve it instead. It was just, I went straight to what we were taught in school more than, like, straight up what I had to think about. *M34*
>
> First, I guess, try to understand what they're asking. And, for me, that's a big thing because, I mean, I don't know if you know this, but I do have dyslexia. So sometimes I do read things differently or wrong. So, like, I don't know if you noticed, but I would reread the question a few times just to make sure I'm understanding what they're asking. *M46*
>
> My general approach in answering questions for multiple-choice just generally, and for this [activity in particular], is more of a process of elimination. If I have a one-in-four chance at the beginning, if there are four answers and I can cut that in half, that's doubling my chances. And then if I can look at those final two answers, get those other two answers out of my brain, so I can focus specifically on those [remaining], then it kind of narrows me in on those, making me think better. *M56*

"Easy" Question Types

3. Was there a particular type of question that you found especially easy to answer? If so, which one and why?

When asked via postexperience question 3 about the types of think-aloud activity questions in Math they found easy to answer, participants most frequently identified problems involving basic algebra, particularly those requiring solving for a single variable, as well as questions with straightforward geometric concepts, such as triangles. Many participants expressed preference for questions that provided relatable real-world contexts, such as the movie theater problem (Math question 11), over more abstract mathematical scenarios. Not surprisingly, they generally perceived multiple-choice questions, with the structure and support of answer choices, to be more approachable than student-produced response questions, with several participants noting that having answer options helped them verify their work. Notably, some participants struggled to identify specific "easy" questions at all, suggesting that even seemingly straightforward problems presented their own types of challenges.

> Some of the graphing questions were pretty easy because all I had to do was put in the numbers. *M26*

> Some of them, like the ones where I could—the triangle one [*Math question 8*], where all I had to do is subtract one side and then I knew 180 minus that, and then whatever was less, because it couldn't be more than that. So I could just eliminate answers. That was pretty easy. Any of them that I could just eliminate answers off the bat are pretty easy because it just narrows it down to a few answers. *M27*

> Well, for me, I like the ones with the multiple-choice [options] were easier, slightly better with multi-choice. So I would like to have, like, things to choose from. It was easier. All the ones with the graph are easier for me because I'm more of, like, a visual person when it comes to learning. *M32*

> Maybe the one where you just had to, like, solve—specifically, where it was just, like, solving something. I think this might have been the first one we did [*actually, Math question 12*]. That was just, it was $(d-30)(d+30)-7=-7$. It was just solving for d, and it was just [a] cut-and-dry, straightforward equation. *M34*

> Like, the movie theater one at first, because I could relate to it. And then it was, like, a, more of like a—I don't know. I'm not good with, like, I'm good with division and stuff. Like, if, if you give me a big number and you need me to divide them into three different parts, I got you. *M58*

"Hard" Question Types

4. Was there a particular type of question that you found especially hard to answer? If so, which one and why?

When discussing the most challenging questions in response to postexperience question 4, participants consistently identified problems containing extensive text as particularly difficult to navigate. Questions involving multiple variables, especially those requiring graphing or coordinate plane analysis and unknown constants, were frequently cited as problematic. Many participants expressed difficulty with percentage problems and questions requiring multiple solution steps, often noting that these types of problems became especially challenging when they couldn't recall specific formulas or procedures to use. Student-produced response questions were also highlighted as particularly challenging, with several participants expressing anxiety about the precise formatting requirements and lack of answer choices to guide their thinking.

> I think the equations with both x and y, where it's two different variables, where it's $20x$ and $30y$ [*possibly Math question 15*]. I just didn't know how to answer those kinds of questions. *M11*

> So, kind of like, just the ones that have a lot of words in them, like the ones in—I think it was in module two—the one that I skipped. It was, like, just a whole bunch of words in it. I tried to reread it to see if maybe, like, the first time I didn't understand it but maybe [would] the second time. But then, it was just, like, the more I kept reading it, the less it made sense. *M16*

> Oh, it was, What did $a + b$ equal [*Math question 6*]? I feel like I didn't have enough information to go off of it. So I was just at a complete loss of what to do. *M34*

> I think, just for me personally, I think I have a hard time with, like, the ones with a lot of reading, like percent. Like, I think I was kind of struggling on the percent ones because it's saying, like, x is greater than, or something about percent [*Math question 5*]. *M44*

> Oh, alphabet questions. Alphabetical, man. Those are dangerous. Very dangerous. Because I sit there and, on the inside—I don't know if you could tell, but this is a little psychology trick I learned about myself, because I took psychology for a little bit in school, and my teacher—she could actually tell whenever I get nervous, I sit there and I breathe a lot. *M58*

> Yeah, there was this question that was asking about angles of a triangle [*Math question 8*]. It really was confusing because I couldn't find my, the answer. The question was also contradicting because, OK, if they were talking about a triangle, maybe they could have indicated an isosceles triangle, a right-angle triangle—those two words would have really made a difference on it. So it wasn't as clear. *M60*

SLDR Symptom Impact

5. Did you encounter anything in the questions that you had difficulty with given that you have a specific learning disorder affecting reading? If so, what was it, and why was it difficult for you?

When discussing, in response to postexperience question 5, how their specific learning disorder affecting reading influenced their simulated test-taking experience, participants consistently reported needing to read questions multiple times to ensure comprehension and prevent missing crucial information. Many described experiencing visual challenges, such as words appearing to move on the page or difficulty maintaining focus on specific numbers and mathematical symbols. A significant number of participants noted particular difficulty with in-context problems, explaining that the combination of processing text and mathematical concepts simultaneously created an additional cognitive burden. Several students described having developed compensatory strategies, such as breaking down complex text into smaller parts, though they noted these strategies often required additional time and mental effort.

> It's just kind of having to read it a couple times. So it just kind of takes up a little bit more time to kind of have to go back through and read all of it to make sure you get everything. Don't skip a word to make sure you're understanding everything. I've missed more questions that way than—like, not reading it properly and not understanding what I have to do—that, than you could imagine it. *M10*

> I have trouble—like, some of the word problems were very wordy. And so I was trying to, I was getting, like, tripped up over the words because there is more words than there was, like, solving and it was, like, hard to, like, break down the words. *M13*

> Probably just reading it, honestly. Kind of just keeping everything from moving around because I have dyslexia. So it's kinda hard to keep all the words from flying off the page and kinda just focus. *M24*

> I think it was mostly just, like, a little bit on every, on all of them. Just keeping numbers straight because I always, yeah, it wasn't a specific one. It was just, like, on all of them, I had to work to not get numbers mixed up. *M34*

> Sometimes, like, as you saw, I would have to go over it another time. Like, I've kind of trained my dyslexia to be, like, less bad. But I feel like with me specifically, I can read it, but my brain won't really register it. I have to read it, like, again and again, sometimes over [again in] my head [or] for you out loud. *M56*

Final Comments

6. Is there anything about your test-taking experience today or about the test-taking strategies you used today that we haven't talked about yet but that you'd like us to know?

In their final reflections on the simulated test-taking experience, elicited by postexperience question 6, participants offered some suggestions for improving the assessment's accessibility. Multiple participants recommended modifications to the visual presentation of questions, including options for alternative fonts, colored backgrounds, and more readable graph formats. Another suggestion involved redistributing the difficulty level of questions throughout the test section rather than, as is current practice for the Math section, clustering easier questions at the beginning and harder ones at the end.

> This is—I don't know if y'all can do this or not. But one thing I find helpful is, like, changing the font, because some fonts are so much easier to read than other fonts. And also the background of the text—like, sometimes doing it in a white background and then, like, the standard, like—what is it?—like Times Roman or Arial font is very hard to read.

[*Bluebook actually uses Noto Serif 15/24, a font selected for its legibility on a range of digital devices and screen sizes.*] But maybe if it was, like, a different font or, like, if you, like, select which color background you would like, it might be a little bit easier to read or for, like, less pressure. I prefer, like, either, like, a blue background just because it makes the letters pop out more and it blends less, if that makes any sense. *M13*

So I try to not let it get to my head about that kind of stuff. It's not a measure of [whether] you are smart or you are dumb. It—to me, it's a measure of, How are you able to perform if we give you this at this time? *M33*

Maybe on, like, the graphing part, I guess, make it a little more like—not accessible, but, like, easier to read the graph, I guess you could say. Like, because when you would try to zoom into the graph or to see it, it would only show certain numbers or certain things when, like, in my head, I need to see all of the numbers so I know what's, like, going on. *M46*

Honestly, scatter the questions. Like, I noticed that, like, the first, second, third, fourth, fifth questions were always easy. The later questions were always harder. *M56*

Section 5: Discussion

Reading and Writing

PARTICIPANT PERFORMANCE

Participant performance levels on individual Reading and Writing test questions used in this study were determined by College Board subject matter experts, who compared transcripts of student verbalizations of their thinking aloud during their question answering to lists of required cognitive behaviors associated with a given question's type (e.g., Central Ideas and Details). Participants who both answered particular questions correctly and demonstrated all required behaviors were assigned the highest performance level (1), while participants who answered incorrectly, failed to demonstrate appropriate behaviors, or both were assigned lower performance levels. A participant differential (D_p) was then calculated for each participant. This differential was determined by subtracting from the total number of correctly answered questions the number of questions for which all required behaviors were demonstrated. This differential was considered "good" if it represented at least 70 percent of correctly answered questions being so answered while the participant demonstrated all required behaviors associated with the question's type.

Nine of fifteen Reading and Writing participants (60 percent) met or exceeded the threshold for a good D_p, providing evidence that students with SLDR are able to demonstrate cognitively complex thinking in line with the question types' constructs. While the remaining participants had low differentials of 2 or 3, they failed to meet the threshold. (For example, participant RW13 had a D_p of 2, but because he answered only four questions correctly and only two of those by demonstrating all required behaviors, his performance didn't meet the 70 percent threshold.) Even participants with a criterion-failing D_p, though, were still able to demonstrate cognitively complex thinking by demonstrating all required behaviors on half to two-thirds of the questions they answered correctly. In general, these results offer evidence that students with SLDR are able to exhibit cognitively complex thinking in line with the question types' expectations.

QUESTION PERFORMANCE

A question differential (D_q) was similarly calculated for each of the fifteen Reading and Writing questions used in this study. This differential represents the arithmetic difference between the number of participants who answered a given question correctly and the number who also demonstrated all required behaviors associated with the question's type (i.e., attained PL 1). A "good" D_q for a particular question was set at 70 percent or more of all correctly answering participants also demonstrating all required behaviors via their verbalizations.

Ten of the study's fifteen Reading and Writing questions (67 percent) met or exceeded the threshold for a good D_q. The remaining five questions exhibited differentials ranging from 1 to 4. The five questions with criterion-failing differentials were also among the least successfully answered questions in the set, with just two to six participants correctly answering them; however, at least one participant was able to attain a PL of 1 on four of these five questions. The overall findings support the claim that the presented Reading and Writing questions are capable of eliciting cognitively complex thinking in line with the question types' constructs from students with SLDR.

PARTICIPANT PERFORMANCE VIGNETTES

Participant performance vignettes (transcript excerpts) exhibiting highly successful (PL 1) outcomes in line with question types' constructs were obtained for fourteen of the fifteen Reading and Writing questions, providing further evidence that the questions are capable of eliciting cognitively complex thinking from students with SLDR. The exception was question 1, a hard (PSB 7) Words in Context question set in a highly challenging (PSR) science context, which only four participants answered correctly and none while also demonstrating both required behaviors. It seems likely that the high level of difficulty of the question, the significant complexity of its stimulus, and the answer key's unfamiliarity ("exploited," in the relatively uncommon sense of merely "used" rather than "took unfair advantage of") account for this outcome.

PARTICIPANT PERCEPTIONS

Reading and Writing participants gave generally positive or neutral assessments of their think-aloud experience (postexperience question 1). They called out a few test-taking strategies (postexperience question 2) as having been used in the activity; these most prominently included rereading as well as incorrect-answer elimination, keyword matching, and relying on a general sense of "fit" between answer choice and question. Participants varied as to whether they would, under normal circumstances, typically read test questions before reading any stimulus material or vice versa, although the study's protocol required them to read out the stimulus first. In terms of question types they found particularly easy to answer (postexperience question 3), participants generally identified three factors contributing to ease: short passages, the blank-completion question format, and literature passages. Conversely, participants tended to consider "hard" (postexperience question 4) those questions that had longer stimulus passages and/or that involved informational graphics (tables and graphs). When asked to identify SLDR-related issues that impacted their performance on the think-aloud activity (postexperience question 5), participants almost always cited difficulties seemingly internal to themselves related to cognitive executive function (specifically, difficulties in maintaining focus and attention, struggles to retain information and ideas in working memory, and lack of stamina to persist through reading challenges) and/or general difficulties with reading rather than issues arising directly from the test section. Across responses to postexperience questions 5 and 6 (the latter being intended as an open-ended "catchall" for any feedback the participants wanted to share but hadn't previously been asked for), participants did, however, sometimes identify elements specific to the testing environment, including the Bluebook platform's lack of a dark mode, lack of access to a dictionary/thesaurus in which to look up words they didn't understand, and the use of italics in certain test questions.

It should also be noted that numerous participants indirectly voiced frustration with the think-aloud protocol's requirement that they read aloud each test question prior to attempting to answer it. This requirement, intended to ensure that each participant fully interacted with the question as written, almost certainly impeded some participants' test-taking performance and, in some cases, laid bare their struggles with decoding and fluency, a topic we return to in the general discussion, below.

Math

PARTICIPANT PERFORMANCE

Seventeen of twenty-one Math participants (81 percent) met or exceeded the threshold for a good participant differential (D_p), thereby providing evidence that students with SLDR are capable of demonstrating cognitively complex thinking in line with the question types' constructs. The remaining four participants exhibited low differentials of 1; while these participants didn't meet the criterion for a good D_p, they were nonetheless able to demonstrate cognitively complex thinking in relation to half to two-thirds of the questions they did correctly answer. In general, these results offer evidence that students with SLDR are able to exhibit cognitively complex thinking in line with the question types' expectations.

QUESTION PERFORMANCE

Twelve of the fifteen Math questions used in this study (80 percent) met or exceeded the criterion for a good D_q. Two of the remaining questions had differentials of 3, while the third had no true differential given that no participant answered it correctly. The criterion-failing questions, all of which were multiple-choice questions set outside of context (the latter fact weakening the supposition that textual processing load was a significant factor), were among the hardest presented to participants in this study, with PSBs of 6 (one question) or 7 (two questions). At the same time, there were three criterion-passing questions with high PSBs of 6 (2 questions) or 7 (one question), thus supporting the claim that students with SLDR are capable of demonstrating cognitively complex thinking for questions in that uppermost difficulty band. Five participants were still able to attain a PL of 1 on one of the three criterion-failing questions, while no participant attained PL 1 on the other two questions. The overall findings support the claim that that the presented Math questions are capable of eliciting cognitively complex thinking in line with the question types' constructs from students with SLDR.

PARTICIPANT PERFORMANCE VIGNETTES

Participant performance vignettes (transcript excerpts) exhibiting highly successful (PL 1) outcomes in line with question types' constructs were obtained for thirteen of the fifteen Math questions, providing further evidence that the questions are capable of eliciting cognitively complex thinking from students with SLDR. One of the two remaining questions—question 5, a hard (PSB 7) multiple-choice Percentages question outside of context—was, as previously noted, answered correctly by no participant, while the other—question 6, a hard multiple-choice Nonlinear Functions: Make Connections question outside of context—was answered correctly by three participants who, in the process, didn't exhibit any expected behaviors. The fact that no participant answered question 5 correctly is likely attributable to its proneness to a common conceptual error—the improper determination of a percentage increase greater than 100 (an error that an incorrect answer choice supported). Transcript evidence furthermore suggests that participants broadly struggled with question 6 due to the fact that its use of unknown constants made finding an entry point into the question highly challenging, a construct-relevant factor.

PARTICIPANT PERCEPTIONS

Relative to the Reading and Writing participants, Math participants expressed a wider range of overall reactions to the study activity (postexperience question 1), with some participants describing the experience negatively as having provoked anxiety and/or frustration. Participants with dimmer views of the activity tended to point to a lack of appropriate mathematics content knowledge and the challenge of parsing what they perceived as difficult contexts for some of the questions. While the former is entirely construct relevant, as the aim of the SAT Suite Math section is to assess students' facility with the mathematics skills and knowledge needed for college and career readiness, the latter suggests possible conflation with literacy achievement as well as impact of SLDR symptoms.

However, before a possibility ossifies into a certainty, with potentially strong implications regarding test fairness and accessibility, three important considerations should be attended to. First, the necessity of students being able to work fluently with math in context is deeply embedded in high-quality, evidence-based standards documents (e.g., NGA Center for Best Practices and Council of Chief State School Officers 2010), often cited as important by postsecondary math educators (e.g., College Board 2019), a routine aspect of math achievement assessment (e.g., Mullis et al. 2021), and a frequent topic of discussion in the research literature (e.g., Sa'diyah et al. 2024). Thus, the issue isn't whether students' ability to solve math problems set in context is important but rather how it should be done effectively. For the SAT Suite, the assessment of math-in-context is ensconced in the Math section's construct definition and claims (College Board 2024b, sections 4.2.1 and 4.2.2), and in-context questions are subject to word count (section 4.1.8) and appropriate language use (section 2.2.7.1) parameters owing to potential conflation with literacy achievement; moreover, SAT Suite test takers with learning differences are eligible for appropriate accommodations, including extended time (section 2.2.7.3). Second, the three studied Math questions with criterion-failing differentials (or, in one case, no true differential, due to the fact that no participant answered the question correctly) were all lacking a context and thus assessed aspects of “pure” mathematics. Participants’ perceptions of difficulties with in-context questions, whether attributable to lack of appropriate content knowledge, impact of SLDR symptoms, or both are thus inapplicable here. Third, as previously noted, the three criterion-failing Math questions were among the hardest (by PSB) presented to participants, which suggests that lack of adequate subject matter knowledge played a role.

Our analysis of participants’ performance on the Math questions and their reported perceptions of the activity itself leads us to conclude that several things can be true at once. First, participants sometimes lacked the requisite content knowledge to efficiently solve certain particularly challenging questions or, indeed, the ability to solve them at all. Second, SLDR symptoms likely did affect some participants’ simulated test-taking performance, though this impact can’t reasonably explain why the only three questions with criterion-failing differentials lacked contexts. Third, as we speculated for Reading and Writing and as attested to by some participants, the activity task of simultaneously thinking aloud while solving math problems inevitably added a level of unfamiliarity and complexity not reflective of actual operational testing conditions.

Participants’ strategy use (postexperience question 2) tended to emphasize close reading of each problem and, when present, its context; the application of classroom-acquired solving strategies; and, for multiple-choice questions, the elimination of one or more incorrect options to enhance chances of success. The last is notable, in part, because it ties in with questions of having (or lacking) adequate content knowledge, as answer option elimination is more useful when students can’t readily affirmatively key given questions. Participants also mentioned the value of breaking down challenging problems into more solvable components and using scratch paper and calculators (both of which are available during operational testing) to parse problems.

Participants’ comments about easy (postexperience question 3) and hard (postexperience question 4) Math question types fall along predictable lines and were largely mirror images of one another. Participants who identified (at least relatively) easy questions from the activity tended to cite those addressing topics typically covered no later than in the first and second years of high school math coursework (basic algebra, geometry), questions with single variables, and questions with real-world contexts (an interesting outcome given the SLDR participants’ often-reported struggles with parsing contexts). Conversely, participants tended to describe as hard those questions that addressed more advanced math concepts, questions with multiple variables and solution steps, and questions in the student-produced response format, for which no answer options are provided and for which specific, precise answers must be entered to earn credit. Some participants were unable to identify any studied question types they found easy, suggesting a lack of adequate content knowledge.

When asked to assess the impact their SLDR symptoms had on successfully completing the activity (postexperience question 5), participants focused on text processing issues, which entailed rereading and took additional time and effort, and the complications that thinking aloud while solving imposed. Suggestions for improvements to the Math section and the testing experience, most often raised in response to postexperience question 6’s invitation for final comments, involved changes to visual presentation and perhaps an alteration of the question order, which currently has questions arranged in each Math test module from easiest to hardest (according to pretest statistics), although it's not immediately clear why this would be deemed beneficial.

As was the case with Reading and Writing participants, Math participants offered some indications of methodological reactivity. Whether they found the process of verbalizing while solving problems beneficial or detrimental, they often expressed some recognition that the think-aloud requirement resulted in deviations from their typical question-answering approaches.

General Discussion

Results from this cognitive lab study involving students with SLDR can be summarized and evaluated quantitatively and qualitatively.

QUANTITATIVE RESULTS SUMMARY: PARTICIPANT AND QUESTION DIFFERENTIALS

Table 5 summarizes the quantitative analyses performed as part of this study in terms of participant (D_p) and question differentials (D_q).

Table 5. Participant and Question Differentials, by Test Section.

	Differential Type	
Test Section	**Participant** (D_p)	**Question** (D_q)
Reading and Writing	9 of 15 (60%)	10 of 15 (67%)
Math	17 of 21 (81%)	12 of 15 (80%)

In terms of D_p, roughly two-thirds (60 percent) of Reading and Writing participants (n = 15) and roughly four-fifths (81 percent) of Math participants (n = 21) met or exceeded the threshold for "good" differentials, which were set at the level of participants demonstrating all required behaviors (Reading and Writing) or at least one expected behavior (Math) for at least 70 percent of the questions they answered correctly. In terms of D_q, two-thirds (67 percent) of the Reading and Writing questions (n = 15) and four-fifths (80 percent) of the Math questions (n = 15) met or exceeded the threshold for "good" differentials, which were set at the level of at least 70 percent of correctly answering participants also demonstrating all required behaviors (Reading and Writing) or at least one expected behavior (Math).

Given the difficulty of many of the Reading and Writing and Math questions, the fact that all participants reported experiencing symptoms of SLDR to one degree or another (modally at the "moderate" level), and, perhaps most importantly, the inherent challenge of and relative or absolute unfamiliarity on the part of participants with thinking aloud while answering test questions, we judge these to be generally good results. Moreover, given that SLDR symptoms, by definition, impair reading facility, the fact that proportionally fewer participants and questions in Reading and Writing than in Math met the criterion for good differentials isn't unexpected, as the textual demands are significantly higher in the former than the latter.

QUALITATIVE RESULTS SUMMARY: PARTICIPANT PERFORMANCE VIGNETTES AND PARTICIPANT PERCEPTIONS

From analysis of individual participant transcripts, we were able to obtain vignettes exhibiting PL 1—the study's highest—from fourteen of fifteen Reading and Writing questions and from thirteen of fifteen Math questions. The fact that the vast majority of Reading and Writing and Math questions analyzed for this study were able to elicit both correct answers and appropriate behaviors from students with SLDR—and under the artificial condition of a think-aloud procedure—we regard as additional evidence that these questions are performing as intended in eliciting cognitively complex thinking, including from students with SLDR.

Participants' perceptions of the study test questions and the think-aloud activity more broadly, as elicited by a standardized set of six postexperience interview questions, coalesced into a few themes that typically applied to both test sections, except as noted:

- The act of thinking aloud while answering test questions added complexity to the usual question-answering task. This was a more prominent concern among Math participants, as their remarks highlighted more negative sentiment, such as anxiety, than did those from Reading and Writing participants.
- Content knowledge was recognized as being critical in Math. Several participants called out either having not yet learned some of the mathematics skills and knowledge being assessed by the test questions or having forgotten how to solve certain types of questions. The fact that this notion came up frequently suggests that participants perceived the Math section as testing their skills and knowledge, the goal of the SAT Suite's achievement-oriented test paradigm.
- Strategy use was somewhat restricted. Reading and Writing participants tended to articulate a limited range of fairly basic question-answering strategies that included rereading, answer choice elimination, keyword matching, and a general sense of the "fit" of an answer option to the question being asked. Similarly, Math participants mentioned using rereading, answer choice elimination (for multiple-choice questions), the identification and application of standard solving strategies, breaking down complex (or seemingly complex) questions into more recognizable and congenial parts, and using scratch paper and a calculator as aids. The general picture that emerges is one in which participants could apply rote, familiar strategies in previously practiced ways to address relatively straightforward questions but often struggled to extend such understandings to more challenging questions or to follow the logical implication of basic principles when asked to apply them in nonroutine ways. This conclusion is reinforced by participants' frequent mention of answer elimination as an important strategy and, in Math, their preference for multiple-choice questions over those in the student-produced response format.
- Perceptions of ease and difficulty fell along predictable lines. Reading and Writing participants tended to find easier those questions that had short stimulus passages, used the blank-completion format, and were set in literature contexts, whereas they tended to find harder those questions that had longer stimuli and incorporated informational graphics. Math participants tended to

find easier those questions that drew from mathematics concepts typically learned no later than the second year of high school (basic algebra, geometry), questions with single variables, questions with real-world contexts, and multiple-choice questions, whereas they tended to find harder those questions that focused on more advanced math concepts, questions with multiple variables and solution steps, and questions in the student-produced response format. These perceptions, first, reinforce the centrality of content knowledge, especially in mathematics, to answering questions correctly and, second, dovetail with self-reports of SLDR symptom impact, discussed next.

- SLDR symptoms had some discernible, global impact. Most participants self-reported having "moderate" SLDR symptoms, meaning that these symptoms had some test-taking impact but were generally manageable with accommodations, such as extended time. Participants from both the Reading and Writing and Math segments of the study often noted that their SLDR symptoms affected their ability to read and process the questions, particularly Reading and Writing questions and Math in-context questions. This is likely why rereading was mentioned so frequently by participants in both segments, as they used this technique as a compensatory strategy. In addition, some Reading and Writing participants called out struggles with focus, attention, short-term memory, and stamina.
- Participants would favor having more tools and customization options. Among the proffered suggestions for interface refinement were the provision of a dictionary or thesaurus for looking up or confirming the meaning of unfamiliar words and phrases, the availability of a dark mode, and a wider range of display options, such as a choice of fonts and colors.

STUDY LIMITATIONS

Several study limitations should be kept in mind when evaluating the results heretofore presented.

The first and most important is small sample size. While typical for cognitive lab/think-aloud studies such as this, small sample sizes (*n* = 15 for Reading and Writing; *n* = 21 for Math) limit the generalizability of findings and increase the risk that idiosyncratic variables impact results. We've attempted to ameliorate such concerns by including diverse (and well-documented) samples within the constraints of the study design, but this study shouldn't be taken as a definitive analysis of the performance of and challenges faced by students with SLDR in large-scale assessment but rather as one set of data and conclusions complementing the work of many other researchers. As a corollary to the above, this study does include shortcomings with respect to full representation of the SLDR population. Notably, members of some racial/ethnic groups are absent altogether, and participants describing themselves as having "moderate" SLDR symptoms are arguably overrepresented, while few reported "severe" symptoms. In addition, higher-achieving students, as indicated by self-reported high school GPA (HSGPA), are probably somewhat overrepresented in the samples, but this may reflect both grade inflation (Sanchez 2024) and self-selection bias, as we'd expect relatively few academically low-achieving students to volunteer to participate in a study of their test-taking performance.

Second, as was discussed extensively throughout this report, the think-aloud methodology itself, though frequently employed for studies of cognition and generally well regarded, entails both a (greater or lesser) degree of artificiality and, in the case of some participants here, a certain psychological cost. That U.S. secondary students aren't routinely asked to think aloud to a stranger while they attempt to answer sometimes very challenging questions almost goes without saying, and, especially in the Math segment, this requirement, though known in advance, seems to have induced some level of anxiety and frustration among at least some participants, which potentially depressed performance. Moreover, while we sought to make the question-answering experience as authentic as possible (e.g., using actual practice test questions, minimizing probes and prompts), it was, fundamentally, an artificial experience under observation. As is intuitively obvious and as responses to the postexperience interview questions make clear, participants to greater or lesser extents altered their typical test-taking approach to accommodate the study format. Notably, the methodology compelled them to begin each question by reading it aloud. Not only did this expose some participants' struggles with decoding and fluency, which may have provoked some anxiety, but it also ensured that participants always began with reading the stimulus, whereas some, in a more naturalistic setting, may have preferred to begin by reading any multiple-choice options first, say, or by examining an included informational graphic. Ultimately, we deem this degree of artificiality as a necessary, inevitable compromise, an exchange of some degree of verisimilitude for the yielded insights into cognitive processes that would otherwise remain hidden. As we detailed in Section 2: Literature Review, the think-aloud methodology, within well-understood constraints and with appropriate safeguards, remains one of the best and only ways in which to peer into otherwise occluded cognitive processes in essentially real time and with minimal retrospective or inferential biases. At the same time, methodological concerns regarding veridicality, reactivity, and demand-induced bias (Kirk and Ashcraft 2001) can't and shouldn't be dismissed.

Finally, as we noted in Section 3: Methodology, technical constraints required that we use a preexisting SAT practice test form as the source for the questions we asked participants to respond to during the think-aloud activity. To minimize the risk that participants would have previously engaged with these questions in their own test preparation, we selected a practice test that was relatively new, in the linear format (whereas students are encouraged to practice in-platform with a digital adaptive practice test, the SAT Suite's standard format, unless they expect to test on paper for accommodations or other reasons), and in the middle of the sequence of practice forms (based on the assumption that the typical student would start their preparation with either the lowest-numbered [oldest] or highest-numbered [newest] practice tests). This concern about prior exposure to the questions on the part of participants seems to have been theoretical rather than actual: no participant in either Reading and Writing or Math gave verbal evidence of having previous experience with any of the questions, and their performance profiles aren't suggestive of such experience either.

Section 6: Conclusion

This report details the results of a verbal protocol study conducted by College Board, with support from vendor Vidlet Inc., involving samples of high school juniors and seniors who have a specific learning disorder affecting reading (SLDR) thinking aloud as they worked through sets of SAT Suite Reading and Writing and Math questions. The research goals of the study were, first, to ascertain, via qualitative and quantitative means, whether these students with SLDR were able to demonstrate cognitively complex thinking in line with the question types' constructs and college and career readiness requirements and, second, to explore whether participants' performance on the questions or their postexperience reflections on the think-aloud activity would uncover any construct-irrelevant barriers to their success on such questions, and in particular barriers not already addressed by the provision of testing accommodations.

With regard to the first goal, the study's findings support the conclusion that students with SLDR are capable of demonstrating cognitively complex thinking via their responses to SAT Suite Reading and Writing and Math test questions. With regard to the second goal, no clear indications of construct-irrelevant barriers residing in the test sections' designs or delivery method were identified, although participants did offer some suggestions, such as having more control over test questions' visual presentation, that might, if implemented, improve their experience or, at the very least, their satisfaction with it.

It's important to note that the study's positive conclusions regarding students with SLDR are predicated on the assumption that these students have access as needed to appropriate testing accommodations. Nearly all participants (n = 36 across the Reading and Writing and Math segments) reported either having received or expecting to receive extended time and/or extended breaks as part of SAT Suite testing. This is necessary and desirable given that SLDR is, by definition, a condition chiefly affecting reading and text processing, and the provision of additional time helps test takers with SLDR level the playing field with respect to their peers without SLDR and the challenges it imposes. It's worth noting, too, that only six participants (three in Reading and Writing, three in Math) indicated either having received or expecting to receive access to assistive technology, such as text-to-speech, and only one participant (RW14, not among those six) used a screen reader during the think-aloud activity. It’s unclear whether this is due to students and their families being unaware that such technology is available, the sometimes cumbersome nature of available screen reader technology, a conscious choice to decline such an option, some combination of the preceding, or a different reason entirely. Nonetheless, the apparent lack of assistive technology use, at least among this small sample, is curious and merits further study. In 2025, College Board made a test application–native text-to-speech option available as an accommodation, and its impact on test-taking performance will be examined. It’s possible, even likely, that a well-developed and well-integrated text-to-speech accommodation may be beneficial to students with SLDR, who face numerous text processing challenges, and allow them an even better opportunity to demonstrate what they know and can do in literacy and math.

References

Al-Maani, Alaa, Bara'ah AlAbabneh, Bassil Mashaqba, and Anas Huneety. 2024. "Investigating Second Language Learning Strategies Using Think Aloud Protocols: Evidence from Jordanian EFL Learners." *Eurasian Journal of Applied Linguistics* 10 (2): 12–22. **https://ejal.info/article-view/?id=724**.

American Psychiatric Association. 2022. *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed., text revision. American Psychiatric Association.

Atman, Cynthia J., and Jennifer Turns. 2001. "Studying Engineering Design Learning: Four Verbal Protocol Studies." In *Design Knowing and Learning: Cognition in Design Education*, edited by Charles M. Eastman, W. Michael McCracken, and Wendy C. Newstetter. Elsevier.

Bainbridge, Lisanne, and Penelope Sanderson. 1995. "Verbal Protocol Analysis." In *Evaluation of Human Work: A Practical Ergonomics Methodology*, 2nd ed., edited by John R. Wilson and E. Nigel Corlett, 169–201. Taylor and Francis.

Bettman, James R., and C. Whan Park. 1980. "Effects of Prior Knowledge and Experience and Phase of the Choice Process on Consumer Decision Processes: A Protocol Analysis." *Journal of Consumer Research* 7 (3): 234–48. **https://www.jstor.org/stable/2489009**.

Biggs, Stanley F., and Theodore J. Mock. 1983. "An Investigation of Auditor Decision Processes in the Evaluation of Internal Controls and Audit Scope Decisions." *Journal of Accounting Research* 21 (1): 234–55. **https://doi.org/10.2307/2490945**.

Bolton, Ruth N. 1993. "Pretesting Questionnaires: Content Analyses of Respondents' Concurrent Verbal Protocols." *Marketing Science* 12 (3): 280–303. **https://www.jstor.org/stable/184025**.

Botsas, George. 2017. "Differences in Strategy Use in the Reading Comprehension of Narrative and Science Texts Among Students with and Without Learning Disabilities." *Learning Disabilities: A Contemporary Journal* 15 (1): 139–62. **https://files.eric.ed.gov/fulltext/EJ1141985.pdf**.

Bowles, Melissa A., and Kacie Gastañaga. 2022. “Heritage, Second, and Third Language Learner Processing of Written Corrective Feedback: Evidence from Think-Alouds.” *Studies in Second Language Learning and Teaching* 12 (4): 675–96. **https://doi.org/10.14746/ssllt.2022.12.4.7**.

Branch, Jennifer L. 2001. “Junior High Students and Think Alouds: Generating Information-Seeking Process Data Using Concurrent Verbal Protocols.” *Library and Information Science Research* 23 (2): 107–22. **https://doi.org/10.1016/S0740-8188(01)00065-2**.

Branch, Jennifer L. 2013. “The Trouble with Think Alouds: Generating Data Using Concurrent Verbal Protocols.” In *Proceedings of the Annual Conference of CAIS / Actes du Congrès Annuel de l’ACSI*. University of Alberta Library. **https://doi.org/10.29173/cais8**.

Cho, Byeong-Young, Lindsay Woodward, and Dan Li. 2018. “Epistemic Processing When Adolescents Read Online: A Verbal Protocol Analysis of More and Less Successful Online Readers.” *Reading Research Quarterly* 53 (2): 197–221. **https://www.jstor.org/stable/26622508**.

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

College Board and HumRRO. 2020. *The Complex Thinking Required by Select SAT Items: Evidence from Student Cognitive Interviews.* College Board. **https://satsuite.collegeboard.org/media/pdf/sat-cognitive-lab-report.pdf**.

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

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

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

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

Deshpande, Divya S., Paul J. Riccomini, Elizabeth M. Hughes, and Tracy J. Raulston. 2021. “Problem Solving with the Pythagorean Theorem: A Think Aloud Analysis of Secondary Students with Learning Disabilities.” *Learning Disabilities: A Contemporary Journal* 19 (1): 23–47. **https://files.eric.ed.gov/fulltext/EJ1295343.pdf**.

Ericsson, K. Anders, and Herbert A. Simon. 1993. *Protocol Analysis: Verbal Reports as Data*, rev. ed. MIT Press.

Goos, Merrilyn, and Peter Galbraith. 1996. “Do It This Way! Metacognitive Strategies in Collaborative Mathematical Problem Solving.” *Educational Studies in Mathematics* 30 (3): 229–60. **https://www.jstor.org/stable/3482842**.

Haffer, Ann G. 1990. “Beginning Nurses’ Diagnostic Reasoning Behaviors Derived from Observation and Verbal Protocol Analysis.” EdD diss., University of San Francisco. ProQuest 9117892.

Isenberg, Daniel J. 1986. “Thinking and Managing: A Verbal Protocol Analysis of Managerial Problem Solving.” *Academy of Management Journal* 29 (4): 775–88. **https://www.jstor.org/stable/255944**.

Johnstone, Christopher J., Nicole A. Bottsford-Miller, and Sandra J. Thompson. 2006. *Using the Think Aloud Method (Cognitive Labs) to Evaluate Test Design for Students with Disabilities and English Language Learners.* Technical Report 44. University of Minnesota, National Center on Educational Outcomes. **https://files.eric.ed.gov/fulltext/ED495909.pdf**.

Johnstone, Christopher, Kristi Liu, Jason Altman, and Martha Thurlow. 2007. *Student Think Aloud Reflections on Comprehensible and Readable Assessment Items: Perspectives on What Does and Does Not Make an Item Readable.* Technical Report 48. University of Minnesota, National Center on Educational Outcomes. **https://files.eric.ed.gov/fulltext/ED499410.pdf**.

Kirk, Elizabeth P., and Mark H. Ashcraft. 2001. “Telling Stories: The Perils and Promise of Using Verbal Reports to Study Math Strategies.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27 (1): 157–75. **https://doi.org/10.1037/0278-7393.27.1.157**.

Kletzien, Sharon Benge. 1991. “Strategy Use by Good and Poor Comprehenders Reading Expository Text of Differing Levels.” *Reading Research Quarterly* 26 (1): 67–86. **http://www.jstor.com/stable/747732**.

Leow, Ronald P., and Kara Morgan-Short. 2004. “To Think Aloud or Not to Think Aloud: The Issue of Reactivity in SLA Research Methodology.” *Studies in Second Language Acquisition* 26 (1): 35–57. **https://psycnet.apa.org/record/2004-11297-002**.

Lundberg, Gustav. 1984. “Protocol Analysis and Spatial Behavior.” *Geografiska Annaler, Series B, Human Geography* 66 (2): 91–97. **https://doi.org/10.2307/490719**.

Magliano, Joseph P., and Keith K. Millis. 2003. “Assessing Reading Skill with a Think-Aloud Procedure and Latent Semantic Analysis.” *Cognition and Instruction* 21 (3): 251–83. **https://www.jstor.org/stable/3233811**.

Montague, Marjorie, and Brooks Applegate. 1993. “Middle School Students’ Mathematical Problem Solving: An Analysis of Think-Aloud Protocols.” *Learning Disability Quarterly* 16 (1): 19–32. **https://doi.org/10.2307/1511157**.

Mullis, Ina V. S., Michael O. Martin, and Matthias von Davier, eds. 2021. *TIMMS 2023 Assessment Frameworks.* TIMMS and PIRLS International Study Center. **https://timssandpirls.bc.edu/timss2023/frameworks/pdf/T23_Frameworks.pdf**.

NGA (National Governors Association) Center for Best Practices and Council of Chief State School Officers. 2010. *Common Core State Standards for Mathematics*. NGA Center for Best Practices. **http://www.corestandards.org/Math/**.

Nguyen, Lemai, and Graeme Shanks. 2007. "Using Protocol Analysis to Explore the Creative Requirements Engineering Process." In *Information Systems Foundations: Theory, Representation, and Reality*, edited by Dennis N. Hart and Shirley D. Gregor. Australian National University Press.

Nisbett, Richard E., and Timothy DeCamp Wilson. 1977. "Telling More Than We Can Know: Verbal Reports on Mental Processes." *Psychological Review* 84 (3): 231–59. **https://doi.org/10.1037/0033-295X.84.3.231**.

Özcan, Zeynep Çiğdem, Yeşim Imamoğlu, and Vildan Katmer Bayraklı. 2017. "Analysis of Sixth Grade Students' Think-Aloud Processes While Solving a Non-Routine Mathematical Problem." *Kuram Ve Uygulamada Eğitim Bilimleri [Journal of Educational Sciences: Theory and Practice]* 17 (1): 129–44. **https://jestp.com/menuscript/index.php/estp/article/view/492/444**.

Özkubat, Ufuk, and Emine Rüya Özmen. 2021. "Investigation of Effects of Cognitive Strategies and Metacognitive Functions on Mathematical Problem-Solving Performance of Students with or Without Learning Disabilities." *International Electronic Journal of Elementary Education* 13 (4): 443–56. **http://dx.doi.org/10.26822/iejee.2021.203**.

Pressley, Michael, and Peter Afflerbach. 1995. *Verbal Protocols of Reading: The Nature of Constructively Responsive Reading.* Erlbaum.

Russo, J. Edward, Eric J. Johnson, and Debra L. Stephens. 1989. "The Validity of Verbal Protocols." *Memory and Cognition* 17 (6): 759–69. **https://doi.org/10.3758/BF03202637**.

Sa'diyah, Mukhtamilatus, Cholis Sa'dijah, and Susiswo Susiswo. 2024. "Students' Ability to Formulate Situation Mathematically from Context-Based Mathematics Problems." *TEM Journal* 13 (2): 1443–51. **https://doi.org/10.18421/TEM132-58**.

Sanchez, Edgar I. 2024. *Changes in Predictive Validity of High School Grade Point Average and ACT Composite Score After the COVID-19 Pandemic.* ACT, Inc. **https://www.act.org/content/dam/act/secured/documents/R2328-Changes-in-Predictive-Validity-of-HSGPA-and-ACT-Composite-Score-After-COVID-19-2024-09.pdf**.

Sanchez, Edgar, and Richard Buddin. 2016. *How Accurate Are Self-Reported High School Courses, Course Grades, and Grade Point Average?* ACT, Inc. **https://www.act.org/content/dam/act/unsecured/documents/5269-research-report-how-accurate-are-self-reported-hs-courses.pdf**.

Stratman, James F., and Liz Hamp-Lyons. 1994. "Reactivity in Concurrent Think-Aloud Protocols: Issues for Research." In *Speaking About Writing: Reflections on Research Methodology*, edited by Peter Smagorinsky. Sage.

Suto, W. M. Irenka, and Jackie Greatorex. 2008. "What Goes Through an Examiner's Mind? Using Verbal Protocols to Gain Insights into the GCSE Marking Process."

British Educational Research Journal 34 (2): 213–33. **https://www.jstor.org/stable/30032828**.

Taylor, K. Lynn, and Jean-Paul Dionne. 2000. “Accessing Problem-Solving Strategy Knowledge: The Complementary Use of Concurrent Verbal Protocols and Retrospective Debriefing.” *Journal of Educational Psychology* 92 (3): 413–25. **https://doi.org/10.1037/0022-0663.92.3.413**.

Vessey, Iris. 1986. “Expertise in Debugging Computer Programs: An Analysis of the Content of Verbal Protocols.” *IEEE Transactions on Systems, Man, and Cybernetics* 16 (5): 621–37. **https://doi.org/10.1109/TSMC.1986.289308**.

Yayli, Demet. 2010. “A Think-Aloud Study: Cognitive and Metacognitive Reading Strategies of ELT Department Students.” *Eurasian Journal of Educational Research* 38 (Winter 2010): 234–51. **https://www.researchgate.net/publication/286547114_A_Think-Aloud_Study_Cognitive_and_Metacognitive_Reading_Strategies_of_ELT_Department_Students**

Appendix

Exhibit 1: Recruitment Solicitation

College Board is seeking a number of high school juniors and seniors to participate in an upcoming research study. Participants will meet one-on-one virtually (via Zoom) with a moderator, who will walk them through an activity and ask follow-up questions. The activity involves reading, thinking aloud through, and answering a series of digital SAT questions in either Reading and Writing or Math and answering some follow-up interview questions. Our goal is to better understand how students interact with our test questions. This activity will take approximately 90 minutes for each student to complete; on successful completion, participants will receive a $150 gift card.

To be eligible to participate, students must

- be either high school juniors or seniors;
- have previously taken the SAT, PSAT/NMSQT, or PSAT 10 tests from College Board in either paper and pencil or digital format;
- have uninterrupted access to an appropriate digital device (desktop computer, laptop computer, tablet; *not* a phone) with a camera; a private space in which to participate virtually in the activity; and an uninterrupted internet connection robust enough for stable videoconferencing;
- commit to spending approximately 90 minutes in working through test questions and answering follow-up interview questions from the moderator to the best of their ability, on a day and at a time mutually agreeable to the moderator and participant; and
- be willing and able to share as much of their thought processes as possible with the moderator while answering test and interview questions.

Participants from all school achievement levels are encouraged to apply. Participants will **not** be evaluated on whether they answer the study's test questions correctly, and participation in this activity will **not** generate a test score, nor will it affect any prior SAT, PSAT/NMSQT, and/or PSAT 10 scores participants may have.

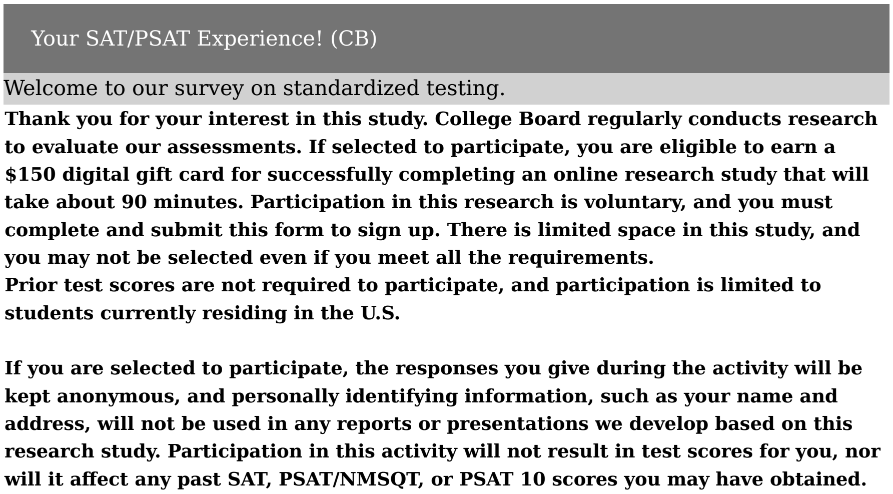
College Board will assign participants to either a reading and writing or a math activity. Participants selected for the math activity should also have access to scratch paper and pencils/pens for use in answering test questions; in addition, they should either be comfortable with the Desmos graphing calculator, which is available as part of the activity, or have their own approved calculator available. For information on acceptable handheld calculators, please visit **https://satsuite.collegeboard.org/sat/what-to-bring-do/calculator-policy**.

This study is for research purposes. Participants' names and other personally identifying information will **not** be used in reports and presentations College Board produces. Sessions will be recorded.

Students (or a parent/guardian, if the student is under 18 years of age) must complete a consent form to participate. This consent form describes the study and its purposes as well as how participants' data will be collected, used, and kept anonymous.

On successful completion of the activity, each participant will receive a $150 gift card, which can be deposited in a bank, deposited into PayPal, or redeemed at one of numerous businesses selected by the participant from a list provided by College Board. Participants may opt out of answering any question or participating in the activity at any time, but successful completion is required to receive the gift card.

Exhibit 2: Recruitment Screener (Survey)

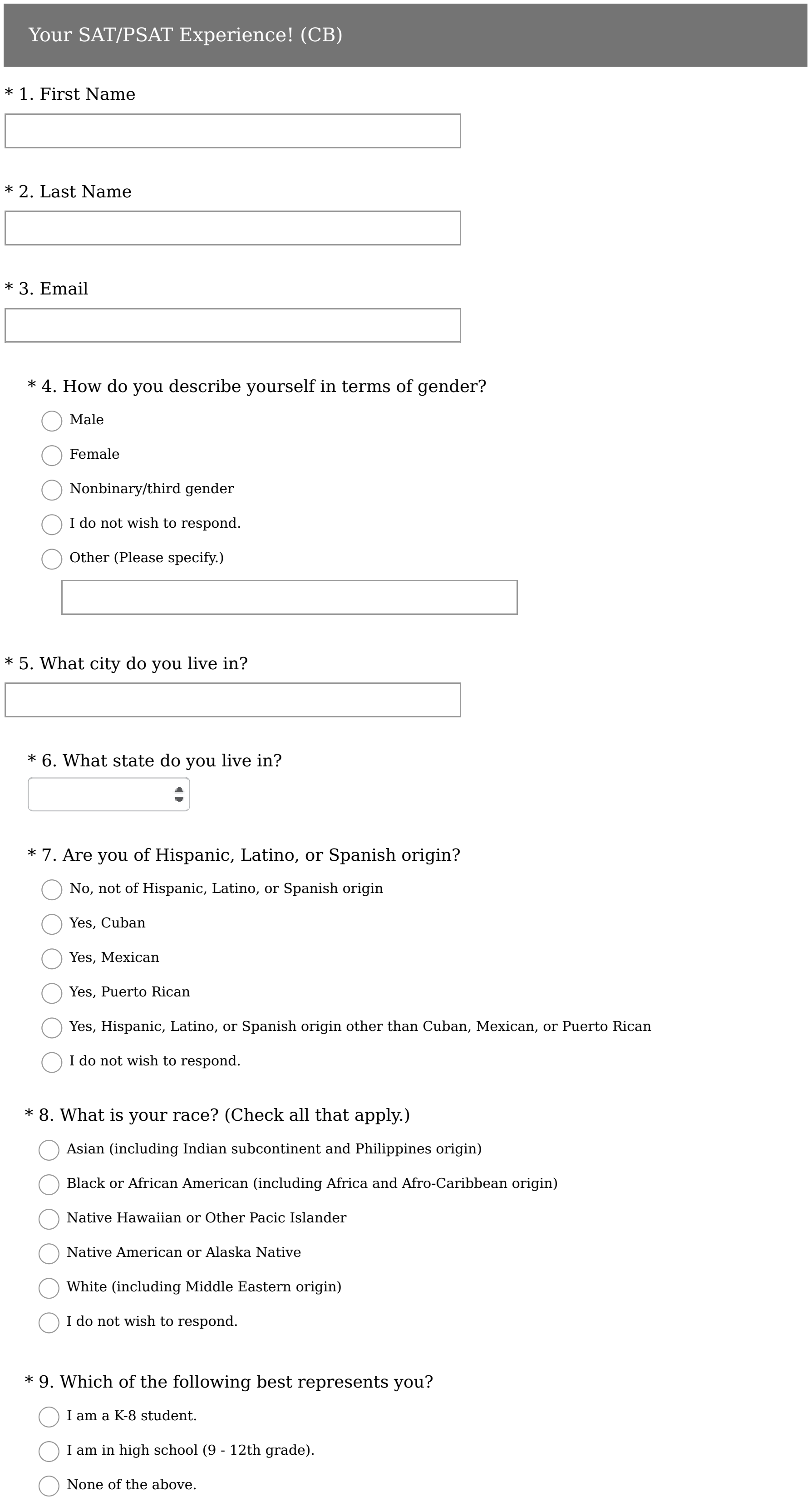


Your SAT/PSAT Experience! (CB)

Welcome to our survey on standardized testing.

Thank you for your interest in this study. College Board regularly conducts research to evaluate our assessments. If selected to participate, you are eligible to earn a $150 digital gift card for successfully completing an online research study that will take about 90 minutes. Participation in this research is voluntary, and you must complete and submit this form to sign up. There is limited space in this study, and you may not be selected even if you meet all the requirements.
Prior test scores are not required to participate, and participation is limited to students currently residing in the U.S.

If you are selected to participate, the responses you give during the activity will be kept anonymous, and personally identifying information, such as your name and address, will not be used in any reports or presentations we develop based on this research study. Participation in this activity will not result in test scores for you, nor will it affect any past SAT, PSAT/NMSQT, or PSAT 10 scores you may have obtained.



Your SAT/PSAT Experience! (CB)

* 1. First Name

* 2. Last Name

* 3. Email

* 4. How do you describe yourself in terms of gender?

○ Male

○ Female

○ Nonbinary/third gender

○ I do not wish to respond.

○ Other (Please specify.)

* 5. What city do you live in?

* 6. What state do you live in?

* 7. Are you of Hispanic, Latino, or Spanish origin?

○ No, not of Hispanic, Latino, or Spanish origin

○ Yes, Cuban

○ Yes, Mexican

○ Yes, Puerto Rican

○ Yes, Hispanic, Latino, or Spanish origin other than Cuban, Mexican, or Puerto Rican

○ I do not wish to respond.

* 8. What is your race? (Check all that apply.)

○ Asian (including Indian subcontinent and Philippines origin)

○ Black or African American (including Africa and Afro-Caribbean origin)

○ Native Hawaiian or Other Pacific Islander

○ Native American or Alaska Native

○ White (including Middle Eastern origin)

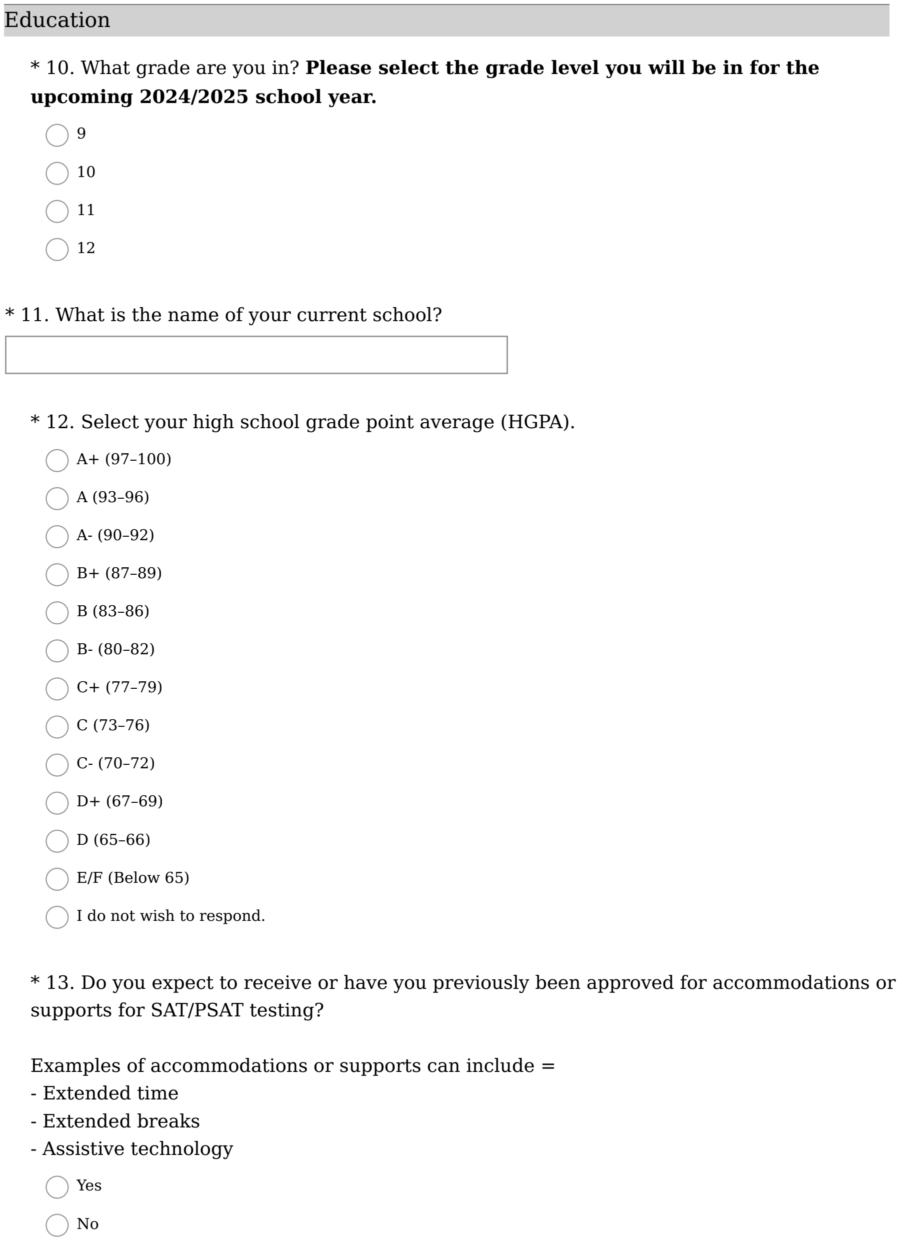
○ I do not wish to respond.

* 9. Which of the following best represents you?

○ I am a K-8 student.

○ I am in high school (9 - 12th grade).

○ None of the above.



Education

* 10. What grade are you in? **Please select the grade level you will be in for the upcoming 2024/2025 school year.**

○ 9

○ 10

○ 11

○ 12

* 11. What is the name of your current school?

* 12. Select your high school grade point average (HGPA).

○ A+ (97–100)

○ A (93–96)

○ A- (90–92)

○ B+ (87–89)

○ B (83–86)

○ B- (80–82)

○ C+ (77–79)

○ C (73–76)

○ C- (70–72)

○ D+ (67–69)

○ D (65–66)

○ E/F (Below 65)

○ I do not wish to respond.

* 13. Do you expect to receive or have you previously been approved for accommodations or supports for SAT/PSAT testing?

Examples of accommodations or supports can include =
- Extended time
- Extended breaks
- Assistive technology

○ Yes

○ No

Accommodations

* 14. For SAT/PSAT testing, what kind(s) of accommodations or supports do you expect to receive or have already been approved for? (Check all that apply.)

□ Extended time on exams

□ Extended breaks

□ Assistive technology (e.g., text-to-speech software)

□ I do not expect to receive any accommodations or supports and have not been approved for any.

□ Other (Please specify.)

* 15. Do you have any specific learning needs or conditions that may impact your test taking experience? (Check all that apply.)

□ Yes, I am an English learner.

□ Yes, I have been diagnosed with ADHD.

□ Yes, I have been diagnosed with a specific learning disorder affecting reading of text.

□ Yes, I am deaf or hard of hearing.

□ Yes, I am blind or have low vision.

□ Yes, I have been diagnosed with autism (ASD).

□ No, I do not have such a need or condition.

□ Other (Please specify.)

Dyslexia

* 16. How were you diagnosed with a specific learning disorder affecting reading of text (dyslexia)?

○ Formal assessment by a specialist (e.g., psychologist)

○ Screening conducted by a teacher or educational professional

○ Self-diagnosis or diagnosis by a family member

* 17. How would you describe the impact of your specific learning disorder symptoms in the context of test taking?

○ Mild: Symptoms are manageable and have minimal impact on test performance

○ Moderate: Symptoms interfere with test taking but can be managed with accommodations

○ Severe: Symptoms signicantly impair test taking ability even with accommodations

ADHD

* 18. How were you diagnosed with ADHD?

○ Formal assessment by a specialist (e.g., psychologist)

○ Screening conducted by a teacher or educational professional

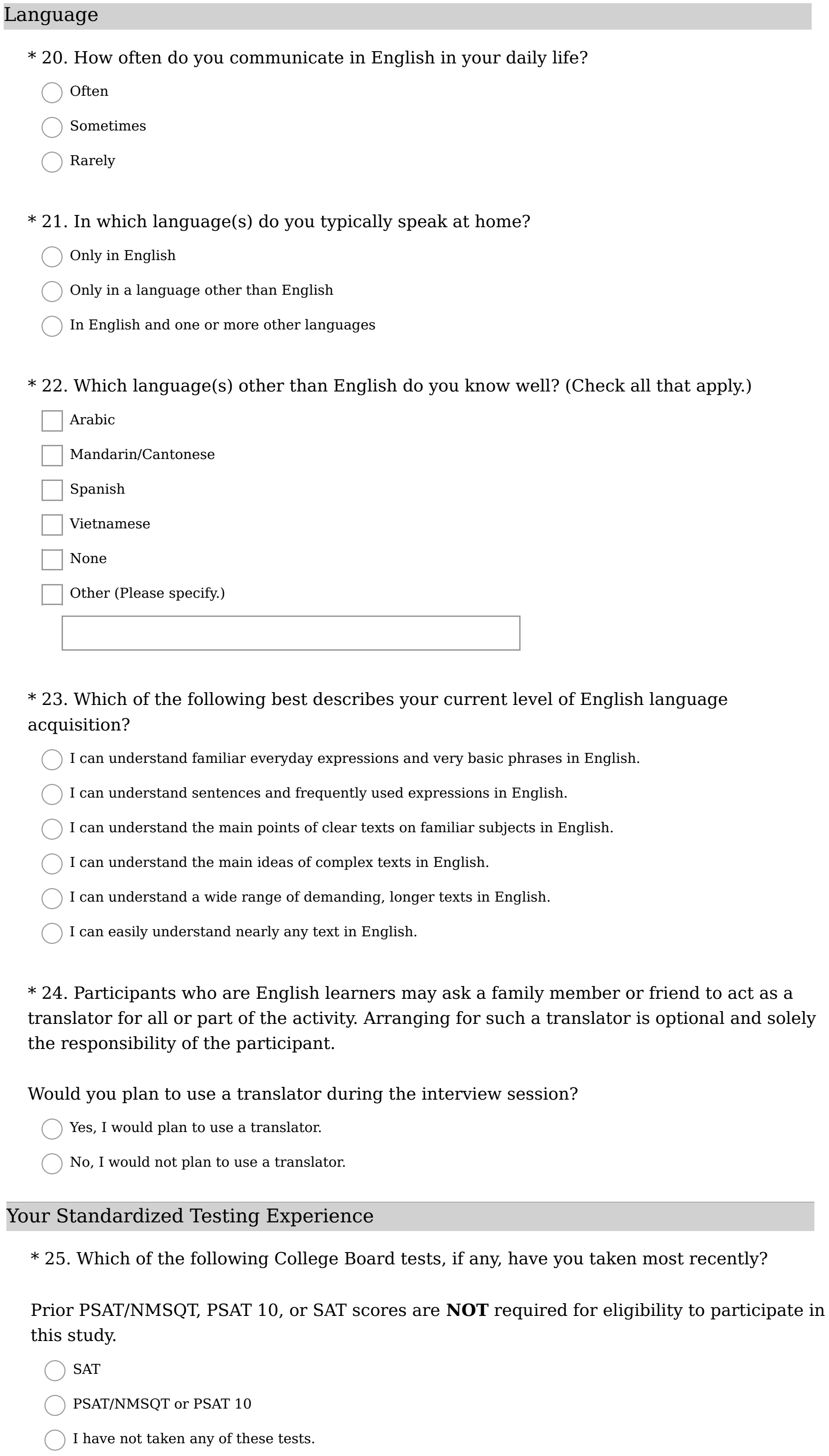
○ Self-diagnosis or diagnosis by a family member

* 19. How would you describe the impact of your ADHD symptoms in the context of test taking?

○ Mild: Symptoms are manageable and have minimal impact on test performance

○ Moderate: Symptoms interfere with test taking but can be managed with accommodations

○ Severe: Symptoms significantly impair test taking ability even with accommodations



Language

* 20. How often do you communicate in English in your daily life?

○ Often

○ Sometimes

○ Rarely

* 21. In which language(s) do you typically speak at home?

○ Only in English

○ Only in a language other than English

○ In English and one or more other languages

* 22. Which language(s) other than English do you know well? (Check all that apply.)

□ Arabic

□ Mandarin/Cantonese

□ Spanish

□ Vietnamese

□ None

□ Other (Please specify.)

* 23. Which of the following best describes your current level of English language acquisition?

○ I can understand familiar everyday expressions and very basic phrases in English.

○ I can understand sentences and frequently used expressions in English.

○ I can understand the main points of clear texts on familiar subjects in English.

○ I can understand the main ideas of complex texts in English.

○ I can understand a wide range of demanding, longer texts in English.

○ I can easily understand nearly any text in English.

* 24. Participants who are English learners may ask a family member or friend to act as a translator for all or part of the activity. Arranging for such a translator is optional and solely the responsibility of the participant.

Would you plan to use a translator during the interview session?

○ Yes, I would plan to use a translator.

○ No, I would not plan to use a translator.

Your Standardized Testing Experience

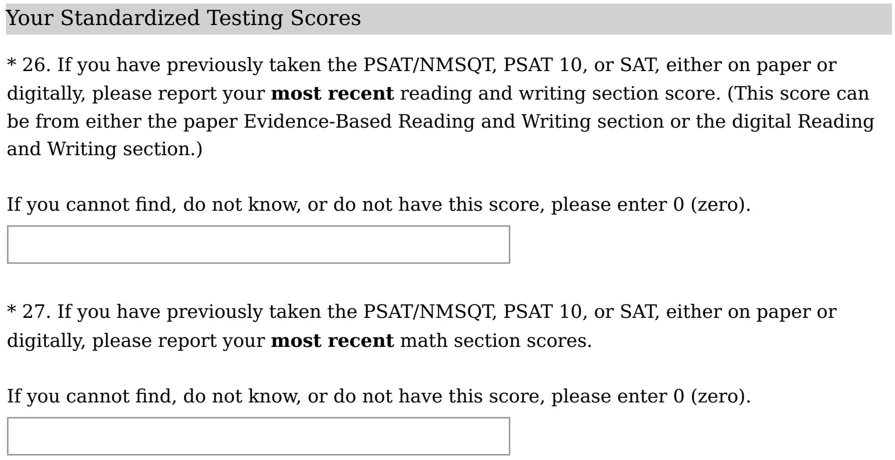
* 25. Which of the following College Board tests, if any, have you taken most recently?

Prior PSAT/NMSQT, PSAT 10, or SAT scores are **NOT** required for eligibility to participate in this study.

○ SAT

○ PSAT/NMSQT or PSAT 10

○ I have not taken any of these tests.



Your Standardized Testing Scores

* 26. If you have previously taken the PSAT/NMSQT, PSAT 10, or SAT, either on paper or digitally, please report your **most recent** reading and writing section score. (This score can be from either the paper Evidence-Based Reading and Writing section or the digital Reading and Writing section.)

If you cannot find, do not know, or do not have this score, please enter 0 (zero).

* 27. If you have previously taken the PSAT/NMSQT, PSAT 10, or SAT, either on paper or digitally, please report your **most recent** math section scores.

If you cannot find, do not know, or do not have this score, please enter 0 (zero).

Exhibit 3: Consent Form

CollegeBoard

Student Research Group Agreement

By signing this agreement, the student identified below ("**Student**"), with consent of their parent/guardian ("**Parent/Guardian**") if the student is under eighteen years of age, agrees to Student's participation in SAT Question Interviews, a research study for College Board ("**Study**"). The Study involves the Student providing feedback to College Board on SAT questions, including but not limited to, providing feedback via a screen-sharing session with a College Board researcher where students may be asked questions or provide feedback about how they answer SAT questions. The study will be conducted entirely online. The activity will take no more than an hour and a half, and on successful completion of the activity, payment will be made via digital payment platform, Tremendous. Student will receive a link from Tremendous to the email address provided which can be used to redeem payment in the form of a bank transfer, PayPal deposit, or a gift card of choice—Tremendous has over 300 gift card options.

Student and Parent/Guardian hereby give their full and complete permission to College Board and its agents to photograph, record (audio and video) Student's participation ("**Images**"). Student and Parent/Guardian grant College Board and its designees, affiliates, agents, subcontractors, and licensees (collectively, "**College Board**") the right to use, transcribe, edit, reproduce, broadcast, publish, exhibit, publicize, and otherwise distribute, without compensation to Student and Parent/Guardian, any Images, along with Student responses, statements and comments Student makes during or in connection with the Study (together with the Images, "**Information**"). The rights hereby granted to College Board are perpetual and worldwide.

Any Images will be stored securely consistent with College Board policies and only College Board personnel involved in the Study and related research and product development will access the recordings. Images will be kept for one year and then securely destroyed. Transcriptions will be kept for two years and then securely destroyed.

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

As the session will include use of live video during the screen-sharing session, please be mindful of your background including, for example, avoid having other individuals in the room, secure any personal items and information from view of the camera and other similar safeguards the Student and Parent/Guardian may wish to consider in their discretion, understanding and acknowledging that the researcher will be able to view the Student's background through the Student's camera.

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

Student and Parent/Guardian understand that College Board is offering to pay Student based on the research activity a US $150 gift card, provided that such payment is permissible under applicable laws and regulations, and the policies and regulations of my employer, if any. Student and Parent/Guardian acknowledge and agree that College Board is not, and that Student and Parent/Guardian is responsible for determining whether Student and/or Parent/Guardian institution's policies and regulations or applicable laws and regulations preclude the Student from participating in the Study or receiving such payment. Student and Parent/Guardian will not consider this agreement an offer to provide this payment if Student and/or Parent/Guardian is prohibited from accepting such payment.

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

Please Print:

Name of Participant Signature Date

Name of Parent/Guardian Signature Date

Student Street Address, City, State

Student Email address

Exhibit 4: Interview Session Training Questions

Note: The following questions were used for participant training purposes prior to the formal start of the think-aloud activity. Session moderators demonstrated thinking aloud for one question using the script included below, after which they gave participants one or (at the moderators' discretion) two questions on which to practice thinking aloud. The training portion of sessions was neither recorded nor analyzed.

READING AND WRITING

Moderator Demonstration Question and Script

> The Younger Dryas was a period of extreme cooling from 11,700 to 12,900 years ago in the Northern Hemisphere. Some scientists argue that a comet fragment hitting Earth brought about the cooling. Others disagree, partly because there is no known crater from such an impact that dates to the beginning of the period. In 2015, a team led by Kurt Kjær detected a 19-mile-wide crater beneath a glacier in Greenland. The scientists who believe an impact caused the Younger Dryas claim that this discovery supports their view. However, Kjær's team hasn't yet been able to determine the age of the crater. Therefore, the team suggests that [blank]
>
> Which choice most logically completes the text?
>
> A) it can't be concluded that the impact that made the crater was connected to the beginning of the Younger Dryas.
>
> B) it can't be determined whether a comet fragment could make a crater as large as 19 miles wide.
>
> C) scientists have ignored the possibility that something other than a comet fragment could have made the crater.
>
> D) the scientists who believe an impact caused the Younger Dryas have made incorrect assumptions about when the period began.

Reading this passage and question, it looks like I'm being asked to figure out how best to fill in the blank with something that makes the most sense in context.

I'm now looking at the answer choices and trying to figure out which is the best answer here. I'm looking for something that logically completes the text.

Choice A says, "It can't be concluded that the impact that made the crater was connected to the beginning of the Younger Dryas." That makes sense because the passage says that the team "hasn't yet been able to determine the age of the crater," so there's still some doubt about whether this crater is even what the team suspects it is. The word "however" also makes me think that Kjær is trying to keep other scientists from jumping to conclusions.

So I like choice A, but I want to look at the other choices before making my decision.

Choice B, “It can’t be determined whether a comet fragment could make a crater as large as 19 miles wide.” This doesn’t make as much sense to me, because the passage doesn’t say anything that would suggest there’s any doubt about whether the crater was made by a comet fragment, only about how old the crater is.

Choice C, “Scientists have ignored the possibility that something other than a comet fragment could have made the crater.” This one seems wrong for the same basic reason choice B was: the passage doesn’t suggest that there’s real doubt about whether the crater was made by a comet fragment.

And choice D, “The scientists who believe an impact caused the Younger Dryas have made incorrect assumptions about when the period began.” No, it’s not this either. The passage doesn’t tell us there’s any real debate about when the Younger Dryas began. There’s a date range, but it’s just presented as a fact. And the passage doesn’t suggest that scientists have made mistakes about dating the period itself. Kjær just seems to want other scientists not to assume that the crater they found is old enough to support some scientists’ hypothesis about how the Younger Dryas started.

So I’ll select answer choice A.

Notice how when I was thinking aloud, I didn’t try to simply summarize what I did after I was done answering. Instead, as I approached this question, I told you exactly what I was thinking as I thought it. I first read the passage and the question aloud and then explained what I thought the question was asking, how I went about answering the question, and why I came up with the answer that I did. I want you to do the same sort of thing when you read and answer test questions today.

Any questions or concerns?

Participant Practice Questions

“The Bet” is an 1889 short story by Anton Chekhov. In the story, a banker is described as being very upset about something: ________

Which quotation from “The Bet” most effectively illustrates the claim?

A) “Then the banker cautiously broke the seals off the door and put the key in the keyhole.”

B) “It struck three o’clock, the banker listened; everyone was asleep in the house and nothing could be heard outside but the rustling of the chilled trees.”

C) “The banker, spoilt and frivolous, with millions beyond his reckoning, was delighted at the bet.”

D) “When [the banker] got home he lay on his bed, but his tears and emotion kept him for hours from sleeping.”

Celebrated Tewa potter Maria Martinez (1887–1980) made her signature all-black ceramic vessels using a heating technique called reduction firing. This technique involves smothering the flame surrounding the clay vessel. ________ the vessel takes on a shiny, black hue.

Which choice completes the text with the most logical transition?

A) On the contrary,

B) For example,

C) Previously,

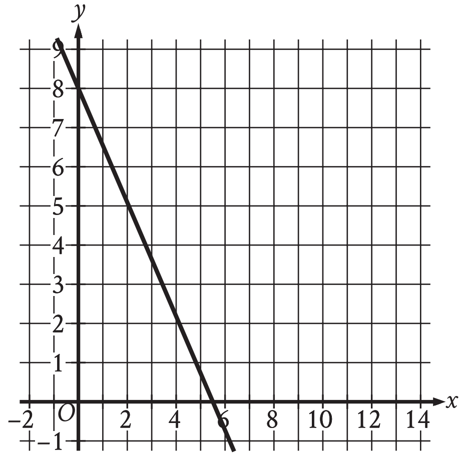
D) As a result,

MATH

Moderator Demonstration Question and Script

This question has a graph in it. The graph shows what looks like a straight line in the *xy*-plane.

Now on to the question.



The graph of the linear function f is shown, where $y = f(x)$. What is the *y*-intercept of the graph of f?

The answer choices are all coordinate pairs.

A) 0, 0

B) 0, negative 16 over 11

C) 0, negative 8

D) 0, 8

This is a question where I need to understand what a *y*-intercept of a graph is. A *y*-intercept of a graph is a point where the graph crosses the *y*-axis. I'm told this is a linear function, so I know there is only one *y*-intercept. From the graph, it appears the line crosses the *y*-axis at the point (0, 8). Since this is a multiple-choice question, choice D is probably my answer.

Let me check the other choices, though. Choice A, (0, 0), isn't right. (0, 0) is the point where the *x*-axis intercepts the *y*-axis. I'm not sure where choices B or C even come from, as (0, negative 16 over 11) and (0, negative 8) don't make any sense here, given the graph we're presented with. So I'm going with my first answer, choice D.

Notice how when I was thinking aloud, I didn't try to simply summarize what I did after I was done answering. Instead, as I approached this question, I told you exactly what I was thinking as I thought it. I first read the passage and the question aloud and then explained what I thought the question was asking, how I went about answering the question, and why I came up with the answer that I did. I want you to do the same sort of thing when you read and answer test questions today.

Any questions or concerns?

Participant Practice Questions

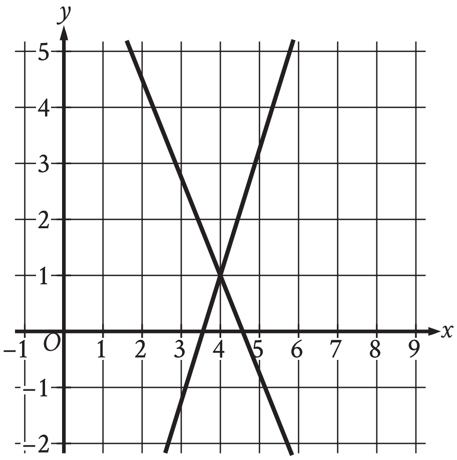
If $4x - 28 = -24$, what is the value of $x - 7$?

A) −24

B) −22

C) −6

D) −1



The graph of a system of linear equations is shown. The solution to the system is (x, y). What is the value of x ?