INFORMING THE DESIGN OF A *GOODNOTES* TESTING INTERFACE

FINAL REPORT SUBMITTED TO *GOODNOTES*

BY

Educational Semiotics and Design Lab (ESD-Lab)

Researchers:

Eunjung Myoung (Research Assistant)
Xinman Liu (Research Assistant)
Maria A. Ruiz-Primo (Professor)
Guillermo Solano-Flores (Professor, Principal Investigator)

Stanford, California, February 19, 2026

Suggested citation:
***Educational Design and Semiotics Lab (2026). Informing the design of a Goodnotes testing interface: Final report submitted to Goodnotes*. Stanford University. Stanford, California, February 19.**

INFORMING THE DESIGN OF A *GOODNOTES* TESTING INTERFACE

FINAL REPORT SUBMITTED TO GOODNOTES

Table of Contents

Sections		Page
	EXECUTIVE SUMMARY Guillermo Solano-Flores, Maria A. Ruiz-Primo	**3**
I	**DESIGNING A PLATFORM FOR IPAD–BASED TESTING: LESSONS LEARNED FROM A RESEARCH AND DEVELOPMENT PROJECT** Eunjung Myoung, Xinman Liu, Maria A. Ruiz-Primo, Guillermo Solano-Flores	**7**
II	**RECOMMENDATIONS** Maria A. Ruiz-Primo, Guillermo Solano-Flores, Eunjung Myoung, Xinman Liu	**12**
III	**BASIC CONCEPTS FOR ITEM DESIGN AND ITEM DESIGN EVALUATION** Eunjung Myoung, Xinman Liu, Maria A. Ruiz-Primo, Guillermo Solano-Flores	**14**
IV	**SEMIOTIC COMPLEXITY IN COMPUTER-ADMINISTERED TESTS: EXAMINING ITEM AND INTERFACE DESIGN IN MAJOR TESTING PROGRAMS** Eunjung Myoung, Xinman Liu, Maria A. Ruiz-Primo, Guillermo Solano-Flores	**24**
V	**IDEAS FOR THE DEVELOPMENT OF A *GOODNOTES* PLATFORM FOR ITEM CREATION** Guillermo Solano-Flores, Eunjung Myoung, Xinman Liu, Maria A. Ruiz-Primo	**46**
VI	**REFERENCES**	**61**

EXECUTIVE SUMMARY

Housed at the Graduate School of Education at Stanford University, the Educational Semiotics and Design Lab (ESD-Lab) conducts research on the improvement of educational artifacts, procedures, and systems (such as instructional materials, tests, computer-based platforms, and curricula) intended to support learning and generate information about learning.

The ESD-Lab's research is grounded in the principles of semiotics, cognitive science, and design. In simple terms, these disciplines respectively examine how society shapes the ways information is represented and understood; how humans process, store, and retrieve information; and how the features of artifacts and environments influence the ways humans interact with them. Accordingly, ESD-Lab's research is guided by the notion that optimal design minimizes unnecessary complexity, thereby maximizing opportunities for individuals to learn and demonstrate learning.

With generous support from *Goodnotes*, in 2023, the ESD-Lab launched an investigation aimed primarily at identifying design factors relevant to creating a platform for iPad-based large-scale testing. *Goodnotes* is an application that allows users to take notes, capture information, and manage files through a wide variety of input modes, including typing, handwriting, and audio recording.

Our project addressed a major concern arising from the current trend of transitioning large-scale testing programs from paper-and-pencil formats to computer-administered tests—the relationship between digital divide and cognitive load. First, differences in the users' access to digital devices may produce differences in digital literacy, including the use of digital devices. Second, when they take tests administered digitally, examinees must use part of their attention figuring out how to interact with the user's interface. This added and potentially unnecessary cognitive load may unfairly impact the performance of individuals from social groups with limited access to or limited familiarity with digital technologies. Ultimately, such unnecessary cognitive load poses a threat to the validity of interpretations of test scores.

Thanks to the intuitive, easy-to-learn, and forgiving design of *Goodnotes*, an iPad-based *Goodnotes* testing platform has the potential to minimize unnecessary cognitive load. In computer-administered tests, examinees must navigate through items, access item content, and enter responses by performing a wide range of actions, such as selecting icons, activating menus, moving the mouse, clicking and double-clicking, typing, and more. This complexity contrasts with the simplicity of the *Goodnotes* interface, in which users' actions closely resemble writing and drawing on paper. An additional advantage of a *Goodnotes* testing platform is its ability to convert handwritten responses into printed text—a feature that potentially can increase scoring objectivity for constructed responses despite individual penmanship differences.

Goals

The project had three goals:

1. To determine the task (item, problem) types that the *Goodnotes* testing platform needs to be able to create and administer, by identifying the wide variety of task types used in large-scale testing programs.

2. To identify the ideal design features of these task types when administered through the *Goodnotes* testing platform and ensure that *Goodnotes*-based tasks are less complex than their conventional computer-administered counterparts.

3. To identify the ways in which students and teachers interact with the *Goodnotes* platform when completing different types of tasks.

Activities Performed

To achieve these goals:

1. We sampled items from three major large-scale testing programs (National Assessment of Educational Progress [NAEP], Programme for International Student Assessment [PISA], and Smarter Balanced Assessment Consortium [SBAC]) that were representative of different task types. We coded and analyzed the frequency of different item features, and the actions users must perform to respond to them.

2. We developed a framework to characterize the complexity of the sets of actions required during test taking. This framework was based on the notions of affordance and constraint and enabled examination of the complexity of computer-administered test items. The framework was also intended to serve as a tool for *Goodnotes* staff to guide the development of the *Goodnotes* testing platform and to evaluate the complexity of user actions involved in test taking.

3. We constructed flowcharts to examine the complexity of the user interfaces used by NAEP, PISA, and SBAC, created *Goodnotes* mockup versions of items, and constructed templates (shells) that allow item creation using pre-established design parameters

4. We met periodically by Zoom which *Goodnotes* staff provided feedback to the ESD-Lab' progress and discussed key concepts guiding the research.

5. We held two in-person meetings during which Goodnotes and ESD-Lab discussed the characteristics of a *Goodnotes* platform supporting classroom assessment activities.

Products and Outcomes

The products and outcomes of the project can be summarized as follows:

- Results from our investigation on the features of items used in large-scale testing programs indicate that, although such programs provide contractors with item specification documents, these documents often lack the level of specificity necessary to ensure high design consistency across item development. Such consistency is critical to minimizing unnecessary cognitive load during test taking. Given its capabilities, *Goodnotes* appears to support the development of test items with superior design quality.

  We developed a conceptual framework for designing and evaluating the design of the users' interfaces used in large-scale testing programs and a coding system to analyze computer-administered items. Using our conceptual framework, we identify basic concepts for item design and item design evaluation. We also provide examples of items from large-scale testing programs and discuss their limitations in terms of affordances and constraints.

- We identified basic design principles relevant to generating and analyzing test items, and to develop a *Goodnotes* platform for item generation.

- Based on the lessons learned, we concluded that a potentially successful strategy for *Goodnotes* to develop innovative products for the testing industry could be based on capitalizing on the important differences between large-scale assessment and classroom assessment. Beyond grading and file management, effective classroom assessment products can support social interaction as a core component of learning and assist teachers in designing, selecting, or customizing assessment activities for their specific classroom contexts. A *Goodnotes* platform for classroom assessment could support learning analytics that not only focus on the accuracy of the students' responses, but also on other aspects of student learning in the classroom, such as tracking students' interactions while responding to a task, engagement time, and ways to support students' metacognition and motivation. In addition, the platform could provide teachers with strategies that support them in understanding student learning progress focusing on critical learning goals. We believe it is important for *Goodnotes* to consider core principles in the development of their classroom products. Examples of those principles are: 1) good instructional/assessment tasks have a level of difficulty that allows students to learn from their errors; 2) effective feedback is effective when it is intended to improve the learner rather than the student's work; and 3) effective assessment design must distinguish between learning and performance.

- We also concluded that *Goodnotes*' initial efforts in the testing industry could include the development of a platform for item creation. Such a platform would enable test developers to systematically manipulate and control many item format and design features.

- Finally, we created flowcharts that allowed examination of the vertical and horizontal complexity of user-interface interaction and put together a series of ideas for *Goodnotes* to develop a platform for item creation. These ideas include the notion of template (shell). A template is a hollow structure that is filled out with textual and non-textual content and which the platform processes and formats according to a set of established design parameters.

Throughout the project, several revisions were made to our activities. Mainly, we de-scoped the analysis of the user-interface interaction phase of the project, since the *Goodnotes* platform had

not been implemented. Also, we paid more attention to classroom assessment and provided *Goodnotes* staff with feedback on their initial efforts to create a *Goodnotes classroom* platform for classroom assessment.

Structure of the Report

This report contains five sections. Section I reports on the development of the conceptual framework guiding the project, the main research activities performed, and key findings. Sections II-V present the project outcomes. Section II offers strategic recommendations to inform *Goodnotes*' development of large-scale and classroom assessment products. Section III discusses basic design concepts that, in our view, are underutilized by test developers, and which *Goodnotes* staff could use to design or evaluate the design of test items. Section IV contains the formal report of the investigation presented in 2025 at the annual conference of the National Council on Measurement in Education. Section V presents a series of ideas for a *Goodnotes* item-creation platform that would allow test developers to systematically create test items according to a rigorous set of design parameters. Section VI contains the references for all the sections. Tables and figures appear at the end of each section.

I.

DESIGNING A PLATFORM FOR IPAD–BASED TESTING: LESSONS LEARNED FROM A RESEARCH AND DEVELOPMENT PROJECT

Motivation for This Project

Important innovations in information technology are propelling a transition in testing practices from the use of paper-and-pencil to the use of digitally- (computer- or iPad-) administered tests. This transition introduces new demands for examinees, who now must navigate complex digital interfaces by interpreting a wide variety of signifiers (e.g., menus, icons) and performing multiple keyboarding, mousing, and gesturing actions (e.g., clicking, hovering, typing, dragging) while simultaneously thinking about their solutions and responses to test items (El-Hashash, 2022).

This added complexity increases test-takers' cognitive load. Beyond thinking about the content itself, examinees must determine how to navigate the interface and enter their responses using digital tools or item formats that may be unfamiliar to them. Potentially, this burden may adversely affect the performance on tests of students with limited access to computers or iPads. Specifically, those with minimal exposure to digital devices may struggle navigating their demands during test taking (Li et al., 2025). This increase in cognitive load concerns test validity—specifically, whether test scores reflect students' ability to navigate the testing platform rather than their actual knowledge and skills (Haladyna & Downing, 2004; Skulmowski & Xu, 2022; Wang et al., 2022).

A *Goodnotes* iPad-based testing platform offers a promising solution to these challenges. Given its handwriting and drawing processing capabilities, *Goodnotes* can simplify testing in both classroom and large-scale assessment contexts. Students can provide responses through natural actions (e.g., writing, drawing, marking) that are akin to those involved in responding to a paper-and-pencil test. Because it can digitize handwritten responses into printed text, such a platform can also ensure the scorers' interpretations of students' responses on constructed tasks are not affected by students' penmanship.

However, realizing this potential requires careful examination of how a *Goodnotes* interface needs to be designed for testing purposes. Understanding the types of tasks used in major assessment programs, the interface capabilities needed to support these tasks, and how students and teachers would interact with such a platform is essential to ensuring that a *Goodnotes* testing interface reduces—rather than introduces—barriers to valid testing.

Original Goals

We aimed to investigate factors critical to designing a *Goodnotes*-based interface for computer- or iPad-administered testing in both classroom and large-scale assessment contexts. We also aimed to provide *Goodnotes* with feedback on interface characteristics needed to effectively

administer assessment tasks and capture test-taker performance data. Our research originally focused on three aspects of a *Goodnotes* platform for large-scale testing:

1. **Survey of Items**: We aimed to identify the wide variety of tasks used in assessment that *Goodnotes* needs to be able to support. Using samples from three major large-scale testing programs: (National Assessment of Educational Progress [NAEP], Programme for International Student Assessment [PISA], and Smarter Balanced Assessment Consortium [SBAC]) We intended to identify the response formats (e.g., fill-in-the-blank, writing, line drawing, free-style drawing, labeling, highlighting) and actions (e.g., keyboarding, mousing, tapping) most frequently used in major large-scale testing programs. We also aimed to examine the correspondence between tasks used in major large-scale tests and those currently used in *Goodnotes* classroom contexts.

2. **Profile of *Goodnotes* Capabilities**: We aimed to identify specific aspects of the *Goodnotes* platform requiring refinement to effectively administer classroom-based assessment and large-scale test tasks. We planned to identify task types that could be developed using current *Goodnotes* capabilities and those requiring further development or refinement. Through an iterative revision process, we intended to script *Goodnotes* versions of each identified task, analyze the complexity of current large-scale assessment items and their *Goodnotes* counterparts, and compare the two interfaces in terms of horizontal and vertical complexity (i.e., the number of signifiers and the length of action sequences examinees must complete to enter responses). The results of this analysis of capabilities were expected to inform the refinement of the *Goodnotes* interface.

3. **User-Interface Interaction Analysis**: We aimed to identify how students interact with the interface and how educators use the *Goodnotes* platform to inform interface design improvements. We planned to have students from different grade levels, cultural and linguistic backgrounds, and socioeconomic statuses respond to sample items in the *Goodnotes* platform and participate in verbal protocols and cognitive interviews. We also intended to interview teachers about their use of *Goodnotes* for formative assessment tasks and any difficulties encountered. The results of this empirical analysis were expected to clarify the extent to which users interact with *Goodnotes* as intended by designers and to identify necessary improvements.

Revised Scope

We decided to de-scope the user-interface interaction analysis due to the lack of an implemented testing platform. However, following *Goodnotes*' request, we expanded our discussion of classroom assessment.

The revised project activities can be summarized as follows.

1. For the task analysis, we examined 660 computer-administered items from public item releases of three large-scale testing programs – NAEP, PISA, and SBAC – administered between 2015 and 2024 across Reading, English Language Arts, Mathematics, and Science. TIMMS (Trends in International Mathematics and Science Study), an important

international large-scale test commissioned by the International Association for the Evaluation of Educational Achievement (IEA), was not included in the study because, at the time, its item public release did not include computer-administered items. Using a social semiotics perspective (see Kress, 2006), we analyzed the selected large-scale testing programs' items according to the functions, actions, and signifiers involved. We presented these results at the 2025 Annual Meeting of the National Council on Measurement in Education. Detailed results are reported in Section IV.

2. Drawing on lessons from the task analysis, we proposed ways in which a *Goodnotes* testing interface could be developed. Our goal was to maximize *Goodnotes*' capabilities while minimizing test-takers' cognitive load and addressing design problems observed in current large-scale assessment interfaces. We created examples of how the *Goodnotes* interface could be improved for different item formats.

3. Drawing on our experience design items, we also articulated a set of ideas for *Goodnotes* to develop a platform for item creation. Such platform would allow test developers to create items of different types according to rigorous format and design specifications

4. Throughout the duration of the project, we met several times with *Goodnotes* staff. In the two in-person meetings, the *Goodnotes* staff shared with us ongoing versions of their platform for classroom assessment. In addition to examining the current usability and functionality of the *Goodnotes* classroom assessment platform, we identified areas for development based on current theory and research on classroom assessment.

Key Findings and Lessons Learned

The knowledge gained from these activities can be summarized as follows:

1. Our investigation of functions, actions, and signifiers across three large-scale assessment programs revealed significant design complexity and inconsistencies in the testing platforms used by major assessment programs. The formats used to deliver items of the same type vary considerably within each program. For example, identical functions may require examinees to interpret and activate different sets of signifiers across similar items. This inconsistency unnecessarily increases cognitive load as, in addition to solving the problem at hand, students must determine how to interact with the interface. These issues stem from a critical gap: Assessment frameworks and item specification documents often lack detailed guidance on digital item delivery characteristics. Without explicit design criteria, the process of item development is not enacted systematically. Rather, it is shaped by idiosyncratic decisions or by the limitations of the software used by test developers to create items.

2. We identified a design paradox in our effort to devise a *Goodnotes* large-scale testing platform: Features valued by *Goodnotes* users are incompatible with large-scale testing practice. While *Goodnotes*' flexibility and its ability to offer multiple pathways to perform actions enhances usability, this same flexibility may constitute a threat to validity of score interpretations and uses. Multiple signifiers and actions (e.g., a back

button and a double-tap touch screen gesture) serving the same function (e.g., undo) increases cognitive load, as users must decide the actions they need to take to enter their responses. In testing contexts, where the goal is to measure content knowledge rather than interface proficiency, providing an abundance of options unnecessarily increases item difficulty. Moreover, the capability for individual customization conflicts with the principle of standardization in test administration—testing conditions must be consistent for all students, except when accommodations are to be provided for students with special needs.

3. Our analysis uncovered fundamental differences in interaction modalities between platforms. While computer-administered tests rely exclusively on mouse-based interactions, tablet implementations must accommodate both Apple Pencil and touch-based input methods. This dual-input capability introduces critical design decisions regarding affordance integration; specifically, whether to leverage the full spectrum of tablet-native interactions or constrain them to maintain assessment standardization. Each choice significantly impacts the resulting combination of signifiers and actions and user experience patterns. The general user interface presents even greater complexity, encompassing considerations that extend beyond item-specific interactions. Design decisions must address typography specifications, navigation paradigms (e.g., scrolling versus page-swiping), zoom functionality (e.g., gesture-based versus button-controlled), page progression methods (e.g., gesture-based swiping versus explicit next/back buttons), and annotation capabilities that mirror traditional paper-and-pencil testing affordances.

4. We concluded that, unlike the existing *Goodnotes* application, a *Goodnotes* testing platform needs to prioritize consistency over flexibility. Specifically, to minimize ability to navigate a digital interface as an extraneous factor in the measurement of student skills, the *Goodnotes* testing platform needs to meet the following conditions:

    - *Standardization of interaction patterns*. Identical tasks should behave identically throughout the platform, allowing examinees to learn interaction patterns once and to apply them consistently across items.

    - *Minimization of signifier redundancy*. Each function should be executed in one single way consistently across items, thus minimizing decision-making unrelated to test content.

    - *Enforcement of constraints*. When assessment tasks specify limits (e.g., number of options to select, maximum word length), the interface should enforce constraints automatically (e.g., by alerting the examinee when the number of options requested is not being met, by disallowing entering text in response boxes when the maximum length of words has been exceeded), rather than assuming that examinees will follow directions on how to enter their responses to items.

    - *Minimization of cognitive load complexity*. Interface design should use the minimal number of signifiers and require the minimum number of actions. Increasing the cognitive load may increase construct-irrelevant variance.

5. We observed that efforts being made to develop a *Goodnotes* classroom assessment platform were focusing mainly on the information collected from classroom activities (e.g., functions that allow teachers to grade students' work, keep students' records, etc.). There is an important set of opportunities for *Goodnotes* to develop features related to social interaction (e.g., how students work together, what kinds of classroom conversations teachers facilitate, how teachers provide feedback). We also observed that efforts to develop a classroom assessment platform appeared to be primarily focused on channeling information from classroom activities into formalized, gradable artifacts (e.g., homework, assignments, tests), with comparatively limited support for documenting and analyzing informal, "in-the-moment" process-oriented assessment practices that emerge through classroom dialogue, collaboration, and responsive teacher feedback. There is a room for a *Goodnotes* classroom assessment platform to expand its support for informal assessment—the form of assessment in which teachers obtain information on their students' progress towards the learning goals, through unstructured activities (e.g., quizzes, short conversations, calling out individual students to participate) that are often improvised and emerge naturally from classroom social interaction.

6. We also concluded that any efforts or actions towards developing a *Goodnotes* classroom assessment platform need to be independent from the efforts and actions being taken to develop a *Goodnotes* large-scale testing platform. Specifically, we concluded that efforts to develop a *Goodnotes* classroom assessment platform need to be grounded on a robust conceptual framework on classroom assessment for providing teachers with:

    - ideas for organizing their teaching through activities with diverse forms of social interaction;

    - resources for conceptual understanding about the importance of the characteristics of instructional and assessment tasks in the classroom (e.g., productive struggle, task variations needed to promote conditional knowledge – knowledge that is more likely to be transferable);

    - resources to understand the distinction between learning and performance during assessment design – appropriate performance in a given task does not necessarily mean that the students have learned the content in a manner that will be applied at a later time in different contexts;

    - resources for promoting the participation of all students in class through multiple forms of formative formal and informal formative assessment and multiple and group configurations (e.g., small groups, pairs);

    - resources and ideas for teaching multiple forms of knowledge (e.g., factual, procedural, schematic, and conditional) and multiple forms of knowledge representation (e.g., verbal, graphic, gestural, written, with formulas, etc.).

II.

RECOMMENDATIONS

This project identified important issues and yielded lessons learned that the ESD-Lab believes *Goodnotes* may be interested in considering in planning its development as a company offering assessment products and services.

1. The development of a *Goodnotes* large-scale testing platform needs be consistent with the principle of standardization in testing—to minimize performance differences attributable to factors unrelated to the knowledge or skills being assessed, students must be tested under the same conditions. Paradoxically, while one of the strengths of the *Goodnotes* platform is the wide range of options available for personal customization, that features is not recommendable in testing platform, except for cases when examinees with special needs need to be provided with accommodations such as larger font size, colored background, highlighting, etc. Moreover, to minimize unnecessary cognitive load, such a testing platform need to use the minimum set of signifiers and actions needed for examinees to navigate the platform and provide their responses.

2. *Goodnotes* may also want to consider developing software and applications to support systematic and effective item creation in the testing industry. The fact that state, national, and international testing programs are administered by a small number of companies limits *Goodnotes*' possibilities to access the testing market, especially because tests are typically computer-based, not iPad-based. However, as our investigation revealed, contractors working for testing programs do not appear to use sufficiently powerful tools that support item creation based on rigorous design parameters. A *Goodnotes* platform for item creation would represent a major advancement in the field of testing by enabling the systematic development of test items. The platform would allow test developers (the users) to select formats for different problem types and specify values for multiple design parameters defining characteristics of text (e.g., length, position, font style, font size), tables (e.g., number of rows and columns, length of table captions), and figures and illustrations (e.g., color, line style, complexity, realism), among many other features. In addition to enabling rapid and systematic item creation, such a platform would ensure design consistency among items within the same grade level and testing program—a property essential for high-quality assessment and one that has yet to be fully achieved by existing testing programs.

3. Future efforts to develop *Goodnotes* platforms for large-scale testing and classroom assessment will be effective to the extent that they are sensitive to the differences between these forms of assessment, as they serve different purposes and involve different users (i.e., examinees vs. teachers). Large-scale assessment is typically summative (assessment *of* learning), focuses on learning as an outcome (i.e., the knowledge and skills acquired at the end of a lesson, unit, course, or academic year), uses standardized tasks administered in the same way to all students, and is intended to generate

information—usually in the form of scores—that compares each student's learning to predefined learning objectives or to a reference group or population. Such assessments are typically used to assign grades. In contrast, classroom assessment may be either summative or formative. Formative assessment (assessment *for* learning) focuses on learning as a process, including how students think and construct knowledge and the processes through which they come to understand content. It uses both formal tasks and informal (often improvised) activities, may involve individual and group work, and is intended to produce qualitative and quantitative information that teachers primarily use to adjust instruction and support student learning. Classroom assessment takes place as an integral part of teaching and learning activities.

4. The development of a *Goodnotes* classroom assessment platform needs to be grounded in a conceptual framework of classroom teaching and learning as a process of social interaction. This framework could guide and support the actions teachers perform in both formal and informal assessment activities. Central to effective assessment is teachers' ability to create conditions for social interaction that allow them to communicate with students individually, in small groups, or with the whole class. The ESD-Lab team believes that an effective *Goodnotes* classroom assessment application needs to be able to support teachers in generating learning and assessment activities tailored to their specific classroom contexts, rather than simply providing highly structured formats for collecting information and scoring student performance. While user customization is antithetical to standardization in a *Goodnotes* large-scale testing platform, it appears to be an essential feature of a *Goodnotes* classroom assessment platform. The conceptual framework for classroom teaching and learning mentioned above would ensure that such customization is sensitive to the forms of social interaction and types of knowledge that are central to teaching and learning across diverse classroom contexts.

III.

BASIC CONCEPTS FOR ITEM DESIGN AND ITEM DESIGN EVALUATION

Introduction

This section introduces basic concepts and reasoning for designing and evaluating iPad-administered items. Originated primarily from the fields of semiotics, design, and cognitive psychology, these concepts are key to minimizing inconsistent user-device interaction patterns that impact user interface usability and constitute a threat to validity in testing because they impose irrelevant cognitive load.

The ideas discussed are supported by findings from our investigation into the design of digitally-administered tests (See Section IV). Our examination of hundreds of items currently used by three major large-scale assessment programs—NAEP, PISA, and SBAC—revealed frequent and serious limitations in item design in current test development practices. The lessons learned could inform *Goodnotes*' efforts to create an item creation platform through a set of rigorous design parameters.

First, we explain three core design principles: **affordance, constraint, and consistency**. Then, we discuss two related cognitive processing concepts—**cognitive demand and cognitive load**. Finally, we illustrate the ideas discussed by evaluating the design of publicly released multiple-choice items from a large-scale testing program according to these three concepts.

Core Design Principles

Affordances: Visual Cues That Guide Action

The term *affordance* refers to the set of attributes of objects, artifacts, or systems that influences how they are used. According to Gibson's (1979) ecological perspective, affordances are action possibilities that the environment provides. Using a cognitive perspective, Norman (2013) further developed this view by emphasizing that the characteristics of objects determine how they are used, whether they are used as intended, and how individuals interact with them.

In simple terms, affordances are cues that show users what they can do. Good affordances make the correct action obvious without explanation.

In the context of assessment design, affordances:

- make items likely to be acted upon through intended functions;
- enable students to both access content and provide responses with ease; and
- lead to proper actions and effective platform interactions, thereby minimizing trial and

error.

Constraints: Design Limits That Prevent Errors

The term *constraint* refers to a set of attributes that limits the ways in which an object, artifact, or system can be used, restricting possible actions to only those critical for correct use. In simple terms, constraints are design features that prevent users from making mistakes and guide their behavior toward specific intended actions.

In the context of assessment design, constraints:

- prevent undesirable or ineffective actions;
- restrict interactions to only those strictly needed for content access and response; and
- minimize unnecessary errors and incorrect actions that are irrelevant to the content being assessed.

In Figure 3.1, the text input box indicates where students need to type their answers (i.e., affordances). Simultaneously, the word counter imposes a limit on response length, and words beyond the limit are not recorded (i.e., constraints).

Consistency: Predictable Patterns Across Items

The term *consistency* refers to the extent to which different objects or artifacts of the same kind provide a common set of affordances and constraints. In simple terms, consistency is attained when the user-interface interaction is the same across items of the same type. Consistency:

- ensures that all items of the same type (e.g., multiple-choice, plotting) have the same set of identical signifiers and require identical sets of actions;
- increases learnability; students can transfer the experience gained from responding to a given item in the ways they respond to other items of the same type; and
- minimizes cognitive load by offering routinely consistent and recognizable patterns.

Figure 3.2 shows two items whose design is inconsistent. In one item, students need to select the right option by filling in a circle; in the other, students need to highlight the option.

Cognitive Processing

Understanding how students process assessment items is crucial for a *Goodnotes* platform design. Poor interface design can unnecessarily increase item difficulty and, ultimately, threaten validity in testing.

Cognitive Demand

Cognitive demand refers to the thinking and reasoning required to respond correctly based on content knowledge and skills.

Items requiring fact recall (e.g., names or dates) are typically less cognitively demanding than items requiring problem-solving with novel scenarios or conflicting information. Higher cognitive demand indicates greater knowledge complexity needed for a successful response.

Cognitive Load

Cognitive load refers to the total mental processing required to access and respond to an item. It comprises two components:

- *Intrinsic cognitive load*: Cognitive load involved of processing information that is relevant to the knowledge and skills being assessed by a specific task.

- *Extraneous cognitive load*: Cognitive load involved in responding to a task, but which is irrelevant to the target knowledge and skills. It includes the mental effort spent determining required actions, navigating the interface (e.g., interpreting signifiers and figuring out how to enter a response). Optimal design minimizes extraneous cognitive load allowing the user to focus their attention and effort on responding to what the item is asking.

Ideally, all the cognitive load imposed by a task should be intrinsic. Extraneous cognitive load increases when students need to interpret unclear signifiers or figure out how to interact with the interface, rather than focus on the task at hand. High-quality item design minimizes extraneous cognitive load through optimal affordances, constraints, and consistency across similar item types.

An Example

In this sub-section we examine two selected-response items from a major large-scale testing program. The examples shown in Table 3.1 illustrate how minor design differences create significant usability issues. In Item 1, examinees need to select one answer from four options (A, B, C, and D) using radio buttons (circles). In Item 2, examinees need to select exactly two options from six options by checking on the corresponding boxes. Despite the specific requirement in the number of options to select, the interface does not restrict the number of options that the examinee can select.

Table 3.2 shows the design analysis of the same items. Item 1 imposes a minimal extraneous cognitive load due to a clear single-selection constraint and an obvious interaction method, allowing students to focus primarily on content evaluation. In contrast, Item 2 imposes unnecessary cognitive load due to the absence of selection constraints, which forces students to self-monitor the number of their selections. The faulty design diverts the examinee's attention, which needs to be partially used in ensuring how to interact with the interface rather than

understanding and processing the item's content.

Table 3.3 compares in detail Item 1 and Item 2 in terms of affordances, constraints, design consistency, and cognitive load, and provides suggestions for design improvements.

Design Recommendations for *Goodnotes*

Based on the ideas presented, we offer several item design recommendations for a *Goodnotes* testing platform.

Constraint and Error Management

- Make constraints automatic: If instructions specify restrictions (e.g., selections, word count, time), the interface should enforce them without assuming that the user will always follow directions on how to respond to the items.

- Design for error prevention: Anticipate common mistakes and make them impossible, rather than correctable.

Consistency and Clarity

- Standardize task patterns across the platform.

- Design for immediate clarity: Interface appearance should communicate function with a minimal explanation.

- Provide meaningful feedback that allows users to immediately see the consequences of their actions.

Cognitive Load and Validity

- Prevent construct-irrelevant difficulty caused by interface complexity.

- Minimize cognitive load.

Accessibility and Technical Implementation

- Maintain cross-item consistency within the same testing interface.

- Support accessibility from the outset rather than retrofitting it later.

Conclusion and Next Steps

Implementing these design principles are critical to positioning *Goodnotes* as a leader in assessment development technology. Consistent, well-designed interfaces not only improve user experience but also ensure valid measurement of student knowledge and skills. Recommended

next steps include:

- evaluating existing *Goodnotes* assessment features according to these principles;
- developing platform-specific design guidelines;
- implementing user testing protocols; and
- training design and development teams on cognitive load principles.

Table 3.1
Two Items from the SBAC

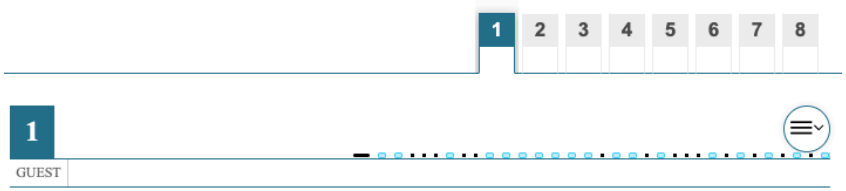
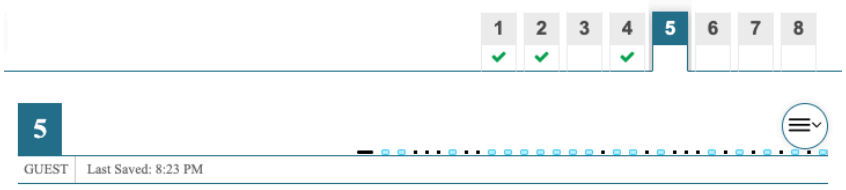
Item 1: Selecting one option	Item 2: Selecting several options
	

Table 3.2
Comparative Design Analysis of the Items Shown in Table 1

Item	Item 1	Item 2
Design Element	Single Choice	Multiple Choice
Option Signifiers	Letters A, B, C, D differentiate options	Squares indicate beginning of each option
Selection Method	Click circle; filled circle + dotted line indicates selection	Click square; checkmark + dotted line indicates selection
Deselection	Automatic when another option selected	Click selected option again to deselect
Constraints	Only one selection possible; cannot deselect all	No constraints—can select any number despite instructions to select two

Table 3.3

Comparison of the Items Shown in Table 1 According to Core Design Principles

Item 1	***Item 2***
Affordances	
● Options signifiers: Letters A, B, C, and D differentiate options. ● Option selection: A black-filled circle in the letter and a dotted line indicate that an option has been selected. ● Option deselection: An option is deselected when another is selected. ● Clickability: The cursor changes from  to  when hovering it over an option that is clickable. Then a green circle with a light-blue border appears after clicking on the option.	● Option signifiers: A square indicates the beginning of each sentence. ● Option selection: A black box with a white check mark and a dotted line indicate that an option has been selected. ● Option deselection: A selected option is deselected by clicking again on it. ● Clickability: The cursor changes from  to  when hovering it over an option that is clickable. Then a green check mark inside a hollow box with a light-blue border appears after clicking on the option.
Constraints	
● Only one option can be selected. ● Actions possible only with clickable signifiers. ● Once an option is selected, it is not possible to deselect all the options.	● Actions possible only with clickable signifiers.
Consistency	
Affordances ● Option signifiers: different. ● Option selection signifiers: different. ● Option deselection signifiers: different. ● Clickability: same logic; same signifiers with slightly different sets of colors.	
Constraints ● Number of selectable options: different. ● Allowed actions only with clickable signifiers: same. ● Deselection actions: different.	
Cognitive Load	
● Minimum extraneous cognitive load while responding to the item, if considered in isolation. However, increased extraneous load in test taking due to inconsistent design across items.	● Increased extraneous cognitive load while responding to the item due to the lack of constraints that ensure that the number of options selected is consistent with that stated in the directions. ● Increased extraneous load in test taking due to inconsistent design across items.
Suggestions for Design Improvement (Both Items)	
● Use different signifiers depending on the allowed number of options selected (e.g., circles when one option must be selected, squares when more than one option can be selected). ● Create constraints to restrict the number of options the directions ask students to select.	

Figure 3.1
An Example of Affordances and Constraints

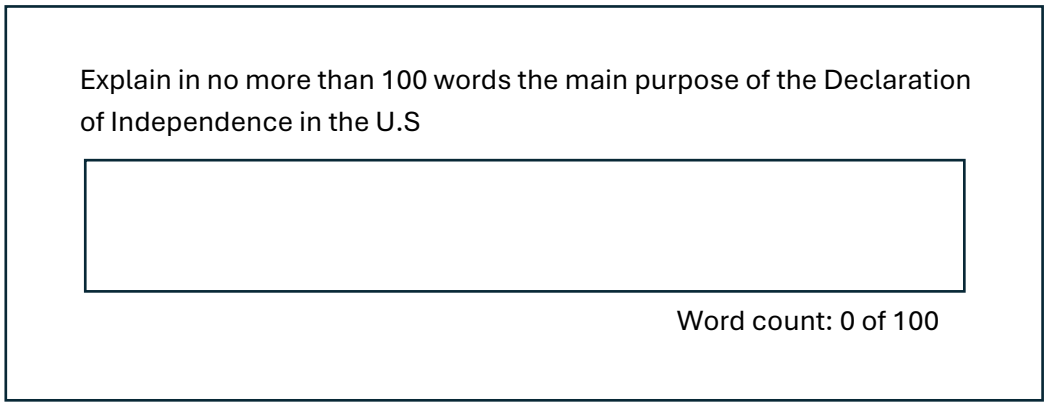
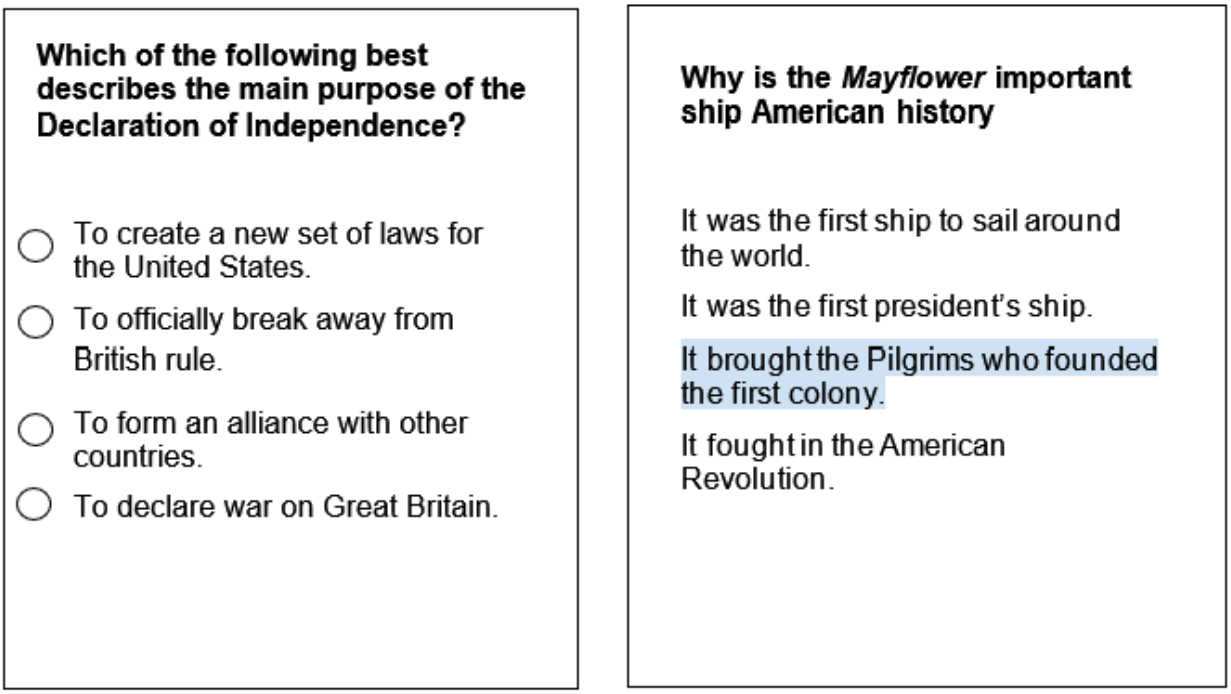


Figure 3.2
An Example of Inconsistent Design



IV.

SEMIOTIC COMPLEXITY IN COMPUTER-ADMINISTERED TESTS: EXAMINING ITEM AND INTERFACE DESIGN IN MAJOR TESTING PROGRAMS

Eunjung Myoung, Xinman Liu, Maria Araceli Ruiz-Primo & Guillermo Solano-Flores
Stanford University

Paper Presented at the Annual Meeting of the National Council on Measurement in Education, Denver, Colorado, April 25th, 2025

Authors' Note: The investigation here reported was possible thanks to generous funding from *Goodnotes*. We are also grateful to Hsiaolin Hsieh, who provided valuable comments at an early stage of the investigation. The opinions expressed are not necessarily those of the funder or our colleague.

NOTE
The format of the original paper presented at the NCME conference has been modified to ensure format consistency with the rest of the report. The references cited in this paper appear in the References section along with other references cited in the report.

Abstract

We address optimal design as critical to minimizing unnecessary cognitive load during test taking and, consequently, minimizing score error variance. In computer-administered testing, students select options, draw lines, generate text, and execute many other functions that involve interpreting signifiers (e.g., icons, labels, and menus) and performing actions (e.g., clicking, dragging-and-dropping, and typing). We examined the functions, signifiers, and actions in science, mathematics and English language arts items from three major large-scale testing programs. We observed disproportionately higher frequencies of option selection over other functions. We observed inconsistencies in the design of items of the same type (e.g., multiple-choice) within each program. Our findings indicate that: (1) digital technology resources are under-utilized, as they could be used to capture a wider variety of student responses; (2) the design of items and testing platforms is not optimal; (3) improved methodologies for systematic, principled computer-based testing design are needed to minimize unnecessary cognitive load during test taking—an important source of error variance.

Introduction

The rapid transition from paper-and-pencil to digital (computer- or tablet-based) modes of administration in large-scale testing increases cost-efficiency but appears to neglect the validity of performance measures. While there is some literature examining the exchangeability of paper-

and-pencil and digital modes of administration (Bugbee, 1996; Neuman & Baydoun, 1998; Scrimgeour & Huang, 2022), little is known about the ways in which mode of administration interacts with other important factors such as the content area assessed and the examinee's level of digital literacy (Reddy et al., 2020). Furthermore, despite the possibilities offered by new digital devices to capture different forms of performance (Perry et al., 2022), multiple-choice continues being the overwhelmingly dominating type of item used by large-scale testing programs. Since different types of items tap into different forms of knowledge (Martinez, 1999, Ruiz-Primo et al., 2009), fact recognition and declarative knowledge may be overrepresented with respect to other, higher-order thinking skills in large-scale testing.

The transition to digital modes of test administration appears to be occurring without adequate consideration of its inherent complexities. For example, when students take computer-administered tests, they need to select options, draw lines, generate text, and execute many other functions. This requires aptly interpreting signifiers such as icons, labels, and menus, and performing a wide variety of instrumental actions such as clicking, dragging-and-dropping, and typing. Such complexities are a source of construct-irrelevant variance when cognitive and technical demands unrelated to the academic skills being assessed, compromise the validity of interpretation of test scores.

The investigation reported here is part of a research agenda committed to optimizing the design of digital testing platforms. We examined, from the perspective of interface design, the information students need to possess and the actions they need to take to navigate testing platforms and to respond to items in computer-administered items. Specifically, we examined the functions, signifiers, and actions involved in responding to items in the testing platforms of three major testing programs.

Conceptual Framework

Our investigation is informed by three theoretical perspectives: social semiotics, cognitive theory, and psychometrics. The perspective of social semiotics (Kress, 2006; the study of how meaning is represented and interpreted according to social conventions), allows examination of identifiers (e.g., icons, labels, menus) which students need to interpret to navigate a testing platform, gain access to the content of items, and respond to those items (Solano-Flores, 2021). The perspective of cognitive theory and design allows examination of cognitive load (Sweller et al., 1998) in problem solving, especially in digital devices (Skulmowski & Xu, 2022). Excessive extraneous cognitive load may result from complex or inconsistent design, as students may need to use part of their working memory not only to reason about the task at hand but also to figure out how to enter their responses.

The perspective of psychometrics alerts us about score variance due to student differences in knowledge and skills that are irrelevant to the content assessed (Haladyna & Downing, 2004). In digitally administered tests, student score differences may be due to not only the differences in the knowledge and skills assessed but also the differences in the ability to navigate the testing platform.

Drawing from these three perspectives, we identify three key concepts—function, signifier, and action—as critical to optimal test design. *Functions* are the things that students can do when they take a test. *General user interface functions* allow students to navigate the testing platform (e.g., moving from one item to the next or to a previous one, activating or deactivating a calculator, scrolling down or up on the screen, writing, typing, deleting, and highlighting); they are typically fixed and do not vary substantially across items of the same item type (e.g., multiple-choice) or content. *Item-specific functions* allow students to access the content of an item and to provide their response to that item (e.g., by selecting an option, entering a number, or typing text). Because they are sensitive to different forms of performance, different types of items tend to have different sets of functions.

Signifiers are affordances that indicate possible uses of the testing platform, and possible actions students can take to navigate the platform and respond to items. A signifier invites the test taker to act on it to execute a function. Signifiers can be visual (e.g., icons such as the image of a calculator), text (e.g., a label), auditive (e.g., a chime, a voice giving an indication), or haptic (e.g., the vibration of a joystick). Icons and buttons are signifiers commonly used in digital testing platforms to indicate affordability. For example, an eraser representing the *delete* function may appear with one of three different background colors: white to indicate the function is not available, gray to show the function is available but inactive, and green to signal that the function is both available and active. Different presentations of the same icon serve as distinct signifiers, each conveying a different meaning.

Actions are manipulations that allow the test taker to execute the functions necessary to navigate the testing platform and to gain access to the content of items and respond to them. For conceptual purposes, it is important to distinguish between *virtual actions*—which reside on the computer screen—and *instrumental actions*—which take place in the real world. Virtual actions are performed through instrumental actions. For example, highlighting a sentence in the text displayed on the screen is a virtual action that requires first, clicking a button in the mouse, then moving the mouse while pressing that button, and then releasing the button.
The following example illustrates how the concepts of function, signifier, and action work together: Computing with a calculator is a function available in a testing platform. The availability of this function and whether it is active is indicated by the calculator icon and its background color (signifier). The actions performed to activate or terminate the computing action are performed by clicking on the icon.

We propose three notions as key to analyzing optimality in the design of digital testing platforms. First, a testing platform's design can be examined based on the sets of functions students need to execute to take a test and the signifiers and actions involved in such functions. Second, a function is executed through a combination of one or more signifiers and one or more actions. Third, a testing platform's design is optimal to the extent that it minimizes the number of signifiers and actions needed to execute functions and to the extent that it consistently involves the same set of functions, signifiers, and actions across items of the same type. We contend that *optimal design* ultimately minimizes unnecessary cognitive load during test taking, thereby minimizing construct-irrelevant score variance.

Research Questions

We formulated two research questions:

1. *What are the commonalities and differences in the sets of functions, signifiers, and actions used in major large-scale testing programs' testing platforms?*

2. *How optimal is the design of these testing platforms?*

Methods

Item Sample. We examined a sample of 660 computer-administered items obtained from the public item releases of three large-scale testing programs platforms—National Assessment of Educational Progress (NAEP; National Center for Education Statistics, n.d.), the Programme for International Student Assessment (PISA; Organisation for Economic Co-operation and Development, n.d.), and the Smarter Balanced Assessment Consortium (SBAC; Cambium Assessment, Inc., n.d.) —administered between 2015 and 2024 across different content area (Table 4.1). (Note 1) In this item count, both stand-alone items and items that were part of item bundles are regarded as separate items. The items belonged to three item type categories: Multiple-Choice, Open Response, and Plotting/Adding, whose percentages varied, respectively between 65% and 69%; 27% and 34%; and 0% and 4% across the item subsamples from the three testing programs.

Coding

We developed a comprehensive coding catalog of the different *functions*, *signifiers*, and *actions* identified in a sub-sample of items from the three testing programs. Then we developed a dichotomous coding system to document the presence or absence of the different functions, signifiers, and actions at both the general interface and item levels. Specifically, while responding to each of the 660 items, we coded (1-0) the functions available to navigate the platform and to respond to the item and the signifiers and actions involved.

Figure 4.1 illustrates the structure of the coding categories used. Two researchers coded the same subset of items independently and resolved any coding discrepancies. However, inter-coder consistency in this form of coding is not an issue, as it does not depend on subjective judgment and is based on examining all the possible sequences of actions taken by the student in navigating the testing platform or responding to an item (Solano-Flores & Martinez, 2023). Figure 4.2 provides an example item from NAEP testing interface and the examples of signifier coding. In this case, the item type is "Multiple-Choice" and the sub-type is "Select an Option from Multiple Options". The item-specific function is "Selection" and the required action is "clicking on a bubble".

Data Analysis

As seen in Table 4.1, the three testing programs focus on different sets of content areas, and target different sets of grades or ages, for different content areas. In addition, they are administered with different periodicities and release different numbers of items. These conditions

make fully crossed designs (e.g., content area x grade x testing program) unattainable. Given these limitations and the scope of our investigation, our analyses focused on the frequencies of functions, signifiers, and actions observed, regardless of content area and grade.
At the general user interface level, we examined the frequencies of different types of items and the frequencies of different functions and their signifiers in the three testing platforms. At the item level, we examined the frequencies of *functions*, and the *virtual actions* involved in responding to items by item type within each testing platform. For simplicity, we refer to these virtual actions simply as *actions*.

Results

To respond to our research questions, we report the frequencies of items of different types in the three testing platforms and the frequencies of functions and signifiers observed at both the user interface and item levels. Then we discuss the optimality of the testing platforms according to the principles stated in our conceptual framework.

Item Types and Sub-Types Across Testing Programs

Table 4.2 provides information about the frequency and percentages of the different types of items identified across the sample of 660 items. Most items were multiple-choice, followed by open response and plotting /adding items (respectively 67%, 31%, and 2% on average across the three testing programs).

Our examination of the items revealed that the testing programs have different sub-types (variations) of multiple-choice items. Five sub-types of multiple-choice items are common in these three large-scale testing programs. “Select an Option from Multiple Options” was the most frequent sub-type of multiple-choice item observed in the three testing programs (42% on average). Beyond this dominant sub-type, each testing program had different varieties of multiple-choice items, with “Change the Location,” “Matrix Table Choice,” and “Multiple Select,” as the second most frequent sub-type respectively for NAEP, PISA, and SBAC. (Note 2) Only SBAC included “Select Sentence(s)” items. Also “Dropdown Choice” and “Binary Choice” items are only in PISA.

Similarly, we identified three types of Open-Response items. “Short-Answer” and “Essay” are common across the three testing programs. The “Create a Row and Select” sub-type was observed only in PISA’s testing platform. To answer this item type, test taker should first run the simulation under different conditions to generate data. Then, click on two rows of data in the table that best support the answer — the selected rows will be marked with a star. The least frequent type of item was “Plotting/Adding,” observed in both NAEP’s and SBAC’s testing platforms, especially that SBAC’s testing platform has a “Adding a Shade” sub-type. The detailed screenshot of the item types and sub-types can be found in Appendix 4.A.

General User Interface Functions and Signifiers. Table 3 shows the general user interface functions and signifiers observed. We identified 28 signifiers across the three platforms. A few of them (mostly observed in the SBAC platform) are accompanied by a label. Some signifiers can be assumed to be easily recognized by any audience, as in the case of the eraser for the *delete*

function. In contrast, other signifiers may not be familiar to all users, as in the case of the signifier used for *delete page cursor* function, which does not represent any concrete object. Only three signifiers were common across the three platforms: "*Next,*" "*Back,*" *and "Help."* In the three programs, when the user clicks on the "Help" button, a window pops up that displays the icons and briefly explains how to activate each. More than half of the 28 signifiers observed in the three testing platforms were unique to the NAEP platform (n = 16; 57%). This means that test takers taking NAEP must interpret more signifiers to respond to computer-based items. Four signifiers are common to different pairs of testing platforms: "*Progress Bar*" (PISA and SBAC), "*Timer*" (NAEP and PISA), "*Zoom in,*" and "*Zoom Out*" (NAEP and SBAC). Such signifiers are similar in appearance but not the same.

Item-Specific Functions and Actions. Table 4 shows the frequencies and the percentages of the item-specific functions and actions involved in the three testing platforms. The functions at the item level fall into three categories: Selection, Typing, and Repositioning/ Plotting, which, on average across the three testing platforms account for 64%, 29%, and 7% of all the items. Within the *Selection* category, "*Click on a bubble*" is the most frequent action in the three programs (35%-46%). "*Click on a box*" is substantially more frequent in SBAC's platform than in the other platforms, which can be at least partially attributed to the high frequency of the "Multiple Select" sub-type.

The frequency of *Typing*, as a category, is similar in the three testing platforms (26%-34%). However, the most frequent actions involved in this function are not the same across testing platforms. "*Typing words*" is the only type of "typing" for PISA, the most frequent for NAEP, and the next to least frequent for SBAC. The most frequent type of typing in the SBAC testing platform is "*Type/click on a keyboard on the screen*."

Discussion

According to our conceptual framework, the number of functions (and their signifiers and actions) at both the general user interface and item levels varies considerably across testing programs. Compared to PISA's and SBAC's testing platforms (respectively with five and 11 functions), NAEP's testing platform makes a considerably higher number of functions (28) available for the user.

Such disproportionality may be a reflection that, to some extent, circumstances that are not related to design (for example, the simple fact that the software used allows inclusion of a wide range of functions and signifiers) shape how testing platforms are built. However, the disproportionality may also reflect how different testing programs implement U.S. legislation that mandates the provision of testing accommodations and accessibility resources for students with disabilities and special needs. As a testing program regulated by the OECD (Organization for Economic Cooperation and Development), PISA does not need to comply with such legislation; NAEP and SBAC do. Clearly, NAEP makes most functions for accommodations and accessibility resources equally available to all students, regardless of special education or special needs status. In contrast, in SBAC's testing platform, many accommodations and accessibility resources appear on it only if the student has selected them in a menu, prior to taking the test (SBAC, 2024). Table 3 shows what students see when they take tests on each platform; it does not show the functions (and the identifiers and actions they involve) SBAC makes available for

students with special needs. However, our argument about optimal design still holds: The fact that NAEP's and SBAC's testing platforms are designed with different approaches about the ways in which students with and without special needs are to be supported speaks to the need for a clear set of design for computer-administered tests. Optimal design is not only about the ways in which testing platforms are built; it is also about how effectively assessment frameworks and item specification documents address design issues.

Our study also revealed that, in some cases, within the same testing program, a function does not consistently involve the same set of signifiers across items of the same type. For example, in PISA's testing platform, the *Back* icon is sometimes displayed in gray, indicating that it is inactive-immutable and becomes clickable only after the *Next* icon has been activated. However, in other items, this icon initially is active-mutable but after clicking on it a window pops up labeled "No Previous Item." The identical icon involves different signifiers and actions. Similarly, on SBAC's platform, the *Back* icon appears blurred, indicating that the icon is inactive-immutable and becomes clickable only after activating the *Next* icon. This logic is not observed with the *Zoom out* icon—which is not initially shown in a blurred or inactive state. Another example of this inconsistency is the display of the *Fraction* icon in NAEP's testing platform. This icon appears inactive for some items, whereas in other items it is shown as active for other items. Not displaying icons that are permanently inactive would increase design optimality.

Final Remarks

Despite the wide variety of functions that, in principle, it is possible to execute, given the possibilities offered by current digital technology, the platforms of the three testing programs depend almost exclusively on clicking and typing. This limited variety of item formats restricts the ability of testing programs to capture performance on different kinds of tasks and, therefore, the ability to tap into a wide range of skills in large-scale testing.

While the three testing platforms rely heavily on multiple-choice items, the format used to deliver this type of item varies considerably across items within each testing program. Such inconsistency may unnecessarily increase cognitive load, as the examinee needs to figure out how to execute the same function with different signifiers across items of the same type. Unfortunately, major testing programs' assessment frameworks and item specification documents lack detailed information on the characteristics of different item types and the ways in which they are to be delivered when tests are administered digitally. Without more specific design criteria, item development and design may be shaped by various idiosyncratic factors or by the set of technical limitations and possibilities of the software used to build each testing platform.

Our findings indicate that the design of testing platforms in the three major large-scale testing programs examined is not optimal. Given their prominence as well-established testing programs, it is reasonable to assume that these findings can be generalized to other testing programs; they speak to the need for a principled methodology for systematic test design. Knowledge gained from this investigation contributes to enhancing the design of testing platforms, ultimately minimizing measurement error due to unnecessary complexity. For now, we conclude that next steps in improving the design of testing platforms need to include: (1) careful analysis of how critical different functions, signifiers, and actions are to responding to different computer-

administered items, (2) selection of the best signifiers to use for each function, and (3) consistent use of signifiers and actions across items of the same type.

Notes

Note 1. While SBAC offers released items on its webpage (Smarter Balanced Assessment Consortium., n.d.) the testing interface varies across states. The sample of SBAC items used in this investigation was drawn from the California State's version of SBAC, the California Assessment of Student Performance and Progress (CAASPP).

Note 2. We acknowledge that using multiple-choice items with one correct option and multiple-choice items with more than one correct option may be primarily a matter of the kind of knowledge or skill being assessed, rather than a matter of design. However, the use of the two item sub-types in the same test has implications for test taking, as it may impose an extra cognitive load—an issue that is relevant to test design.

Table 4.1

Sample of Items by Assessment Program, Year, and Content Area

	Assessment Program and Year							Total [n = 660]
	NAEP			PISA			SBAC	
Content Area	2018	2019	2022	2015	2018	2022	2024	
Mathematics			74			13	230	317
Reading			10		31			41
Science	23	27		24				74
English Language Arts							228	228

Table 4.2

Types and Sub-Types of Items by Testing Program: Frequencies and Percentages

Item Types and Sub-types	NAEP Grades 4 & 8 (n = 134)		PISA Age 15 (n = 68)		SBAC Grade 8 (n = 458)	
	n	*%*	*n*	*%*	*n*	*%*
Multiple Choice	***87***	***65***	***46***	***68***	***317***	***69***
Select an Option from Multiple Options	59	44	31	47	161	35
Select an Image	8	6	0	0	6	1
Select Sentence(s)	0	0	0	0	32	7
Multiple Select	4	3	1	1	67	15
Dropdown Choice	0	0	4	6	0	0
Matrix Table Choice	3	2	5	7	36	8
Change the Location	13	10	4	6	15	3
Binary Choice	0	0	1	1	0	0
Open Response	***45***	***34***	***22***	***32***	***123***	***27***
Short Answer	17	13	0	0	105	23
Essay	28	21	18	26	18	4
Create a Row and Select	0	0	4	6	0	0
Plotting / Adding	***2***	***1***	***0***	***0***	***18***	***4***
Plotting a Line	1	1	0	0	10	2
Adding a Point	1	1	0	0	3	1
Adding a Shade	0	0	0	0	5	1

Note. The bold indicates the subtotal frequency and percentage of each type.

Table 4.3

General User Interface Functions and Their Signifiers in The Testing Platforms

Function	Signifiers		
	Testing Program		
	NAEP	PISA	SBAC
Back			
Calculator			
Change Theme			
Context Menu			
Delete Page			
Delete Page Cursor			
Erase			
Erase Cursor			
Fraction			
Hand Cursor			
Help Button			
Highlight			
Highlight Cursor			
Line Reader			
Next			
Pause			
Pencil			
Progress Bar			
Questions Drop-Down			
Read Aloud			
Reading Section			
Save			
Scratch Book			
Scroll Down			
Scroll Up			
Timer			
Zoom In			
Zoom Out			

Table 4.4

Actions by Item-Specific Function in the Testing Programs: Frequencies and Percentages

Item-Specific Function and Actions Involved	NAEP Grades 4 & 8 (*n* = 134)		PISA Age 15 (*n* = 68)		SBAC Grade 8 (*n* = 458)	
	n	*%*	*n*	*%*	*n*	*%*
Selection (Clicking on a…)	***75***	***56***	***46***	***67***	***310***	***68***
bubble	62	46	36	53	161	35
number/image	9	7	1	1	6	1
box	4	3	1	1	103	22
sentence	0	0	0	0	32	7
Line in a drop-down menu	0	0	4	6	0	0
row	0	0	4	6	0	0
an area to shade/ mark	0	0	0	0	8	2
Typing	***45***	***34***	***18***	***26***	***123***	***27***
numbers	17	13	0	0	5	1
words	28	21	18	26	18	4
/Clicking on a keypad on the screen	0	0	0	0	100	22
Repositioning / Plotting	***14***	***10***	***4***	***6***	***25***	***5***
Dragging and dropping (click on object, push, release)	14	10	4	6	25	5

Note. The bold indicates the subtotal frequency and percentage of each type.

Figure 4.1

Taxonomy of Signifiers. Each Taxon's Code is Shown in Parentheses

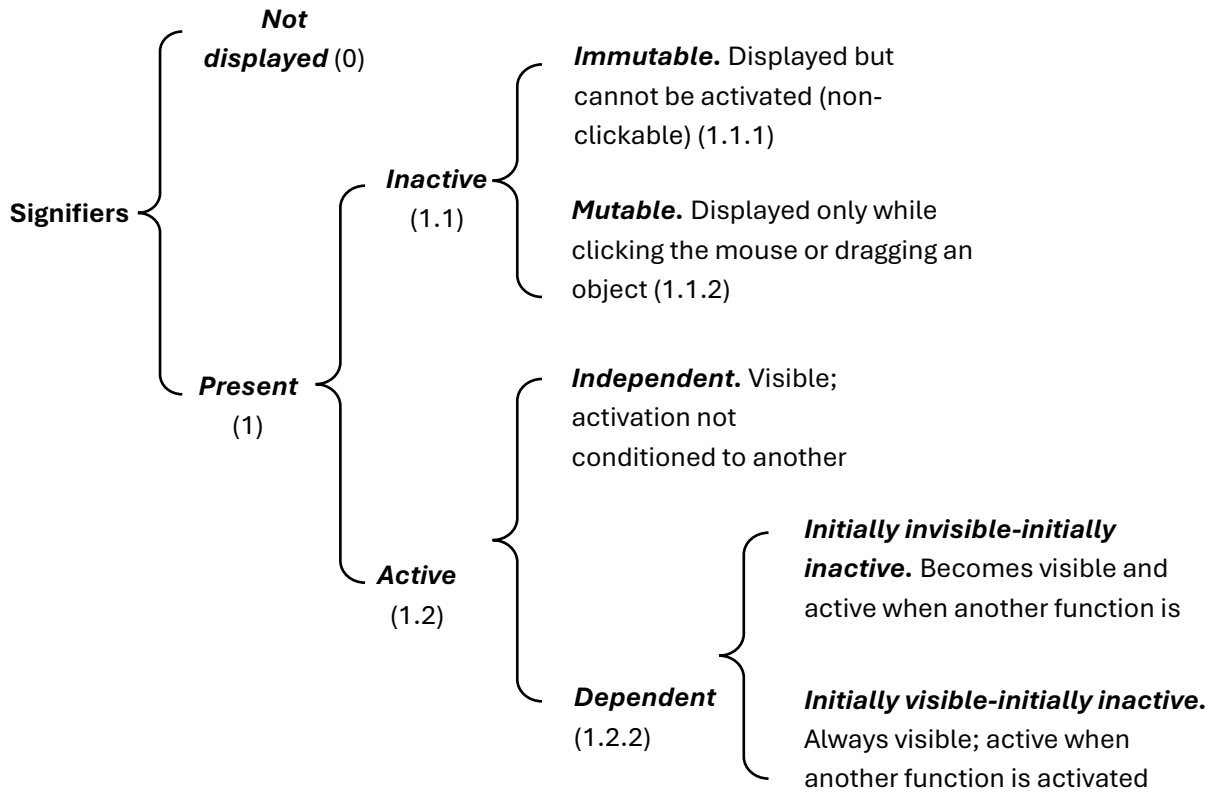
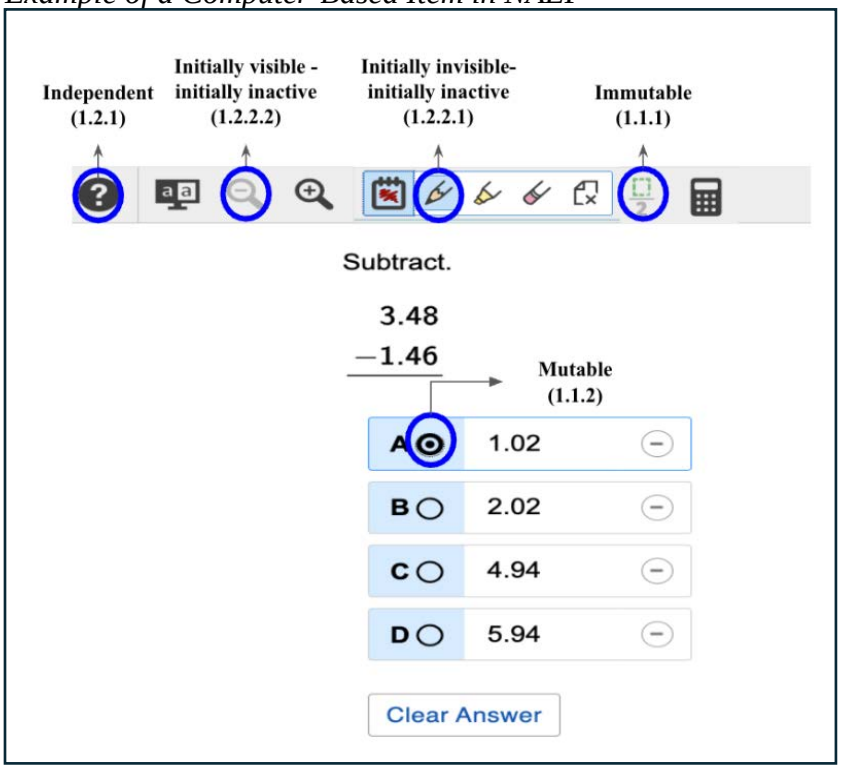


Figure 4.2

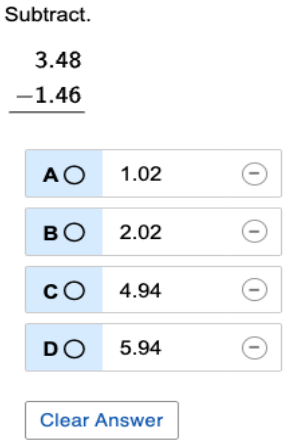
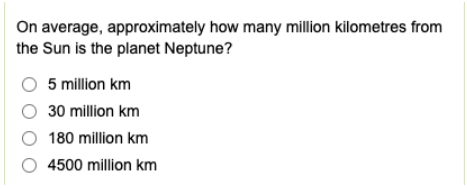
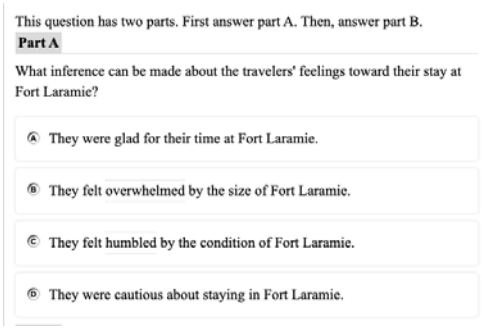
Example of a Computer-Based Item in NAEP



Note. The link for this item: https://cotw.naep.ed.gov/student/grade4/MAT/VH098638/toolbarOn

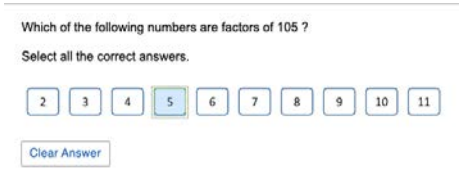
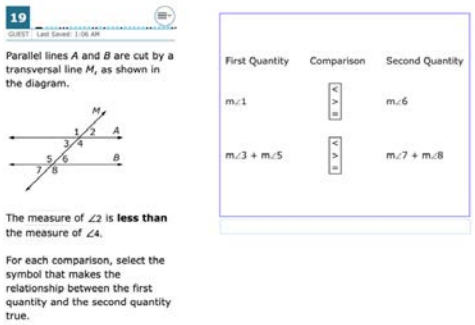
Appendix 4.A

An Example of Item in Each Testing Program Based on Item Sub-types Note. Cases in which item sub-types were not observed in the corresponding testing program are indicated with "–" and their frequencies and percentages are reported as "0" in Table 2.

	NAEP	PISA	SBAC
Select an Option from Multiple Options			

(continues)

(continuation)

	NAEP	PISA	SBAC
Select an Image		-	
Select Sentence(s)	-	-	This question has two parts. First, answer part A. Then, answer part B. **Part A** Click on the statement that **best** provides an inference that can be made about the author's opinion of the Perlman Music Program. A. The Perlman Music Program is limited in its value because of how few students it accepts. B. The Perlman Music Program is most suitable for students who would not do well in a competitive environment. C. The Perlman Music Program offers students a unique opportunity to learn from accomplished musicians in a nurturing setting. D. The Perlman Music Program offers students a unique opportunity to learn new instruments from experienced orchestral musicians.

(continues)

(continuation)

<table>
<tr><th></th><th>NAEP</th><th>PISA</th><th>SBAC</th></tr>
<tr><td>Multiple Select</td><td>Which of the following directions could Paula take to get from her house to school?

Select all the correct answers.

A ☑ 7 blocks south, 3 blocks west
B ☐ 7 blocks south, 3 blocks east
C ☑ 7 blocks north, 3 blocks east
D ☐ 2 blocks west, 7 blocks south, 1 block west
E ☐ 2 blocks east, 7 blocks south, 1 block east

Clear Answer</td><td>Which statements about the golden plover's migration do the maps support?

✓ Remember to select one or more boxes.

☐ The maps show a decrease in the number of golden plovers migrating southward in the past ten years.
☐ The maps show that northward migratory routes of some golden plovers are different from southward migratory routes.
☐ The maps show that migratory golden plovers spend their winter in areas that are south and southwest of their breeding or nesting grounds.
☐ The maps show that the migratory routes of the golden plover have shifted away from coastal areas in the past ten years.</td><td>Select all possible values for x in the equation $x^3 = 375$.

☐ $5\sqrt[3]{3}$
☐ $\sqrt[3]{375}$
☐ $75\sqrt[3]{5}$
☐ $125\sqrt[3]{3}$</td></tr>
<tr><td>Dropdown Choice</td><td>-</td><td>In the table below, answer each question by selecting a country from the corresponding drop-down menu.

<table><tr><th>Question</th><th>Country</th></tr><tr><td>In terms of percentage points, which country had the greatest gain between 2005 and 2015?</td><td>Select</td></tr><tr><td>Which country had no overall change between 2005 and 2015?</td><td>Select</td></tr><tr><td>In terms of percentage points, which country had the greatest loss between 2005 and 2015?</td><td>Select</td></tr></table></td><td>-</td></tr>
</table>

(continues)

<table>
<tr><th></th><th>NAEP</th><th>PISA</th><th>SBAC</th></tr>
<tr><td>Matrix Table Choice</td><td>
Three statements about comparing fractions are shown in the table.

For each statement, determine whether it is True or False.

Make one selection for each statement to show your answer.

<table>
<tr><th>Statements</th><th>True</th><th>False</th></tr>
<tr><td>$\frac{1}{2} < \frac{3}{4} < \frac{5}{6} < \frac{7}{8}$</td><td>○</td><td>○</td></tr>
<tr><td>In two fractions with the same denominator, the fraction with the greater numerator always has a greater value.</td><td>○</td><td>○</td></tr>
<tr><td>In two fractions with the same numerator, the fraction with the greater denominator always has a greater value.</td><td>○</td><td>○</td></tr>
</table>
Clear Answer
</td><td>
Some posts on a forum can be relevant to the topic while some posts are not. Click on either Yes or No to indicate whether the posts in the table below are relevant to Ivana_88's problem.

<table>
<tr><th>Is the post relevant to Ivana_88's problem?</th><th>Yes</th><th>No</th></tr>
<tr><td>NellieB79's post</td><td>○</td><td>○</td></tr>
<tr><td>Monie's post</td><td>○</td><td>○</td></tr>
<tr><td>Avian_Deals' post</td><td>○</td><td>○</td></tr>
<tr><td>Bob's post</td><td>○</td><td>○</td></tr>
<tr><td>Frank's post</td><td>○</td><td>○</td></tr>
</table>
</td><td>
Select True or False to indicate whether each comparison is true.

<table>
<tr><th></th><th>True</th><th>False</th></tr>
<tr><td>$\sqrt{37} < 5\frac{1}{4}$</td><td>□</td><td>□</td></tr>
<tr><td>$3\pi > 3\sqrt{3}$</td><td>□</td><td>□</td></tr>
<tr><td>$\sqrt{5} > \frac{5}{7}$</td><td>□</td><td>□</td></tr>
<tr><td>$\frac{15}{\sqrt{10}} > 8.38$</td><td>□</td><td>□</td></tr>
</table>
</td></tr>
<tr><td>Change the Location</td><td>
Talat has a jar.

She has three questions she needs to answer about the jar.

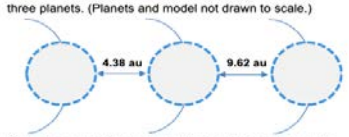
What measurement could be used to answer each of Talat's questions?

Drag a measurement into each box to show your answer.

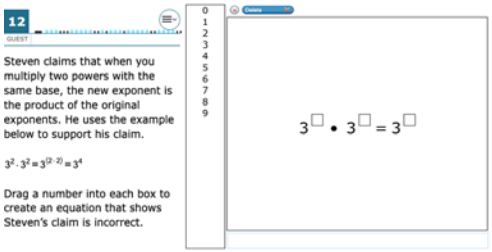
Height | Volume | Weight

How heavy is the jar? | How tall is the jar? | How much liquid can the jar hold?

Clear Answer
</td><td>
The following model shows the average distances between three planets. (Planets and model not drawn to scale.)



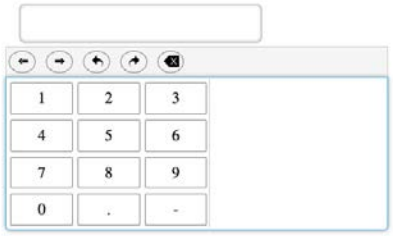
Based on the distances given, which planets belong in the model? Drag the correct three planets in the correct order. To change an answer, first drag the previous planet out.

</td><td>


</td></tr>
</table>

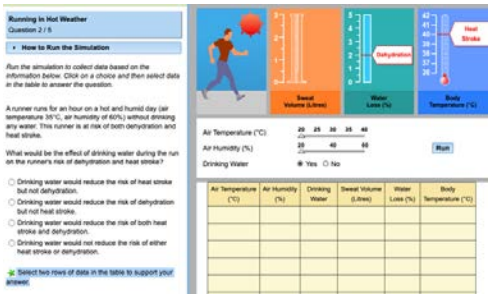
(continues)

(continuation)

	NAEP	PISA	SBAC
Binary Choice	-	**Please read the examples. The correct answers are highlighted. Then click on the NEXT arrow to try some practice sentences.** A. The red car had a flat tire. YES NO B. Airplanes are made of dogs. YES NO C. The student read the book last night. YES NO	-
Short Answer	 What number is represented by point X on the number line?	-	Enter the value of n for the equation $5^n = 5^{11} \cdot 5^3$. 

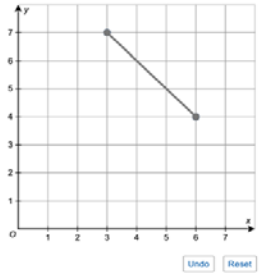
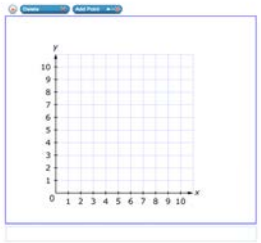
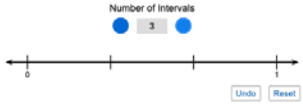
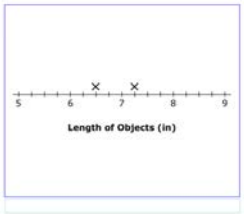
(continues)

(continuation)

	NAEP	PISA	SBAC
Essay	Many cooking pans are made of metal with a plastic handle. Why is metal a good material for a cooking pan? Why is plastic a good material for the handle of a pan?	Identify a factor that might make the volunteers' counts of migrating birds inaccurate, and explain how that factor will affect the count.	
Create a Row and Select	-		-

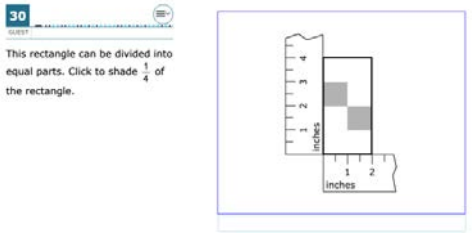
(continues)

(continuation)

	NAEP	PISA	SBAC
Plotting the Line	A line segment is shown on the grid. Create a line segment that is parallel to the given line segment. To show your answer, select two points on the grid.  Undo Reset	-	1 Use the Add Point tool to plot each point on the coordinate plane. Part A: Plot the point (3, 2). Part B: Plot the point (6, 4). Part C: Plot the point (8, 1). 
Adding a Point	Plot a point on the number line that is 0.15 away from 0.75. To show your answer: • Use ⊖ or ⊕ to change the number of intervals of equal size. • Then, select a location on the number line to plot the point.  Undo Reset	-	24 Students pulled classroom objects from a bag and measured them in inches. They used this data to make a line plot. Objects / Length (in) Pen $5\frac{1}{2}$ Scissors $7\frac{3}{4}$ Stapler $7\frac{1}{4}$ Calculator $6\frac{1}{2}$ Notepad $8\frac{1}{4}$ Click above the tick marks to complete the line plot that displays the data. 

(continues)

(continuation)

	NAEP	PISA	SBAC
Adding a Shade	-	-	

V.

IDEAS FOR THE DEVELOPMENT OF A *GOODNOTES* PLATFORM FOR ITEM CREATION

Introduction

During our discussions with *Goodnotes* staff, we identified an important problem not sufficiently addressed by major players in the testing industry—the lack of a robust principled practice for systematic development of items according to pre-established design criteria. We recommended *Goodnotes* to consider the possibility of developing a platform for item creation as a promising way to participate in the testing industry. This section elaborates and illustrates the ideas proposed.

As explained in Section IV, our analysis of items from large-scale testing programs revealed serious inconsistencies in the features of items—a serious potential source of measurement error. For example, two multiple-choice items from the same testing program, grade, and content may have different layouts, involve different sets of signifiers and actions, and have different levels of visual complexity. This variation suggests that the software and procedures used by contractors for large-scale testing programs to develop items do not allow detailed control of format and design features. It is important to mention that large-scale testing programs provide contractors with item specifications documents that prescribe and illustrate the characteristics of the different types of items to be developed. Clearly, such specifications are not sufficiently detailed and leave too much room for interpretation.

Text length is an example of how important format and design features are not sufficiently considered in current item development practice. An investigation on the textual and visual features of a sample of SBAC mathematics items revealed similar textual complexity (as measured by text length) across items of different school grades (Solano-Flores et al., 2023). Since the reading demands of textual material are typically lower for lower grades, this finding indicates that text complexity is not sufficiently addressed by item specifications documents as a source of measurement error that needs to be minimized.

Flowchart Comparative Analysis

The ideas proposed here for the development of a *Goodnotes* platform for item creation originated from our efforts to identify the design limitations of testing interfaces used by major large-scale testing programs. First, we selected two computer-administered SBAC items—an English Language Arts multiple-choice item and a mathematics plotting item. The items are shown in Figures 5.1(a) and 5.1(b).

Second, we created *Goodnotes* mockup versions of Items 1 and 2, as shown in Figures 5.2(a) and 5.2(b).

Third, we constructed general and item-specific user-interface interaction flowcharts for the selected SBAC items and their *Goodnotes* mockup versions. These flowcharts, shown in Figures

5.3 and 5.4, map all possible functions, signifiers, and actions involved in responding to Items 1 and 2 and navigating the interface.

Fourth, we conducted a flowchart comparative analysis (see Solano-Flores & Martínez, 2023) to identify how different were the sets of functions, signifiers, and actions involved in responding to the items administered in the SBAC interface and in the *Goodnotes* interface. As shown in Figures 5.3 and 5.4, the user-interface interaction is more complex both horizontally and vertically in SBAC than in *Goodnotes*.

As shown in Figure 5.5, for Item 1, the *Goodnotes* version has the following improvements:

- *Elimination of redundant signifiers.* In the SBAC interface, two different signifiers (letters in circles and hollow boxes) are available for the same function—select. In contrast, *Goodnotes* uses a single, unified system of signifiers (letters with gray circles) that ensures functional clarity and reduces visual noise.

- *Touch-native optimization.* The *Goodnotes* versions of the items eliminate unnecessary mouse-specific affordances, such as hover states (cursor changes), which do not exist in tablet devices. By removing these redundant signifiers, the interface becomes more intuitive and easier to interpret.

- *Reversible decision-making.* Unlike the SBAC interface, which allows users to switch options without deselecting, the *Goodnotes* version introduces direct deselection. Users can click a selected option again to deselect it (represented by a new green "action" box). This provides students with greater flexibility and a more natural decision-making process.

As shown in Figure 5.6, for Item 2, the *Goodnotes* version has the following improvements:

- *Integrated tool activation.* In SBAC, students must click an "Add Arrow" icon to activate the drawing function. *Goodnotes* eliminates this step. The default pencil tool is always ready for interaction, thus eliminating unnecessary actions from the workflow.

- *One gesture.* SBAC's interface requires a discrete three-step sequence for plotting: Click (point 1) → Drag → Click (point 2). By leveraging *Goodnotes*' built-in "Pencil-to-Shape" technology, the same function is completed with a single continuous gesture. Students click and drag, and the platform automatically ensures geometric accuracy. This shift from vertical (discrete) to horizontal (continuous) actions on the flowchart minimizes unnecessary cognitive load.

- *Unified deletion workflow.* SBAC separates the deletion of lines and dots into two different functions. *Goodnotes* simplifies this through a unified "Eraser" tool. Students can erase any part of a line to remove the entire element, mimicking the familiar action with a real eraser consistently across all item types.

Basic Elements for Item Creation

Based on the results of our comparative flowchart analysis, we provide here ideas that the ESD-Lab *Goodnotes* may be interested in considering to develop a platform for item creation: interface design specifications, item design parameters, and shells.

Interface Design Specifications. Table 5.1 shows the functions, signifiers and actions needed to create any type of test item using a *Goodnotes* platform. These are the only nine functions that the ESD-Lab thinks are needed for examinees to effectively interact with the iPad *Goodnotes* testing platform and successfully access test items.

Item Design Parameters. Table 5.2 provides a list of textual and visual characteristics of items administered with the iPad *Goodnotes* testing platform. This list is far from being complete; it is intended to show how important textual and visual features of test items and the testing interface can be systematically pre-determined.

Templates. The *Goodnotes* platform for test creation can be supported by templates (shells). Templates or shells can function as "programming environments for developers" and "conceptual tools that help assessment developers communicate effectively (Solano-Flores et al., 2001, p. 48). For the purposes of item design, a template can be formally defined as the hollow structure of an item of a given type that is filled out by the developer with the textual and visual content of an item. A template ensures that the layout of the item and its format have the values previously selected for a series of design parameters such as text length, font size, font style, location of illustration, illustration size, color, gray tone range, text box location, table location, and color, among many other features. Specifically, a template provides: (1) formal specifications of structural properties for tablet-administered tests, (2) authoring environments that standardize user interfaces across different item formats, and (3) regulatory frameworks that guide the systematic development of tablet-based platforms.

Templates can be created to process some design features automatically. For example, any text in the stem of a multiple-choice item entered by the test developer will be automatically displayed in Times New Roman font style and font size 12, if those are the values selected for the font style and font size design parameters. Other design features require direct interaction with the developer according to certain design constraints. For example, the text entered by the developer to fill out the box for a multiple-choice item's stem is rejected if it has more than three sentence or more than 30 words, or if it uses passive voice, if those are the values selected for text length and grammar design parameters.

Figures 5.7(a), 5.8(a), and 5.9(a) show respectively examples of three types of items and their corresponding templates: multiple-choice with text stem and four text options; multiple-choice with text stem and four illustration options; and open-ended with text and illustration stem and four text options. Figures 5.7(b), 5.8(b), and 5.9(b) show templates for the corresponding types of items, as they would appear in the *Goodnotes* item creation platform. These templates are only three of the many possible templates that the platform could contain. They are intended to show how all items belonging to the same type can be created with the same format and design features.

The templates proposed for *Goodnotes* to use have a rectangular layout consistent with the shape of an iPad screen and are organized according to a grid of 30 rows and 22 columns. While large-

scale testing programs such as SBAC, NAEP, and PISA rely on a landscape orientation designed for mouse-based navigation, the *Goodnotes* portrait-oriented, stacked layout platform has the advantage of mirroring the standard paper-and-pencil format, which produces a more intuitive user experience. For each type of item, the orange lines demarcate the zones in the grid that should contain different components such as item stem, options, illustrations, open-ended response box, etc. The templates already include three operation/navigation buttons: Push to respond/Push to undo (toggle), Zoom in/ Zoom out (switch), and Back/Next (switch).

Table 5.1

Functions of Items Administered with Goodnotes

Type of function	Function	Instrumental actions and signifier / operandum	Formats
Operation	1. Enable / undo response	Touch toggle button "Push to Respond" / "Undo"	1 "Push to Respond" label on white. Response is enabled when touched. Turns to "Push to Undo" on pink. 2. "Undo" label on pink. Response is erased when touched. Turns to "Push to Respond on white.
	2. Increase / decrease font size	Touch on Zoom in (left side) / normal (center) / or zoom out button	Size of item is reduced, increased, or put back to original form respectively by touching the left side, the center, or the right side of the button.
Navigation	3. Move back or forth	Touch "Previous" end of button or "Next" end of button	Previous item or the next item displayed.
Responding	4. Selection	Touch one of several radio buttons	Enabled when "Push to Respond" button is touched. Response is provided by touching one radio button.
	5. Drawing	Touch and slide fingertip / move stylus on screen	Enabled when "Push to Respond" button is touched. Response is provided by drawing a line.
	6. Typing	Type on screen keyboard	Enabled when "Push to Respond" button is touched. Screen keyboard is displayed. Response is provided by typing. Screen keyboard disappears when student touches screen outside.
	7. Writing letters or symbols	Touch and slide fingertip / move stylus on screen	Enabled when "Push to Respond" button is touched. Response is provided by writing on screen with finger or stylus.
	8. Repositioning / rotating	Touch and slide fingertip on screen	Enabled when "Push to Respond" button is touched. Response is provided by sliding the fingertip on the screen.
	9. Plotting	Touch and slide fingertip / move stylus on screen	Enabled when "Push to Respond" button is touched. A dot is created at the point of touch; continued dragging draws a continuous line until release.

Table 5.2

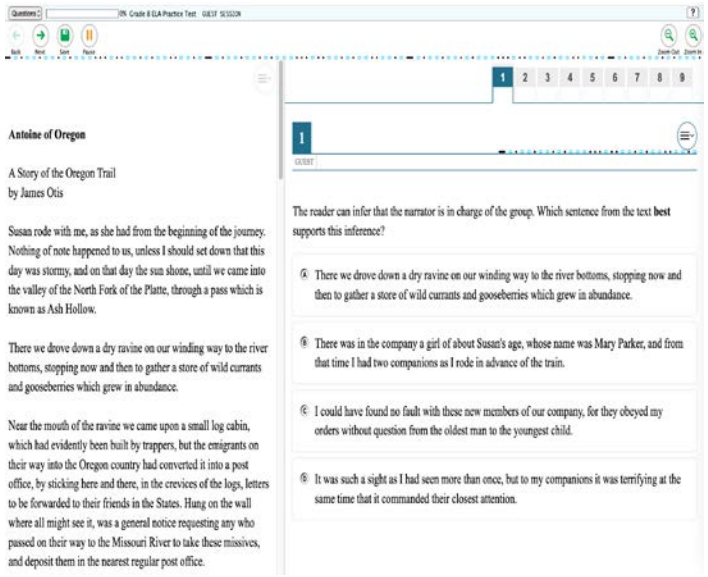
Textual and Visual Characteristics of Multiple-Choice Items Administered with Goodnotes: Examples

Feature	Variable	Value
Stem	Framing	No
	Text font	Times New Roman, plain
	Text size	12 pt
	Text color	Black
Illustrations	Framing	Yes
	Background	No
	Colors	Black and white
Response box	Framing	Yes
	Font	Calibri
	Size	11 pt, plain
	Color	Black
	Weight	1 pt
	Box line: Color	Black
Working space	Framing	Yes
Multiple choice buttons	Colors	Non-selected in white; Selected in black
Options	Framing	Style
	Style	A), B), C), D)
	Font	Aptos 11 pt, bold
	Selection modality	Radio button, left to option letter
	Exclusivity	Mutually exclusive; clicking on one option erases previous option

Figure 5.1

Screenshots of Two SBAC Items

(a) Item 1
Grade 8, English Language Arts, Multiple Choice: Select one option from four items



(b) Item 2
Grade 3, Mathematics, Plotting/ Adding: Plotting a Line

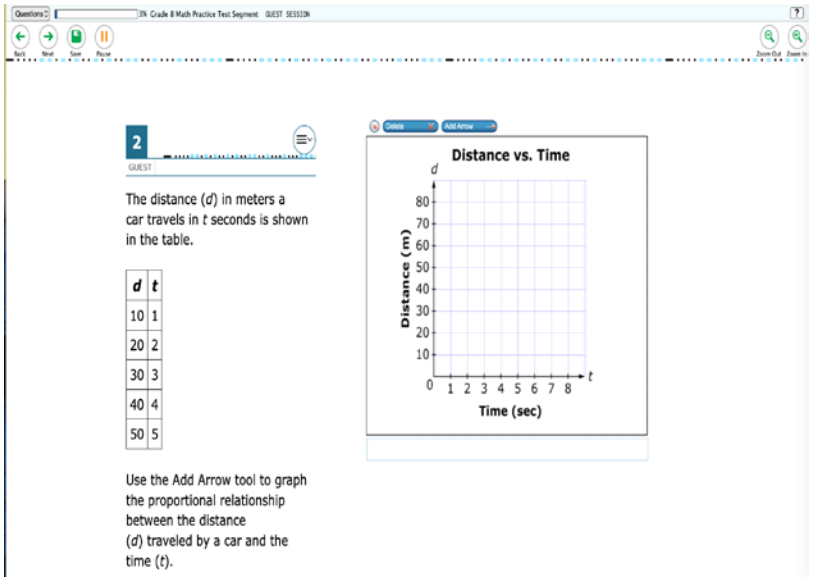
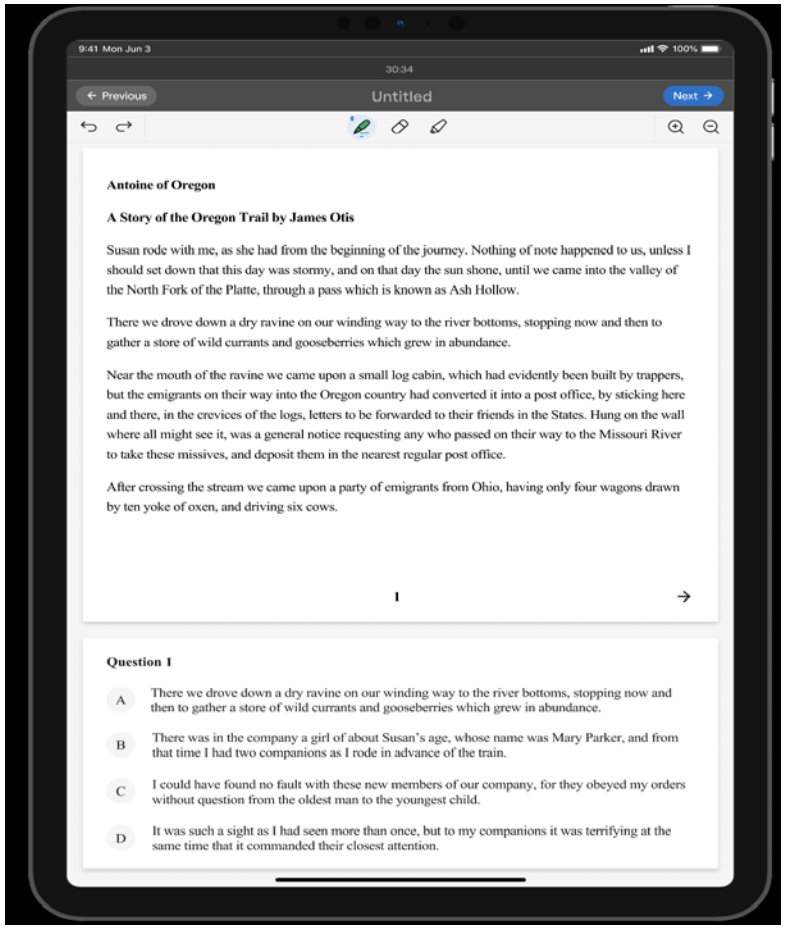


Figure 5.2

Screenshots of Two Goodnotes Items: Mockups. Hyperlinks to the Interactive Versions Shown in Blue.

(a) **Item 1**

Grade 8, English Language Arts, Multiple-Choice (Select one Option from Four Options)



(b) **Item 2**

Grade 8, Mathematics, Plotting/Adding: Plotting a Line

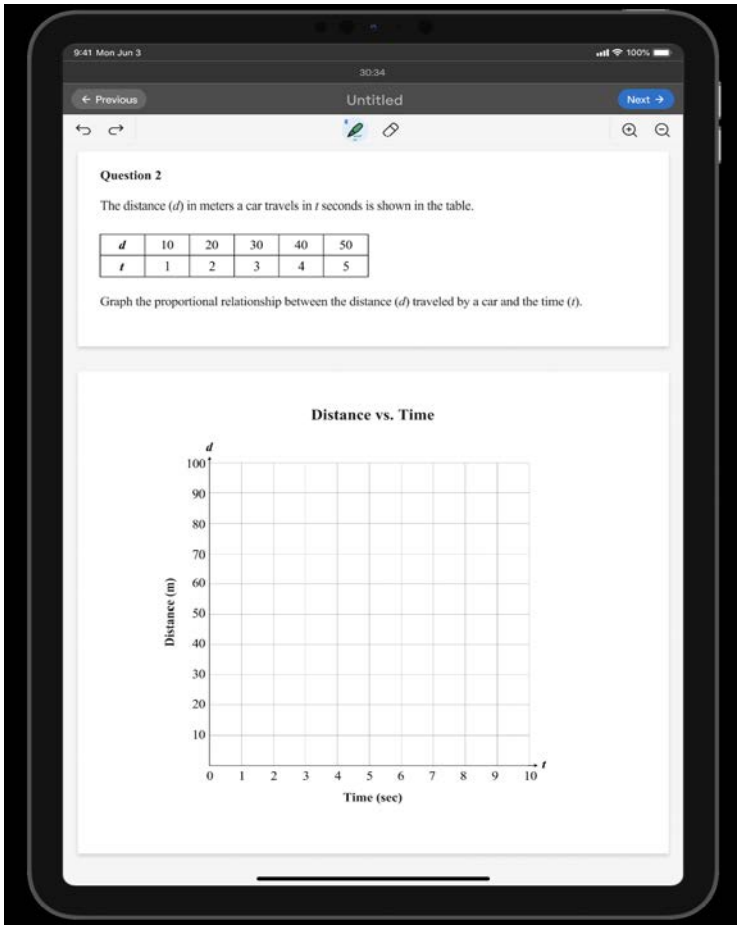


Figure 5.3

User-Interface Interaction Flowchart: Item 1 (Grade 8 - ELA - Multiple Choice). Functions, Signifiers, and Actions Shown Respectively in Pink, Orange, and Green. General and Item-Specific User Interfaces Shown in the Upper and Lower Parts, Respectively.

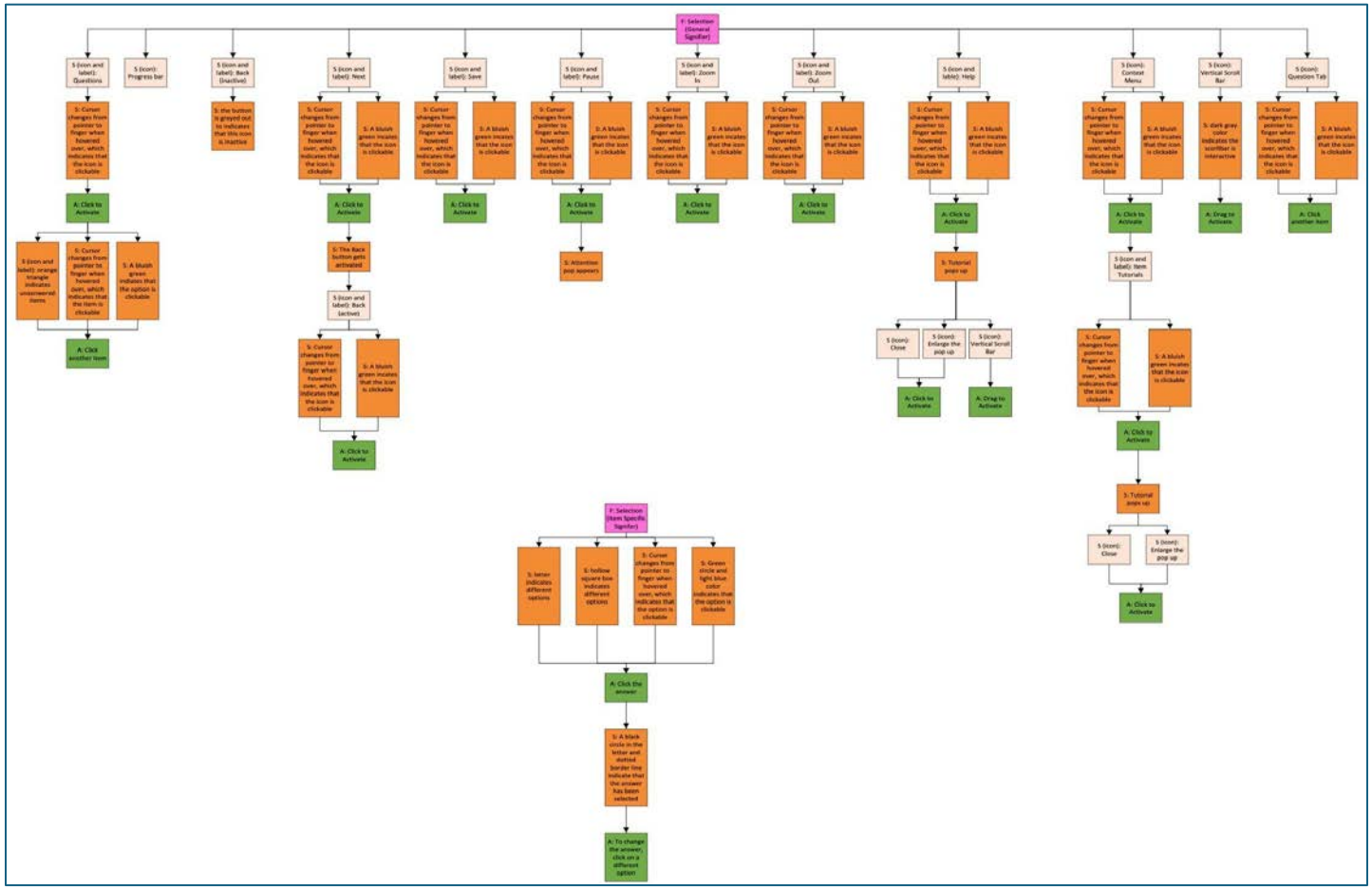


Figure 5.4

Student-Interface Interaction Flowchart: Item 2 (Grade 8 - Math - Plotting/ Adding: Plotting a Line). Functions, Signifiers, and Actions Shown Respectively in Pink, Orange, and Green. General and Item-Specific User Interfaces Shown in the Upper and Lower Parts, Respectively.

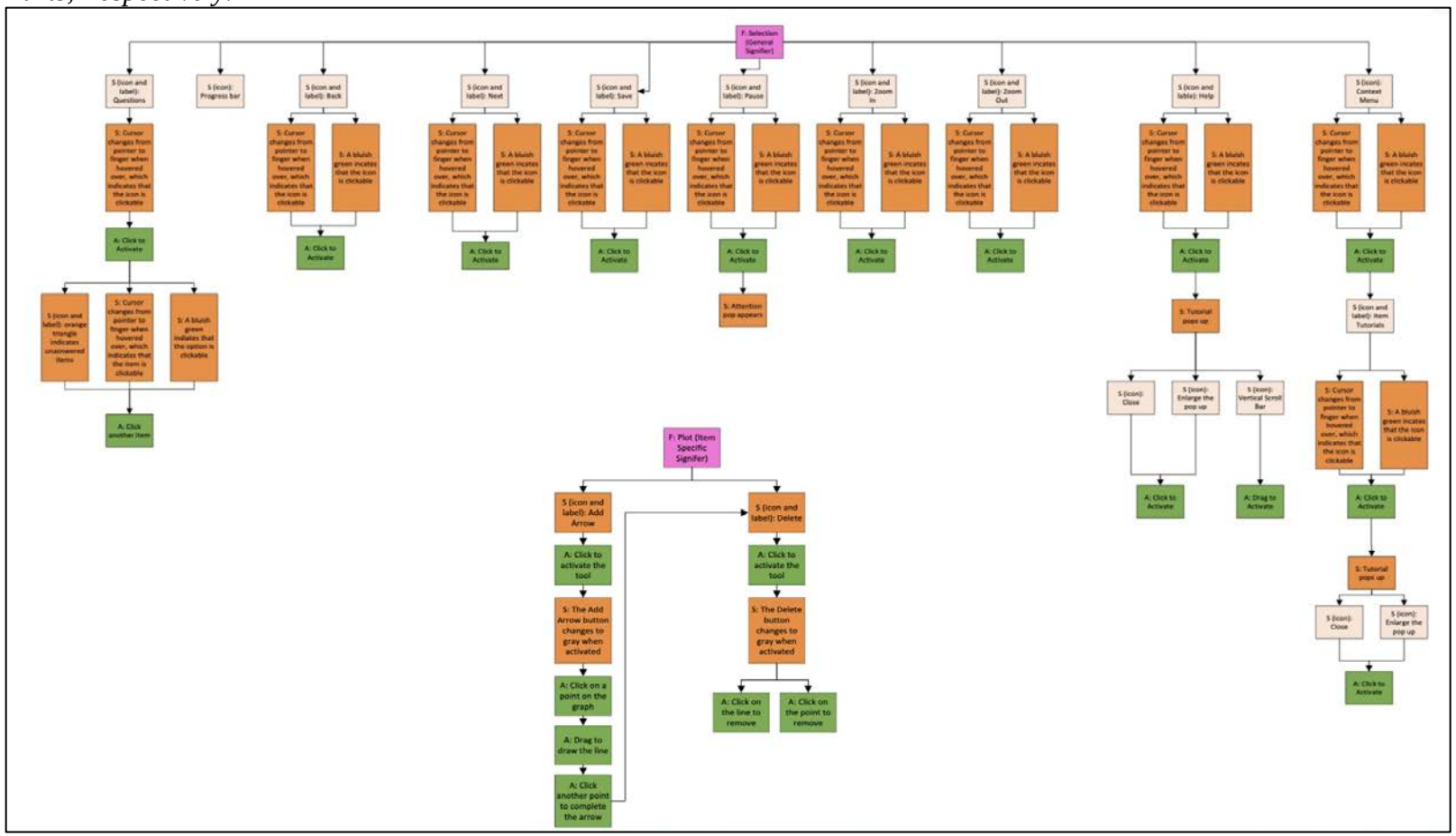


Figure 5.5

SBAC and Goodnotes User-Interface Interaction Flowcharts for the Same Multiple-Choice Item: Item-Specific Interface for Grade 8, ELA, Select One Option From a Set of Four Options. Functions, Signifiers, and Actions, Shown Respectively in Pink, Orange, and Green.

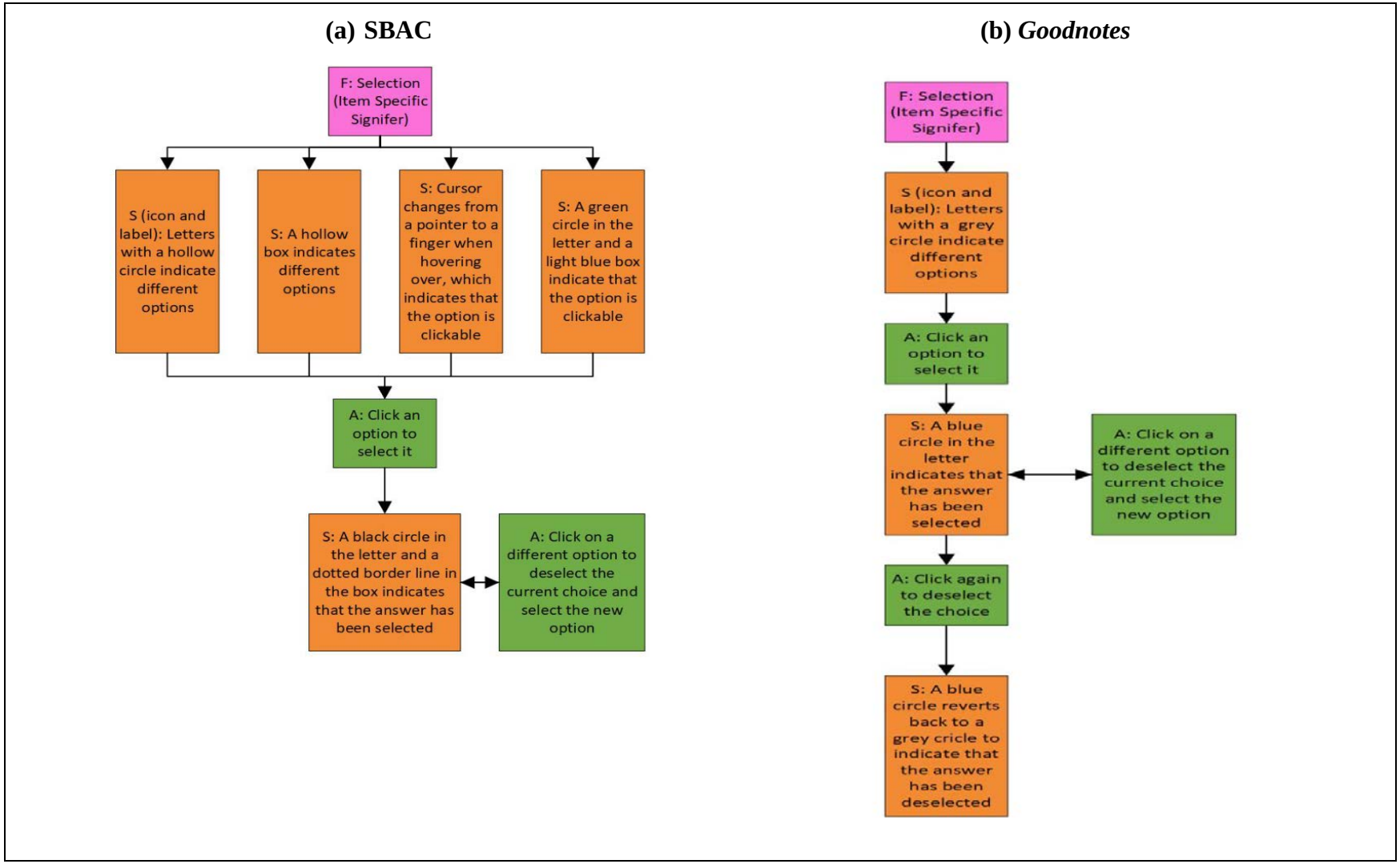


Figure 5.6

SBAC and Goodnotes User-Interface Interaction Flowcharts for the Same Constructed-Response Item: Item-Specific Interface for Grade 8, Math, Adding/Plotting: Plotting a Line. Functions, Signifiers, and Actions, Shown Respectively in Pink, Orange, and Green.

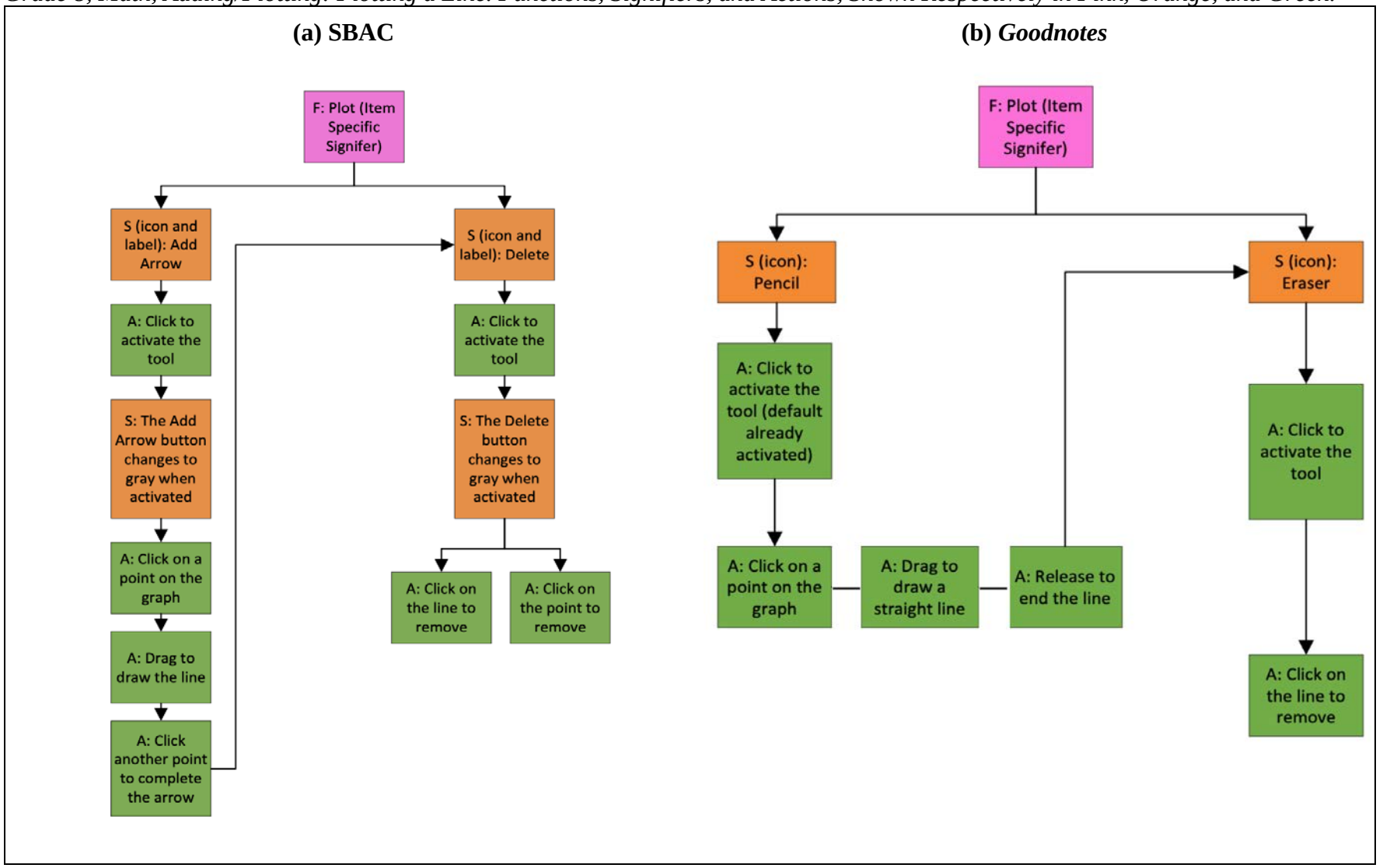


Figure 5.7

Multiple Choice with Text Stem and Four Text Options: Item and Template

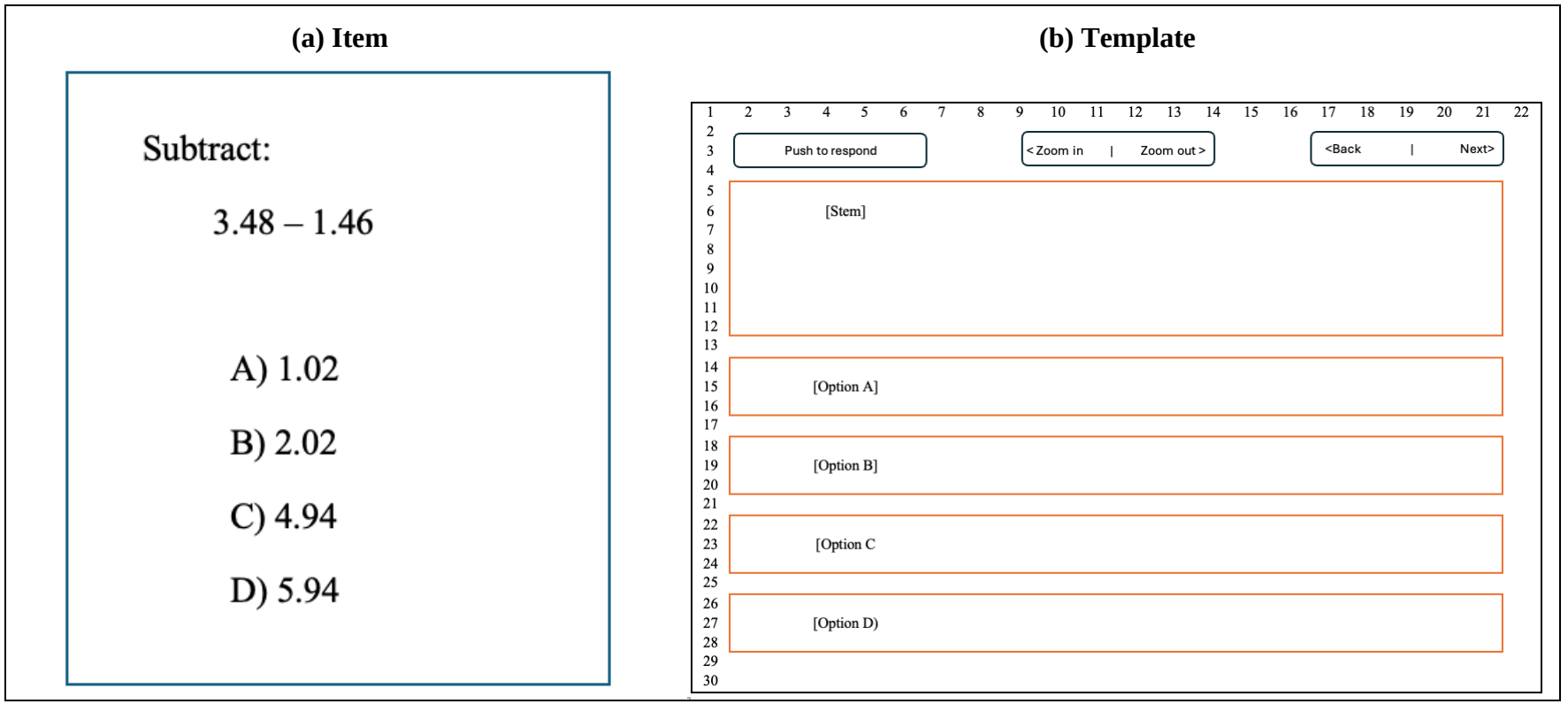


Figure 5.8

Multiple-choice with Text Stem and Four Illustration Options: Item and Template

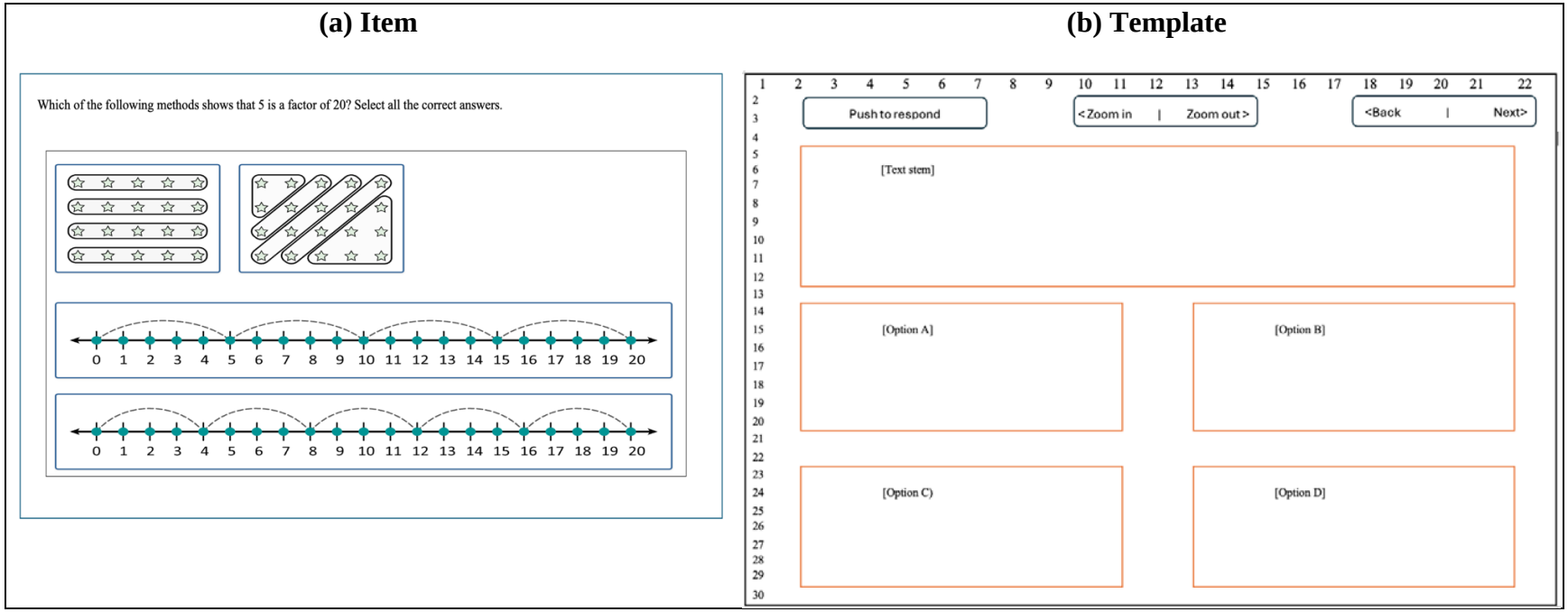
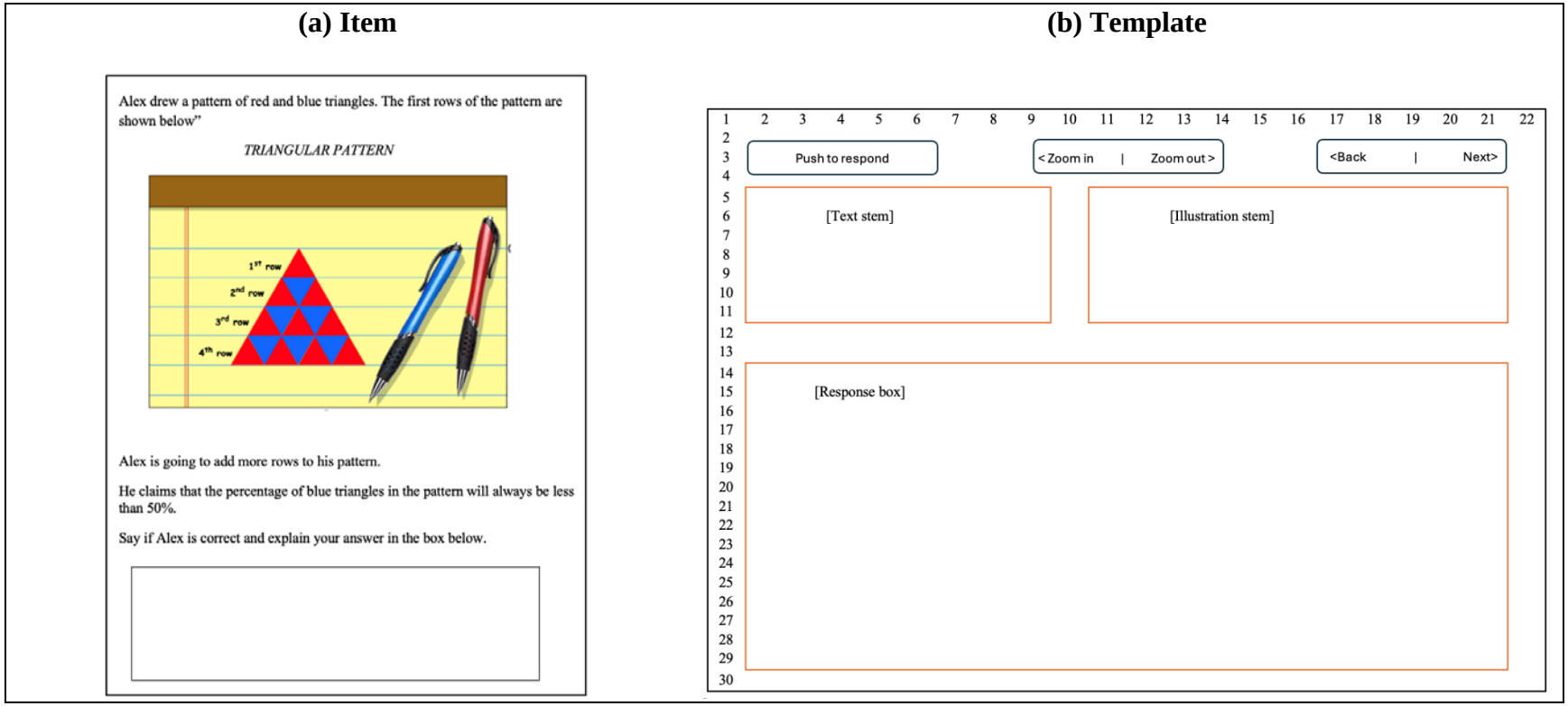


Figure 5.9

Open-Ended Item with Text-and-Illustration Stem and Four Text Options: Item and Template



VI.

REFERENCES

Bugbee, A. C. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education, 28*(3), 282–299. https://doi.org/10.1080/08886504.1996.10782166

Cambium Assessment, Inc. (n.d.). *Student login page*. Retrieved April 1, 2025, from https://login5.cambiumtds.com/student_core/V147/Pages/LoginShell.aspx?c=California_PT&a=Student

El-Hashash, A. (2022). Role of digital formative assessment in improving the assessment and monitoring of students' learning and their significance during the COVID-19 Pandemic. *Open Journal of Educational Research, 2*(1), 9–12.

Gibson, J. J. (1979). *The theory of affordances*. In *The ecological approach to visual perception* (pp. [page range of the chapter]). Houghton Mifflin.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Kress, G., & Leeuwen, T. V. (2006). *Reading images: The grammar of visual design, 2nd edition*. Routledge.

Li, F., Cheng, L., Wang, X., Shen, L., Ma, Y., & Islam, A. Y. M. A. (2025). The causal relationship between digital literacy and students' academic achievement: A meta-analysis. *Humanities and Social Sciences Communications, 12*(1), 1–12. https://doi.org/10.1057/s41599-025-04399-6

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207-218.

National Center for Education Statistics. (n.d.). *NAEP questions tool*. Retrieved August 9, 2024, from https://www.nationsreportcard.gov/nqt/searchquestions

Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement, 22*(1), 71–83. https://doi-org.stanford.idm.oclc.org/10.1177/01466216980221006 (Original work published 1998)

Norman, D. A. (2013). *The design of everyday things* (Revised and expanded ed.). Basic Books.

Organisation for Economic Co-operation and Development. (n.d.). *PISA test*. Retrieved August 9, 2024, from https://www.oecd.org/en/about/programmes/pisa/pisa-test.html

Perry, K., Meissel, K., & Hill, M. F. (2022). Rebooting assessment. Exploring the challenges and benefits of shifting from pen-and-paper to computer in summative assessment. *Educational Research Review, 36*, 100451. https://doi.org/10.1016/j.edurev.2022.100451

Reddy, P., Sharma, B., & Chaudhary, K. (2020). Digital Literacy: A review of literature. *International Journal of Technoethics, 11*(2), 65–94. https://doi.org/10.4018/IJT.20200701.oa1

Ruiz-Primo, M. A. (2009). *Towards a framework for assessing 21st Century science skills*. Commissioned paper for The National Academies. February.

Scrimgeour, M. B., & Huang, H. H. (2022). A comparison of paper-based and computer-based formats for assessing student achievement. *Mid-Western Educational Researcher, 34*(1), 69–92.

Skulmowski, A., Xu, K.M. (2022). Understanding cognitive load in digital and online Learning: A new perspective on extraneous cognitive load. *Educational Psychology Review, 34*(1), 171–196. https://doi.org/10.1007/s10648-021-09624-7

Smarter Balanced Assessment Consortium (2024). *Usability, accessibility, and accommodations guidelines*. Prepared with the assistance of the National Center on Educational Outcomes. June 27. The Regents of the University of California.

Smarter Balanced Assessment Consortium. (n.d.). *Sample items*. Retrieved August 22, 2024 from https://sampleitems.smarterbalanced.org/

Solano-Flores, G. (2021). The semiotics of test design: Conceptual framework on optimal item features in educational assessment across cultural groups, countries, and languages. *Frontiers in Education, 6*. https://doi.org/10.3389/feduc.2021.637993

Solano-Flores, G., & Martinez, R. (2023). Trans-semiosis and fairness in the design of testing accommodations/accessibility resources for students with special needs. Technical report submitted to the Learning Differences Initiative and the Stanford Accelerator for Learning. June 1.

Solano-Flores, G., Shyyan, V., Chía, M., & Kachchaf, R. (2023) The design of mathematics testing accommodations for second language learners: Semiotic exchangeability of translation and illustration pop-up glossaries in computer-administered tests. *International Multilingual Research Journal, 17*(3), 177-190. https://doi.org/10.1080/19313152.2023.2178216

Sweller, J., van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–252. https://doi.org/10.1023/A:1022193728205

Wang, J.-F., Wang, T.-H., & Huang, C.-H. (2022). Investigating students' answering behaviors in a computer-based mathematics algebra test: A Cognitive-load perspective. *Behavioral Sciences, 12*(8), 1–14. https://doi.org/10.3390/bs12080293