Effects of Visual Representations and Associated Interactive Features on Student Performance on National Assessment of Educational Progress (NAEP) Pilot Science Scenario-Based Tasks

Richard P. Durán
University of California, Santa Barbara

Ting Zhang
American Institutes for Research

David Sañosa
University of California, Santa Barbara

Fran Stancavage
American Institutes for Research

March 2020
Commissioned by the NAEP Validity Studies (NVS) Panel

The NAEP Validity Studies (NVS) Panel was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

Panel Members:

Peter Behuniak
Criterion Consulting, LLC

Jack Buckley
American Institutes for Research

James R. Chromy
Research Triangle Institute (retired)

Phil Daro
Strategic Education Research Partnership (SERP) Institute

Richard P. Durán
University of California, Santa Barbara

David Grissmer
University of Virginia

Larry Hedges
Northwestern University

Gerunda Hughes
Howard University

Ina V.S. Mullis
Boston College

Scott Norton
Council of Chief State School Officers

James Pellegrino
University of Illinois at Chicago

Gary Phillips
American Institutes for Research

Lorrie Shepard
University of Colorado Boulder

David Thissen
University of North Carolina at Chapel Hill

Gerald Tindal
University of Oregon

Sheila Valencia
University of Washington

Denny Way
College Board

Project Director:

Frances B. Stancavage
American Institutes for Research

Project Officer:

Grady Wilburn
National Center for Education Statistics

For Information:

NAEP Validity Studies (NVS) Panel
American Institutes for Research
2800 Campus Drive, Suite 200
San Mateo, CA 94403
Email: fstancavage@air.org

EXECUTIVE SUMMARY

The National Assessment of Educational Progress's (NAEP's) transition to an entirely digitally based assessment (DBA) began in 2017. As part of this transition, new types of NAEP items have begun to be developed that leverage the DBA environment to measure a wider range of knowledge and skills. These new item types include the science scenario-based tasks (SBTs) that are the focus of this report.

The present project was one study in a program of research designed to gather evidence relevant to claims that students' performances on NAEP's DBA tasks are valid measures of constructs defined in the NAEP content frameworks. More specifically, our goal was to provide information that could help inform design considerations for construct-relevant use of visual and interactive features in future science items and tasks.

Like other NAEP DBA-enabled items, the science SBTs assess students through their interaction with multimedia tasks in which information is presented in two or more forms, such as on-screen text, audio narration, and visual representations. The latter include both static pictures (e.g., graph, charts, maps) and dynamic visual representations (e.g., animations, videos, interactive illustrations). Interactive features, which can be used by students to explore phenomena or enter responses, include typing text into designated text boxes, selecting radio - button options, manipulating slider bars, and selecting and dragging visual objects to different permissible portions of the screen.

The study involved a cognitive lab investigation of 31 eighth-grade students performing five science SBTs that had been used in the 2015 NAEP pilot assessment. Both the development of the cognitive lab protocols and the subsequent analyses of cognitive lab data were informed by principles for the use of multimedia in learning tasks developed by Mayer (2009, 2014) and graphical user interface (GUI) principles based on a synthesis of human-computer interface research (Watzman & Re, 2012). The main research question was:

> Which key visual and associated interactive features of NAEP science SBTs used in the 2015 pilot assessment might inhibit or enable the ability of students to accurately demonstrate their actual level of mastery of target knowledge and skills?

Our cognitive lab procedures yielded verbal and performance evidence that tested a series of conjectures about whether students would comprehend or, conversely, have issues with comprehending arrangements of visual and interactive features that were present in the five SBTs used in our study. Most of these conjectures originated from recommendations from our expert panelists and the study team; the remainder were identified in the process of reviewing our data.

Findings

Our analysis of students' performance, retrospective think-alouds, and video-recorded actions supported conclusions that students (a) generally, but not always, tended to comprehend the visual and interactive features used in the SBTs, and (b) had favorable views of the SBTs and found them engaging. For most of the students in our sample, however, the

SBTs required more time than was allotted in the 2015 pilot assessment. This finding is consistent with the pilot-assessment results.

We found a positive relationship between item performance and comprehension of SBT features; conversely, when students did have problems with particular features, we found a negative relationship. We also investigated and discovered potentially interesting relationships between student contextual variables and two measures of overall SBT comprehension as well as item performance.

There were mixed results concerning our a priori conjectures about particular features that would affect students' ability to navigate SBT scenes and address the solution of scored SBT items as intended. In cases where conjectures were not supported, it is possible that a contributing factor was the limited sample size used in our cognitive lab investigation (31 students in total, or about six per SBT). There were many instances in which only a few students had issues, and it was not possible to tell whether these issues were idiosyncratic or systematically associated with certain contextual characteristics.

Design Recommendations

Although our study focused on visual and interactive features in particular, these features are only one part of the multimedia context that must be considered when evaluating cognitive load and the demands placed on students. On the basis of our study, therefore, we offer a range of design recommendations for future development of science SBTs and similarly complex multimedia tasks.

All recommendations reflect the judgements of our expert panelists, who reviewed the study SBTs in light of multimedia learning principles. As explained in the text below, some, but not all, of the recommendations also were supported by the study findings. We have listed recommendations supported by clear evidence from our cognitive labs first, followed by those with null or mixed evidence.

Soft Feedback

Within a task, certain inputs from students can be designed to trigger "soft feedback," typically in the form of low-intensity visual signaling such as a flashing color shift, which signals to students that their input has been registered by the system. The use of soft feedback is consistent with Mayer's *signaling principle,* which states that key information should be highlighted.

We found that most of the interactive buttons in the study SBTs made good use of visual features, such as color changes, to signal different button states (i.e., enabled versus disabled, current tab versus hidden tab). There were other instances, however, in which the students might have benefited from soft feedback that had not been included in the design. For example, in one simulation with an interactive feature that allowed students to place barriers into specific locations, students had trouble determining when a barrier had been successfully manipulated into one of those locations. Soft feedback would have been helpful here.

We also noted a few instances in which soft feedback was distorted due to what we assumed to be unintended software "glitches."

We suggest reviewing a wider range of scenarios in which interactive features are used in order to determine which would benefit from adding soft feedback for student inputs. Soft feedback can be delivered visually but also in an audio format, such as momentary "clicks."

Scene Navigation

The SBTs used in our study were designed with a variety of means for students to navigate from scene to scene. These include features such as the *Next* button on the eNAEP toolbar at the top of the screen; the *Submit* button embedded within the task; and, in some cases, *Tab* buttons to move between associated scenes.

We observed a few instances in which students, given multiple options for navigation to the next scene, made errors in navigation or expressed uncertainty about the consequences of choosing a particular navigation option. We suggest standardizing the use of the navigation tools across scenes within a given SBT, and ideally across SBTs. This should include standardizing the relationship between the within-task navigation tools and the navigation tools built into the eNAEP toolbar. Furthermore, navigation tools should follow standard conventions of consumer-available digital platforms when possible. If nonstandard navigation tools are necessary or desirable for whatever reason, we suggest providing clear instructions (e.g., by visually signaling or highlighting the button to be used to navigate to the next scene). These suggestions are consistent with Mayer's *signaling principle*.

Screen Layout for Visual Features

Another tactic that can help reduce construct-irrelevant cognitive load is to employ a consistent and readily interpretable layout when positioning visual features within scenes (e.g., use the same screen position for the same or similar visual features that appear across multiple screens within an SBT).

Our experts found that the layouts of features within or across scenes in the study SBTs were generally consistent and sensible, but there were problems in some places. One example is the use of layouts that placed the text instructions for pressing a button on one side of the screen and the button itself on the other side. Scenes with this layout led some of our students to click on the text instructions themselves or to delay their progress to search for the relevant button. Mayer's *spatial contiguity principle,* which recommends that wording and icon labels be placed next to their graphical referents, applies in this case.

Task Instructions and Textual References to Other SBT Features

The wording of text features is critical for comprehension of SBTs, especially when such text refers to or explains other visual and interactive features. In their review, our experts identified several instances of confusing or overly complex wording; this is especially concerning when the text directly relates to how students are intended to interact with scored items because it raises the possibility of construct-irrelevant variance in these items.

Supporting the concerns expressed by our experts, our findings identified instances in which students were, in fact, confused by complex or unclear answer choices or text that was intended to be explanatory, leaving the students unsure of what action to take to respond correctly to the item. Review procedures for SBTs should include explicit consideration of all text features.

Insufficiently Specified Assessment Items

The pilot SBTs used in the study contained some scored items that our experts considered to be so open ended conceptually that it was doubtful that students would be able to understand the author's intent—and therefore respond appropriately—without further prompting. This conjecture was supported by our latency data analysis, in which we found that students spent the highest average time out of all scenes across SBTs (527 seconds) on one underspecified item.

This item required three extended-constructed responses, all of which we judged to be excessively open ended because of the wide range of plausible responses. In addition to spending an excessive amount of time on this item, students lost points for not meeting rubric requirements that were relatively narrow and specific given the open-ended nature of the question. We suggest prompting students further on open-ended items so that they better understand what is being asked of them. Another option is to adjust the rubrics to accommodate a wider range of responses, but this would not alleviate the excessive amount of time students spent on the underspecified items.

Data Representations

The clarity of data representations, including displays such as tables, graphs, and virtual instrument readings, is critical for understanding their content. This might be especially true in the case of dynamic representations, or representations that are not consistent with standard graphical conventions.

The SBTs in our study contained only a few instances in which students reported difficulty interpreting data representations. Outside of these cases, our students were able to comprehend the wide range of data representations used in the SBTs. Nonetheless, in developing future SBTs, we suggest minimizing construct-irrelevant difficulty by evaluating each data representation in light of the specific measurement objective(s) for which the data representation is being used. Also, if pretest data are collected, such as from cognitive labs, it would be desirable to probe students' perceptions of the data displays to confirm clarity.

Time

In our cognitive labs, few students completed the study SBTs within the time limits used in the 2015 science pilot assessment. Although we added 15 minutes to the time limit for each of the SBTs, some students still failed to complete their assigned SBTs within the allotted time limits. Our close-in analysis of the visual and interactive features of tasks suggested that the pace with which students progressed though SBTs was associated, not only with some of the visual and interactive features of the SBTs, but also with the difficultly of the scored items—in particular, items that required constructed responses. It is an open question as to

whether more careful design, both in terms of item wording and visual and interactive features associated with the items, could ameliorate the cognitive load and speed students' progress while still assessing the intended cognitive targets.

Put another way, the design challenge will be to develop SBTs that students can complete within a reasonable amount of time (e.g., 30 minutes) and that measure a range of cognitive targets—including targets that require the use of constructed-response items—while also incorporating engaging visual and interactive features and avoiding construct-irrelevant variance in item scores. Meeting this challenge will require further research of the type reported here.

Everyday Design Conventions

Task design should be informed by the design conventions that a student is likely to encounter in everyday life outside of the testing environment, given that these conventions play a role in shaping student expectations and, in turn, student performance. Research has shown that the expectations of an interface can structure user interactions, and that violating the conventions of digital interfaces can have short-term effects on task performance (Still & Dark, 2010). Although users can adapt to violations of these conventions (if they are consistent), the existence of this initial gap might have unintended effects in an assessment context, especially in the absence of an extended adaption or training period. It may not always be advantageous or beneficial to align with user conventions, but we suggest that task designers be parsimonious and purposeful in choosing to violate them. To fully leverage current conventions also would require keeping up to date with trends in the design of interfaces that students are likely to have experienced going into the assessment.

Avatars/Agents

A pedagogical agent was used in each of the study SBTs to give directions and provide context for the student, and all but one of the tasks used a visual representation (i.e., an avatar) for the agent. Our experts expressed concerns that the visual properties of some of the avatars (e.g., image size or extraneous animation) might unnecessarily increase cognitive load (Mayer & Moreno, 2003).

Despite the experts' concerns, we found that students generally comprehended the nature and intent of the avatar (e.g., they reported feeling connected to the investigation by the avatar), and they did not dwell on the avatar's appearance. There were instances in which students reported finding something odd about an avatar—such as the way the avatar was dressed or the avatar's apparent age—which may have been sufficiently distracting to impact performance. However, students did not indicate that they found any of the avatars to be distracting, and our study design did not allow us to isolate the effect of the avatars on performance.

Despite these null findings, we recommend that the design of avatars be informed by multimedia research. Several of Mayer's principles (e.g., *personalization, voice principle*) refer to properties of such agents and suggest that their depiction be personable and informal. Other principles suggest that the design of the avatars minimize the amount of extraneous information that the student needs to process. In addition, in order for item developers to

better understand and control the impact of different instantiations of avatars, we recommend that avatars be treated in a consistent manner across SBTs or that any variation be purposeful.

Extraneous Information

After reviewing the five SBTs targeted for inclusion in our study, our expert panelists flagged many instances where they felt irrelevant, or marginally relevant, information was intrusive and potentially distracting. These included instances of marginally relevant static images taking up a disproportionate amount of the visual field as well as instances of animation that served no clear purpose (e.g., an avatar animated to rock back and forth).

Even though our cognitive lab data did not yield evidence that supported the experts' concern, findings from other studies (e.g., Clark & Mayer, 2011; Harp & Mayer, 1998; Mayer, Heiser, & Lonn, 2001) suggest that visual and interactive features that introduce excessive irrelevant information—that is, information not critical to comprehending a scene or performing actions within it—should either be modified or eliminated. This recommendation follows Mayer's *coherence principle,* which states that unnecessary or extraneous information be excluded so as not to distract students from attending to more critical features.

Amount of Information

A related point, also aligned with the *coherence principle*, concerns the density of information within a scene, regardless of relevance to the task. Based on the suggestions of our experts, we targeted multiple information-rich scenes in the study SBTs for investigation. Although, overall, we found that students were successful in handling SBT scenes with large amounts of information and could focus attention on the critical features of the tasks, there was some evidence from two SBTs that some of the information-rich scenes caused comprehension issues. We recommend that task designers explicitly consider the amount of information presented on each screen in light of students' attentional resources and the potential for cognitive overload.

In Summary

The results of our study suggest that, with relatively few exceptions such as those noted above, students in our small sample generally understood the visual and interactive features of SBTs as intended, suggesting that NAEP science SBT development procedures are working well. That said, our study also suggests that better formulated principles, or guidelines, for visual and interactive features should be developed and that quality control focused on those principles should be integrated into the critical path. Such a strategy would likely have caught—earlier in the development process—features of the 2015 pilot science SBTs that we found to be problematic. The exact means for defining such principles remain to be developed, but we suggest that close-in investigations of students' interactions with these types of features could be a component, particularly if these investigations were informed by previous research on multimedia principles as applied to assessments.

As noted, our investigation was based on pilot versions of SBTs because NAEP was just beginning to explore the use of SBTs when our study began and pilot versions were the only ones available. Based on the results from the 2015 pilot, as well as interim feedback from our study, many of the SBTs were significantly revised, and many of the issues we identified were corrected or became irrelevant. These include issues with students failing to finish the SBTs in the allotted time. Consistent with standard NAEP practice, any SBTs that underwent significant revisions were repiloted before being used operationally.

CONTENTS

STUDY RATIONALE

Visual representations, such as charts, diagrams, tables, animations, and videos, are an important component of materials utilized in science instruction and are becoming pervasive in major large-scale science and mathematics assessments. These visual representations often are embedded in multimedia tasks (in which information is presented in two or more forms) and associated with interactive features that can be used by students to explore phenomena or enter responses. Developers of instructional materials and assessment items are utilizing interactive multimedia technologies in increasingly sophisticated ways.

As the National Assessment of Educational Progress (NAEP) is transitioning to digitally based assessments (DBAs), new types of NAEP items have begun to be developed that leverage the DBA environment to measure a wider range of knowledge and skills. An initial set of interactive computer tasks (ICTs) was fielded in 2009, and a new generation of interactive item types is being developed that includes the science scenario-based tasks (SBTs) that are the focus of this report.

Like other NAEP DBA-enabled items, the science SBTs assess students through their interaction with multimedia tasks. One general belief driving this assessment design is that tasks with vivid and interesting multimedia are more likely to engage students and thus facilitate their comprehension and performance. However, only a limited number of research studies in NAEP have been done to collect evidence in support of this belief, and there remain concerns about the extent to which the cognitive processing evoked by at least some of the visual and interactive features of tasks may be construct irrelevant rather than construct relevant.

In light of previous research findings on the impact of visual representations (as well as the impact of other multimedia features) in contexts other than NAEP (e.g., Mayer, 2014; Shah, Mayer, & Hegarty, 1999), this study aimed to examine—in an exploratory manner—the effects of various multimedia representations on students' performance on NAEP science SBTs. More specifically, our focus was on the extent to which students showed evidence of *comprehending* key visual and interactive features of tasks or, conversely, evidence of encountering *challenges* in comprehension. The former would presumably enhance performance, while the latter would inhibit performance.

Our goal was to generate information that could help inform design considerations for construct-relevant use of visual and interactive features in future science items and tasks. Achieving this goal required the development of an innovative cognitive lab methodology that allowed us to administer the SBTs, record and analyze students' performance on those tasks, and collect and analyze retrospective verbal accounts of how the students interpreted key visual and interactive features as they performed the tasks.

Although not a focus of the research, concern also was given—within the limitations of sample size—to possible interactions between key examinee demographic characteristics (gender, socioeconomic status, language-minority status) and item or task features, particularly interactions that might differentiate performance in ways that would not reflect the measurement of intended constructs.

THEORETICAL FRAMEWORK

Computerized multimedia SBTs of the sort examined in this study, by their very nature, employ a complex, interconnected sequence of scenes rich in visual and interactive features. This format allows students to comprehend a scientific investigation as taking place in a meaningful, real-world context. In addition, it expands the potential for measuring the target constructs because the ability to appropriately use certain visual and interactive features is an essential component of many of the target constructs undergoing assessment.

The format, however, also carries the risk of cognitive overload if students are required to process an amount of multimedia information that exceeds their available cognitive capacity—in particular, their attention span and working memory capacity (Baddeley, 1998; Chandler & Sweller, 1991; Mayer & Moreno, 2003; National Academies of Sciences, Engineering, and Medicine, 2018).[1] In order to provide appropriate responses to scored items embedded in the SBTs, students need to understand and process visual and interactive features in a designer-intended manner. Misinterpretation of visual and interactive features may not only distort students' comprehension of the intended content, but also may result in students expending their assessment time inefficiently as they struggle to make sense of representations and their interconnections.

The current work builds on past NAEP experience measuring science performance with hands-on tasks (HOTs) and, more recently, with the earlier generation of ICTs (Carr, 2012; National Center for Education Statistics, 2012), and it is grounded in learning sciences research on multimedia and human-computer interaction (Mayer, 2009, 2014) as well as research on evidence-centered design (Mislevy, 2008). Our approach can be described as an exploratory *response process* validity study (Ercikan & Pellegrino, 2017) focusing on students' comprehension of visual and interactive features in the manner intended by item developers.

Response process validity analyses of multimedia-based assessments is an emerging new field of research and still in its infancy (*op. cit.*). It is best thought of as an important subcomponent of the overall validity argument for an assessment—one that adds information on the extent to which a student's observable behavior and unobservable mental processes appear to align (or not align) with the intentions of the assessment developers. Evidence is based on analysis of data that capture external behaviors of students. These data may come from many sources, including process data that track students' use of time during an assessment, forced or free answer choices entered by the student, mouse or cursor movements, and recordings that track the moment-by-moment focus of the student's attention. Other sources of data on students' behaviors can derive from cognitive labs in which students provide verbal reports (think-alouds) about their actions and perceptions, either concurrently, as they work their way through an assessment, or retrospectively, following completion of the assessment. The validity analyses also may include examination of other variables that show an association with students' response process—variables such as students' item scores, demographic characteristics, and self-reports on topics including

[1] Unfortunately, current research does not support identification of a specific threshold for cognitive capacity, which would, in any case, differ by age, extent of subject area knowledge, and other test-taker characteristics. Rather, empirical investigations are necessary to develop guidelines for specific populations.

familiarity with multimedia formats, access to technology, and previous experience with the content and practice dimensions of specific assessment tasks.

STRUCTURE AND PURPOSE OF NAEP SCIENCE SCENARIO-BASED TASKS

Since the mid-2000s, NAEP has been experimenting with DBA-enabled tasks that target students' inquiry skills in science. An early example is the 2009 science ICTs that test students' problem solving in a computer-simulated environment. For instance, in the eighth-grade Bottling Honey ICT, which was designed to measure scientific inquiry and technological design, students were asked to conduct investigations on the effect of temperature on the flow of liquids. In a simulated laboratory setting, students conducted experiments to determine the optimum temperature for bottling honey in the least amount of time with the least expenditure of energy. Examples of the 2009 ICTs, as well as DBA-enabled tasks developed for the Technology and Engineering Literacy (TEL) assessment, can be found on the NAEP website (nces.ed.gov/nationsreportcard).

To understand the objectives of the current study, it is helpful to first review the structure and purpose of the NAEP science SBTs, which represent the next generation of DBA-enabled science tasks. Two-thousand-fifteen pilot versions of the new SBTs are the basis for the study reported here.

Structure. All science SBTs used in the 2015 NAEP pilot begin with an introductory scene that describes a simulated real-world scenario involving a complex scientific problem that is to be solved. A student examinee takes the role of a participant in the investigation, a role that requires the student to utilize both scientific inquiry practices and science content domain knowledge.

Following the introductory scene, the SBT proceeds to further elaborate on the problem to be investigated and to introduce steps, scientific tools, and inquiry procedures needed to solve the problem and its subproblems. Scored items are incorporated into many of the scenes and subscenes, sometimes including the introductory scene. Occasionally, a new scene is introduced that provides students with the correct answer to a previous item. This allows all students to progress to subsequent items with the relevant information required for solution. A final, concluding scene closes each SBT.

Measurement Targets. Each of the scored items embedded in the SBTs is intended to assess students' problem-solving proficiency along one of two dimensions: a science content dimension or an inquiry practice dimension. The measurement targets for both types of proficiency, as specified in the blueprint for the NAEP science assessment, are set for the assessment as a whole, which includes a range of item types—discrete items, items embedded in hybrid HOTs, and items embedded in SBTs.

Incorporation of Scientific Inquiry Practices. The NAEP science framework describes four scientific inquiry practices that are practical to measure in the NAEP science assessment:

1. Design or critique aspects of scientific investigations (e.g., involvement of control groups, adequacy of sample).

2. Conduct scientific investigations using appropriate tools and techniques (e.g., selecting an instrument that measures the desired quantity—length, volume, weight, time interval, temperature—with the appropriate level of precision).
3. Identify patterns in data and/or relate patterns in data to theoretical models.
4. Use empirical evidence to validate or criticize conclusions about explanations and predictions (e.g., check to see that the premises of the argument are explicit, notice when the conclusions do not follow logically from the evidence presented).

(National Assessment Governing Board, 2014, p. 69)

It is important to understand that these practices do not describe a strict ordering of the steps required to conduct an investigation; therefore, the practices are flexibly embedded in the presentation of a given SBT. For example, actions related to the practice of designing or critiquing aspects of a scientific investigation may be introduced *after* scenes involving identifying patterns in data, not just at the beginning of an SBT.

Platform. Like other NAEP DBA assessments, the science SBTs are administered using the eNAEP platform. During the assessment, the eNAEP platform displays a toolbar at the top of the screen, which is shown in Figure 1 as it was configured for the 2015 pilot SBTs.[2] Of particular interest to the current study are the navigation tools included in the toolbar—left- and right-facing arrows that allow navigation between SBT scenes. Within the text of the SBTs, these arrows were referenced, respectively, as the *Previous* button and the *Next* button (e.g., "Press the *Next* button to continue").

Figure 1. The eNAEP Toolbar, as Used in the 2015 Pilot Science SBTs

Task Length. The five SBTs examined in the study—which comprised all the eighth-grade SBTs administered in the 2015 pilot—were of two lengths: extended tasks that were designed to take up to 30 minutes to complete and short tasks that were intended to take up to 15 minutes. The number of scenes, subscenes, and scored items in a given SBT varied accordingly.

[2]Although not shown in Figure 1, the toolbar also displayed the scene ID for the scene currently being viewed.

DEVELOPMENT OF THE FRAMEWORK FOR DATA CODING AND ANALYSIS

The main research question posed by the study was:

> *Which key visual and associated interactive features of NAEP science SBTs used in the 2015 pilot assessment might inhibit or enable the ability of students to accurately demonstrate their actual level of mastery of target knowledge and skills?*

Two expert panel meetings were convened to advise the study; the panelists, who were experts in science instruction and learning sciences research on multimedia, guided the development of the cognitive interviews and the analyses used to address the main research question.

An initial expert panel meeting was convened to review the five eighth-grade science SBTs that were the focus of our study in light of the stated goals, content, and structure of the SBTs as well as the design specifications and evidence-centered design guidelines used in development. Because of the small number of pilot SBTs per grade level, fourth- and 12th-grade SBTs also were made available to the panel members to give them a broader sense of the variation across NAEP SBTs.

The three panelists were Richard Mayer (expertise: learning sciences, experimental design, and multimedia learning), Mary Hegarty (expertise: learning sciences and visual multimedia problem solving), and Danielle Harlow (expertise: science and engineering instruction, science standards, and teacher preparation). All three were faculty at the University of California, Santa Barbara (UCSB).

The panelists were asked to evaluate the visual and interactive features of each eighth-grade SBT (including associated language representations of task information), focusing on the comprehensibility of each scene from the perspective of a student asked to work on the task. More specifically, panelists were asked to consider the following question derived from graphical user interface (GUI) principles, principles that are based on a synthesis of human-computer interface research (Watzman & Re, 2012):

Overall, as a student proceeds through an SBT, given your expertise, do you sense that the student will be clear about each of the following?

- What to look at
- What to do
- When to do it
- Where to do it
- Why to do it
- How to do it

In addition to the GUI principles, the expert panel also was provided with a brief literature review of work by Richard Mayer and colleagues laying out a cognitive framework for multimedia learning and associated, research-based principles for the design of multimedia learning tasks (see Mayer, 2009, 2014). We suggested to the panel that these same multimedia principles were potentially relevant to the design of multimedia assessment tasks. More details on Mayer's multimedia principles and their implications for the NAEP SBTs can be found in Appendix A.

Panelists were asked to provide written summaries of their observations about aspects of the SBTs relevant to the goals of the study. In his summary, Richard Mayer noted that the visual and interactive SBT features that might affect performance could be viewed through the lens of principles derived from his long-term research program on multimedia learning (Mayer, 2009, 2014). More specifically, Mayer identified specific scenes and items from the SBTs that adhered to or violated the following principles for effective multimedia learning:

- **Coherence principle**: Delete extraneous material.
- **Spatial contiguity principle**: Place printed words next to the part of the graphic to which they refer.
- **Personalization principle**: Use conversational style rather than formal style.
- **Segmenting principle**: Break a complicated screen into smaller parts.
- **Signaling principle**: Highlight key information.
- **Modality principle**: When the material is complex, the presentation is fast paced, and the learners are familiar with the words used; present instructions in spoken form rather than printed form.
- **Voice principle**: Speak in a human voice rather than a machine voice.
- **Interactivity principle**: Do not provide unconstrained interactivity.

Following the meeting of the first expert panel, our research team created a Visual and Interactive Feature Data Matrix to associate features of the five target SBTs with panelists' observations. More specifically, we created a two-dimensional matrix in which the rows listed the key visual and interactive features that occurred in each scene of each SBT (e.g., radio buttons and corresponding text instructions in SBT 1, scene 10), and the columns represented each of the multimedia principles identified by Mayer (e.g., spatial contiguity principle), the GUI principles, and other categories of concern raised by panelists.

Based on the observations of the expert panelists, coupled with the insights of the research team, we then coded each occurrence of a visual or interactive feature, in a binary manner, as either potentially relevant or not relevant to use in investigating each of the principles identified in the columns. That is, if a feature was executed in a manner that aligned with (or conflicted with) a given principle we would code the feature in the Data Matrix as potentially relevant to investigating that principle through focused probing in the cognitive interview. In this way, the Data Matrix was used to guide the development of specific conjectures to be evaluated in the cognitive interviews.

Due to the size of the matrix, we do not present it in this report; however, we draw on the matrix in our discussion of the study's analyses.

DEVELOPMENT OF THE COGNITIVE LAB DESIGN AND PROTOCOLS

Contributions of the Expert Panel

The protocols for conducting the cognitive interviews were tailored to the specific SBTs and established with the advice given at a second expert panel meeting. The expert panelists included our three original consultants, Danielle Harlow, Mary Hegarty, and Richard Mayer, as well as two new experts, Nora Newcombe (Temple University/Spatial Intelligence and Learning Center; expertise: visual information processing and learning) and Tracy Noble (TERC; expertise: math and science assessment of diverse students and cognitive lab studies). The panelists were provided with a review of findings from the first panel meeting (and associated implications for study design), and then asked to discuss student sampling procedures and pilot protocols for the cognitive labs.

Following the meeting, the advice of the panelists was synthesized and incorporated into a set of draft protocols tailored to the specifics of each of the SBTs. The development process took account of the fact that several of the points raised by the panelists were suggestions and recommendations, rather than strong prescriptions. Within the constraints of Office of Management and Budget (OMB) preapproval sample size guidelines, we conducted pilot tryouts of these draft protocols before finalizing them for inclusion in the OMB submission that defined the actual study.

Key takeaways from the second expert panel meeting follow:

Sample. The expert panelists endorsed a sampling procedure that ensured demographic heterogeneity among participants; they also recommended sampling students who varied in terms of their exposure to educational technology and to instruction and activities related to science, technology, engineering, and mathematics (STEM).

Overall Cognitive Lab Design. The panelists agreed that an appropriate starting point could be a focus on the relationship between students' comprehension of SBTs and the extent to which the SBTs adhered to the Mayer multimedia learning principles. They also indicated that it was important to recognize that the study was exploratory in nature and that, although it could generate conjectures for further research, other research designs (e.g., experiments) would be necessary to test such conjectures scientifically.

The panelists recommended that, given the constraints of the study and its methods, attention should focus on the overall comprehensibility of tasks as experienced by students, especially as comprehension is related to the assessment targets of the scored items embedded in the SBTs. They further recommended that the study use data from both parts of the cognitive lab protocol: the student's self-administration of the SBT on a tablet computer, and the subsequent retrospective think-aloud and probing of the student's experience. Recordings of the self-administration would yield "observable" performance data based on the student's manipulation of inputs and generation of responses.

Panelists noted that consideration of cognitive-processing factors—such as cognitive load, language-processing demands, and perceptual and figural clarity—could inform the study's interpretation of students' think-aloud and performance data. In addition, panelists

suggested that preliminary conjectures about students' cognitive strategies could be derived from analyses of students' use of time and their response-choice and scene-navigation behaviors, coupled with their verbal reports.

The panelists also recommended that probes used during the retrospective think-aloud be selective and based on prioritization of issues. They suggested that it would not be a productive use of time to extensively probe students on less consequential violations of certain multimedia learning principles, such as whether students were bothered by the fact that text boxes obscured features of the background scene (unless information from the background scene was central to comprehending the text).

Think-Aloud and Prompt/Probe Procedures. Panelists considered a variety of options for think-aloud and prompt/probe questioning procedures. There was a consensus that, before students begin the retrospective think-aloud, they should be given an orientation that clarified what was expected of them. In particular, panelists suggested that students be asked to take on the task of assisting the assessment designers by helping the designers understand how students perceive the SBTs and their demands.

The panelists also suggested that the interview protocol include a mix of general prompts and more focused probes targeting specific SBT scene characteristics. In general, the questions suggested by panelists were focused on students' reactions to the user-comprehensibility issues and multimedia learning principles incorporated in the study design—articulated in a "kid-friendly" fashion. Furthermore, the panelists suggested asking students how the overall SBT presentation and the scored items could have been made clearer, and whether they had previous experience with the kinds of investigative situations represented by the SBTs.

eNAEP Tutorial. At the beginning of the testing session, students participating in NAEP are given a standardized eNAEP tutorial that introduces them to the full range of multimedia and interactive tools available in eNAEP (e.g., calculator tool, text or figure highlighting tool, note-taking tool), without consideration of whether these tools are activated in the specific subject-area assessment that the student will be taking. The expert panelists recommended that, in the interest of time, we administered a shortened tutorial that focused only on those tools applicable to the SBTs used in the study.

Measure of Contextual Information. We also discussed with the panelists a set of student contextual questions—to be administered at the end of the cognitive lab—that could inform the interpretation of the study results, if only in an exploratory manner. General areas of contextual information that panelists suggested collecting from students included availability and use of digital devices in the home, self-rated competency in employing technology and interest in acquiring scientific knowledge, science courses taken, participation in science clubs and informal science learning activities, and exposure to media, such as television programs, with a science content focus. However, panelists also recognized that only a limited number of questions would be feasible within the time constraint of a 2-hour cognitive lab and that priorities would have to be set.[3]

[3] In addition to contextual data collected from students, demographic data would be collected during recruitment.

Pilot Administration of the Cognitive Lab Protocols

Our study design required that each student complete one of the five eighth-grade science SBTs. As noted above, protocols were customized to fit the specifics of each SBT. Before finalizing our five protocols, we piloted them using six students residing in the Santa Barbara region. These included four regular eighth-grade students and two advanced ninth-grade students who were enrolled in an engineering academy at a local high school.

Our pilot data showed that, on average, our pilot subjects could complete no more than half of an SBT within the time limits used in the 2015 pilot and, correspondingly, built into the eNAEP software (15 or 30 minutes, depending on the SBT). This restriction resulted in our pilot subjects only reaching SBT scenes that covered about half of the conjectures that we intended to investigate. Given this experience, and with the assistance of the NAEP technology contractor, we altered the eNAEP software to add an additional 15 minutes to the time allocation for each SBT. With the extended administration time, we could substantially improve the completion rate for SBTs while still staying within the OMB-approved cognitive lab session limit of 120 minutes.

FINAL STUDY DESIGN

Cognitive Lab Protocol

Our general approach was to first have the student self-administer his or her assigned SBT. Following the self-administration, we played back a recording of the student's performance and asked the student to retrospectively comment on what he or she was thinking and doing during the self-administration with regard to the visual and interactive features of that SBT. This retrospective approach avoided the significant cognitive load that would have been imposed if we had asked the student to describe his or her thoughts and actions while working through the SBT for the first time. A three-ring notebook of screenshots from the SBT also was available for additional reference during the retrospective think-aloud and probes.

Interview Script. Based on input from our expert panelists, our team reviewed the SBTs in detail and identified 9 to 11 points within each individual SBT that involved visual and interactive features prioritized for investigation and could provide evidence of whether a student understood what he or she was seeing or being asked to do. During the think-aloud, when each of these points was reached, the administrator asked the prompt-and-probe questions that were built into the interview script and designed to focus the student on the feature of interest. For selected scenes, students also were presented with additional prompt-and-probe questions assessing students' understanding of key contextual information and interactive features that were central to interpreting the content of the scene and its connection to information presented elsewhere in the SBT.

Order of Activities. The cognitive lab began with an orientation in which the student first was given a brief introduction to the study. We then provided a think-aloud demonstration and practice session in which the administrator demonstrated the think-aloud procedure and the student practiced the technique.[4] Finally, the student viewed a shortened version of the eNAEP tutorial that, as noted earlier, introduced only those eNAEP tools that were active in the SBTs.

Following this orientation, the student worked through his or her assigned SBT using a Microsoft Surface Pro tablet computer (the standard device used in NAEP administrations during this time period). Computer software called Camtasia was used to record how the student responded to the task shown on the screen. After the self-administration was completed, and while the administrator played back the Camtasia recording of the student's performance, the student was (a) asked to describe what he or she was thinking while working through the task and (b) taken through the prompts and probes built into the interview script. The playback was paused as needed to give time for the student to finish talking about a feature before the feature disappeared from the screen.

[4] Given the lack of appropriate DBA items that could be used for practice, paper-and-pencil items were used in this exercise.

The cognitive lab ended with a brief set of contextual questions asking about the student's experiences and attitudes pertaining to technology and science education (see Appendix B). Each cognitive lab session was designed to take no more than 120 minutes.

Sample

Our study design called for the acquisition of usable data from at least 30 students—six for each SBT. In the summer of 2016, we assembled a convenience sample of 32 students who had just completed the eighth grade in the Santa Barbara and San Francisco Bay Area regions, and we successfully collected data from 31. (Data from the remaining participant were dropped due to a technical problem with the recording of the self-administration that impacted the retrospective think-aloud and probing.) Our sample size goal was thus attained, with an extra student run for one of the SBTs.

The study design also set a goal of recruiting and including students with diverse demographic characteristics, though population representativeness was not intended. The variables on which we sought participant diversity were gender, participation in the National School Lunch Program, and exposure to a non-English language at home. Table 1 shows the distribution of participants by these demographic characteristics for each SBT and overall.

Table 1. Distribution of Participant Demographics by Scenario-Based Task (SBT)

SBT ID	Number of Participants	Gender		Participation in National School Lunch Program		Language Spoken at Home Other Than English	
		Female	Male	Yes	No	Yes	No
SBT 1 (Short)	6	3	3	0	6	0	6
SBT 2 (Short)	6	3	3	2	4	1	5
SBT 3 (Short)	6	3	3	1	5	0	6
SBT 4 (Long)	6	3	3	1	5	2	4
SBT 5 (Long)	7	3	4	2	5	1	6
Total	**31**	**15**	**16**	**6**	**25**	**4**	**27**

Data Preparation and Coding

Data Preparation. The eNAEP software automatically collected data on students' response inputs and latencies in working through their assigned SBTs. In addition, the cognitive labs generated three types of raw data:

1. Camtasia video/audio recordings of students' performance during the self-administration of the SBTs,
2. Video camcorder recordings of the cognitive lab sessions, and
3. Digital audio recordings of the cognitive lab sessions.

Students' vocalizations as they recalled how they performed on the SBTs and in response to prompt-and-probe questions from the cognitive lab administrator were captured by both the video and audio recordings. This redundancy was intentional and done to help ensure that

equipment or software failure in one component would not necessarily compromise data for subsequent analysis.

The audio recordings were transcribed by a professional transcriber, and each transcript was annotated with the relevant scene IDs, time stamps, and prompt/probe question IDs.

The Camtasia and video camcorder recordings of each student were entered into a unified dual-video system (Dual Video Layout) that combined and synchronized the two types of video recordings and the audio track of our digital audio recorder. This allowed the coders to simultaneously display a screen that had the two types of video positioned side by side on a common time stamp grid with:

- The left-side video showing the Camtasia recording of specific SBT scenes that students saw in the playback as they went through the retrospective think-aloud, and
- The right-side video showing students retrospectively thinking aloud and gesturing spontaneously as well as responding to prompts and probes.

The purpose of this layout was to facilitate analysis of students' take-up of visual and interactive features from multiple perspectives—what students did when they first self-administered their assigned SBT without thinking aloud, and what students subsequently did and said when they viewed their original performance.

Estimation of Students' SBT Scores. Based on the Camtasia recordings, we scored students' responses to NAEP items embedded in the SBTs (using the most up-to-date scoring guides for the pilot SBTs) and calculated an SBT total score for each student. We subsequently used those scores to explore relationships between item difficulty (as estimated in the 2015 pilot assessment) and (a) the performance of students in the cognitive labs and (b) the occurrence of particular visual and interactive features of interest to our study.

Due to the limited sample size and sampling method (convenience sample) for our cognitive labs, it was not possible to estimate students' SBT total scores using the standard NAEP methodology, which entails item response theory (IRT) models and multiple imputations. Instead, students' SBT scores were estimated using *weighted percentage scores,* a procedure that allowed us to capture a student's overall task performance while considering the difficulty of individual items (as estimated from the 2015 pilot administration). The formula we used is presented as follows:

$$x_i = \frac{\sum_{j=1}^{k} u_{ij}\,(1 - P_j^+/100)}{\sum_{j=1}^{k} U_j\,(1 - P_j^+/100)}$$

Where x_i indicates the for individual student *i*, u_{ij} denotes the score that student *i* received on item *j*, and U_j represents the maximum possible score that a student can obtain from item *j*, with $j \in [1, \ldots, k]$ where *k* is the last item that a student reached with a response.

$P_j^+/100$ is the proportion of pilot assessment students who responded correctly on item *j,* and (1- $P_j^+/100$) is the proportion of pilot assessment students who responded incorrectly. In other words,

$$P_j^+ = \frac{number\ of\ correct\ responses}{number\ of\ correct\ responses + number\ of\ incorrect\ responses} * 100$$

For polytomous items, difficulty per point is 1 minus the average item score normalized to lie between 0 and 1.

Example of How to Calculate a Weighted Percentage Score

Suppose a test with four dichotomous items was administered to three students. The table below presents students' scores (1=correct; 0=incorrect), the maximum score for each item, and the item difficulty (i.e., $P_j^+/100$). In this simplified case, item difficulty is estimated from the three-student sample, and not from an external source.

Item #	Student1	Student2	Student3	Maximum Score	$P_j^+/100$
Item 1	1	1	0	1	.67
Item 2	1	1	1	1	1.00
Item 3	1	0	0	1	.33
Item 4	0	0	1	1	.33

P_1^+ /100 for item 1 is calculated as:

2/(2+1) =.67

The weighted percentage score for student 1 is calculated as:

(1*.67+1*1.00+1*.33+0*.33)/ (1*.67+1*1.00+1*.33+1*.33) = .8

In our study, we computed each student's raw and maximum-possible SBT scores as sums of the item scores for the items to which the student responded, weighted by difficulty per point. The ratio of those respective sums was the difficulty-weighted percentage score.[5] For SBTs for which students did not complete all items, the unreached items were excluded from both percentage calculations.

Claim Scoring. We developed procedures for coding and analyzing cognitive lab data based on input from the expert panel as compiled in the Visual and Interactive Feature Data Matrix mentioned earlier. We called these procedures *claim scoring*. The goal was for coders to render principled judgments regarding the degree of evidence that students understood and knew how to navigate specific visual and interactive features in scenes targeted for analysis. We also remained open to the possibility of examining data on visual and interactive features that extended beyond targeted scenes.

More specifically, each SBT was divided into the scenes (or sequences of related scenes) that we had targeted for investigation. Within each scene, we further noted each visual or interactive feature of interest and coded the evidence for a student's understanding of that feature in the form of a claim score.

[5] As noted, item difficulties were derived from the 2015 pilot administration.

Table 2 presents the four types of claim scores used to classify students' take-up of visual or interactive features.

Table 2. Claim Score Types

C	(C)omprehension	The student **navigated** visual and interactive features in a target scene (or sequence of related scenes) **with no evidence of confusion, inability** to focus on critical information, or **inability** to perform the actions required to proceed through a scene (or related scenes).
I	(I)ssues With Comprehension	The student showed evidence of encountering challenges in navigating visual and interactive features in a target scene (or sequence of related scenes) with **emergent or clear evidence of confusion, inability** to focus on critical information, or **inability** to perform the actions required to proceed through a scene (or related scenes).
M	(M)ixed	There was **contradictory evidence that the student had comprehended or had issues comprehending** visual and interactive features in a scene (or sequence of related scenes) that were required to respond to critical information-processing demands.
A	(A)mbiguous	There was **insufficient evidence that the student had comprehended or had issues comprehending** visual and interactive features in a scene (or sequence of related scenes) that were required to respond to critical information-processing demands.

The following data were used in coding the visual and interactive features claim scores for each SBT:

- Recording of the think-aloud interview as displayed in the dual-video layout
- Transcript of the think-aloud interview
- Participants' item responses and item scores

Two coders were involved in the claim score coding—the primary research assistant on the project (a graduate student at UCSB) and an American Institutes for Research (AIR) employee who was familiar with the SBTs through other assignments at AIR. The coding of each SBT started with the two coders calibrating their coding on the same cognitive lab session. To do this the coders went through the session one scene at a time, discussing each coding decision until they reached agreement before moving on to the next scene. After calibration, the rest of the cognitive lab sessions for the calibrated SBT were split between the coders, who worked independently. The team adjudicated the final coding; we did not implement a formal reliability study.

Analysis of Claim Score Data

Once the claim score coding was completed, the claim scores for all the students who took a given SBT were recorded and summarized in a Claim Score Distribution Table. Table 3 shows an example Claim Score Distribution Table for a representative SBT. There is one row in the table for each visual or interactive feature for which a claim score was coded,

organized according to the scenes (or series of related scenes) in which these features occurred. The scenes and features are identified in the first two columns of the table.

In the next set of columns, the claim scores of the students who took that SBT are recorded. There is one column per student, with individual students identified by unique identifier codes that take the form Psxx.

The final set of columns summarizes across students to show the total number of claim scores of each type that were "earned" by a given feature.

Data on students' performance on scored items are added in the final two rows of the table. The second-to-last row displays each student's total raw score (the sum of item scores); the last row displays each student's weighted percentage score.

Table 3. Example of a Claim Score Distribution Table for a Representative Scenario-Based Task (SBT)

SBT Scene	Visual and Interactive Features	Claim Scores of Students						Code Frequencies			
		Ps02	Ps08	Ps14	Ps15	Ps21	Ps28	C	I	M	A
Introduction(2), Introduction(3)	Graphic (including text labels) and text instructions	I	M	C	C	M	M	2	1	3	0
Introduction(4)	Measurement tool and corresponding text instructions	I	C	C	C	C	C	5	1	0	0
	Background image	A	A	I	A	A	M	0	1	1	4
	Data table and corresponding text instructions	M	C	C	C	C	C	5	0	1	0
Location(2)	Measurement tool, interactive features, and corresponding text instructions	C	C	M	C	C	I	4	1	1	0
	Background image	A	A	A	A	A	A	0	0	0	6
	Data table and corresponding text instructions	C	C	C	C	C	C	6	0	0	0
SampleData(2), SampleData(3)	Graph and corresponding text instructions	C	M	C	C	C	I	4	1	1	0
	Graphic (including text labels)	I	C	C	A	I	M	2	2	1	1
	Check boxes and text area	M	C	C	A	C	I	3	1	1	1
Simulation(3), Simulation(4)	Interactive features and corresponding text instructions	I	C	C	I	I	I	2	4	0	0
DataCompare(1)	Graphic (including text labels)	C	C	C	C	A	C	5	0	0	1
DataCompare(3), DataCompare(4)	Measurement tool and corresponding text instructions	I	I	C	C	C	I	3	3	0	0
	Data table	C	C	C	C	C	C	6	0	0	0
FinalItems(1)	Radio buttons (interaction) and corresponding text instructions	C	C	C	C	C	C	6	0	0	0
	Graphic (including text labels)	C	C	C	C	C	C	6	0	0	0
	Data table (including text labels)	M	C	C	C	C	I	4	1	1	0
General	Avatar	M	C	C	M	C	M	3	0	3	0
	eNAEP toolbar (e.g., includes *Next* and *Previous* buttons, time)	C	C	C	C	I	I	4	2	0	0
	Submit button	C	C	I	C	C	M	4	1	1	0
	Text area response and corresponding text instructions	C	C	I	C	C	I	4	2	0	0
Code Frequencies	C	11	17	17	16	15	8				
	I	5	1	3	1	3	8				
	M	4	2	1	1	1	5				
	A	2	2	1	4	3	1				
Sum of Raw Item Scores		25	26	27	28	30	23				
Weighted Percentage Scores		0.69	0.71	0.79	0.86	0.91	0.65				

Note: The scene names correspond to labels given in the task metadata. Claim scores are codes that classify the student's take-up of the visual or interactive features. C = (C)omprehension, I = (I)ssues with comprehension, M = (M)ixed, and A = (A)mbiguous. Ps<##> is the participant ID. Sum of raw item score = the sum of the students' scores on the items they completed. Weighted percentage score = the ratio of a student's total item score to the maximum possible score, weighted for item difficulty and based on the items the student completed.

Claim score data were analyzed in two ways. First, for each feature in the Claim Score Distribution Table for a given SBT, the conjectures that had led us to include the feature in the study were evaluated in light of the patterns of *C* and *I* claim scores earned by that feature. From this analysis, we drew conclusions as to whether or not each conjecture had been upheld.

For example, for one SBT, one row in the Claim Score Distribution Table referenced a series of laboratory-simulation scenes. The expert panel had raised concerns about the amount of irrelevant visual information present in the background of these scenes, particularly a table in the background for which the line defining the top of the table was visible behind two clear beakers shown in the foreground. The relative levels of liquid in each of these beakers was central to scene comprehension, and the expert panel hypothesized that the line of the tabletop, seen through the beakers, would interfere with the student's ability to read these levels.

The pattern of *C* and I scores, however, indicated that most students had no trouble comprehending and navigating the laboratory-simulation scenes despite the prominence of background images not central to SBT subtask demands; therefore, the conjecture was not upheld.

The second type of claim score analysis involved correlating the percentages of *C* and *I* scores with weighted percentage scores, both within and across SBTs. The purpose was to determine if there was evidence for a relationship between comprehension of visual and interactive features and performance.

Other Analyses

In addition to the analyses of claim scores, our study included several ancillary analyses. These included analyses of latency data, student contextual data, and responses to three general prompts/prompt sequences that covered students' overall comprehension of the SBTs, exposure to relevant science instruction, and evaluation of the SBT experience. These analyses are discussed in the Findings section.

FINDINGS

Because revised versions of the SBTs examined in this study will be used in future NAEP assessments, the content of the SBTs is confidential. Findings reported here are limited to those that can be described without exposing the content of the SBTs. This primarily affects the detail with which we can report the findings relating claim score data to conjectures.

Relating Claim Score Data to Conjectures

Our findings suggested that students understood how to respond to visual and interactive features of the pilot SBTs included in the study about two-thirds of the time, and that we could identify specific instances of visual and interactive features that created comprehension or response issues for multiple students in our sample.

There were mixed results, however, concerning our a priori conjectures about particular features that would affect students' ability to navigate SBT scenes and address the solution of scored SBT items as intended. In cases where conjectures were not supported, it is possible that a contributing factor was the limited sample size of our cognitive lab investigation (31 students, or about six per SBT). There were many instances in which only a few students had issues, and it was not possible to tell whether these issues were idiosyncratic or systematically associated with particular contextual characteristics.

That said, we summarize and discuss below the findings for several classes of visual or interactive features, as derived from our analysis relating claim scores to conjectures.

Soft Feedback. Within a task, certain inputs from students can be designed to trigger "soft feedback," typically in the form of low-intensity visual signaling, such as a flashing color shift, which signals to students that their input has been registered by the system. The use of soft feedback is consistent with Mayer's *signaling principle,* which states that key information should be highlighted.

We found that most of the interactive buttons in the study SBTs made good use of visual features, such as color changes, to signal different button states (i.e., enabled versus disabled, current tab versus hidden tab). There were other instances, however, in which the students might have benefited from soft feedback that had not been included in the design. For example, in one simulation with an interactive feature that allowed students to place barriers into specific locations, students had trouble determining when a barrier had been successfully manipulated into one of those locations. Soft feedback would have been helpful here.

Scene Navigation. The SBTs used in our study were designed with a variety of means for students to navigate from scene to scene. These include features such as the *Next* button on the eNAEP toolbar bar at the top of the screen; the *Submit* button embedded within the task; and, in some cases, *Tab* buttons to move between associated scenes.

We observed a few instances in which students, given multiple options for navigation to the next scene, made errors in navigation or expressed uncertainty about the consequences of choosing a particular navigation option. In keeping with Mayer's *signaling principle*, this

confusion could have been averted by clearer instructions, such as visually signaling or highlighting the button to be used to navigate to the next scene.

Screen Layout for Visual Features. Our experts found that the layouts of features within or across scenes in the study SBTs were generally consistent and sensible, but there were problems in some places. One example was the use of layouts that placed the text instructions for pressing a button on one side of the screen and the button itself on the other side. Scenes with this layout led some of our students to click on the text instructions themselves or to delay their progress by as much as 50 seconds to search for the relevant button. Mayer's *spatial contiguity principle,* which recommends that wording and icon labels be placed next to their graphical referents, applies in this case.

Task Instructions and Textual References to Other SBT Features. The wording of text features is critical for comprehension of SBTs, especially when such text refers to or explains other visual and interactive features. In their review, our experts identified several instances of confusing or overly complex wording; this is especially concerning when the text directly related to how students were intended to interact with scored items because it raises the possibility of construct-irrelevant variance in these items.

Supporting the concerns expressed by our experts, our findings identified instances in which students were, in fact, confused by complex or unclear answer choices or text that was intended to be explanatory, leaving the students unsure of what action to take to respond correctly to the item. For example, in one SBT, five out of seven students provided incorrect responses to an item that required interpretation of a data chart. From a follow-up probe, we collected evidence that the students understood the data chart; the problem arose due to a rhetorical mismatch between the wording of the item stem and the configuration of the chart.

Data Representations. The clarity of data representations, including displays such as tables, graphs, and virtual instrument readings, is critical for understanding their content. This might be especially true in the case of dynamic representations, or representations that are not consistent with standard graphical conventions.

For example, in one SBT, some students reported difficulty interpreting the graph that was associated with one of the scored items and, consequently, gave incorrect responses. There were features of the graph that may have contributed to the students' problems, including the faintness of the lines marking the graph intervals and the nonstandard labeling of the y-axis, which ascended in intervals of 4 kph rather than 5 kph.

Avatars/Agents. A pedagogical agent was used in each of the study SBTs to give directions and provide context for the student, and all but one of the tasks used a visual representation for the agent (i.e., an avatar). Our experts expressed concerns that the visual properties of some of the avatars (e.g., image size or extraneous animation) might unnecessarily increase cognitive load (Mayer & Moreno, 2003).

Despite the experts' concerns, we found that students generally comprehended the nature and intent of the avatar (e.g., reported feeling connected to the investigation by the avatar), and they did not dwell on the avatar's appearance. There were very few instances in which students reported finding something odd about an avatar—such as the way the avatar was

dressed or the avatar's apparent age—which may have been sufficiently distracting to impact performance. However, students did not indicate that they found avatars to be distracting, and our study design did not allow us to isolate the effect of the avatars on performance.

Extraneous Information. Our expert panelists flagged many instances across the five SBTs where they felt irrelevant, or marginally relevant, information was intrusive and potentially distracting, thus violating Mayer's *coherence principle*. These included instances of marginally relevant static images taking up a disproportionate amount of the visual field as well as instances of animation that served no clear purpose (e.g., an avatar animated to rock back and forth). However, our cognitive lab data did not yield evidence that supported the experts' concerns.

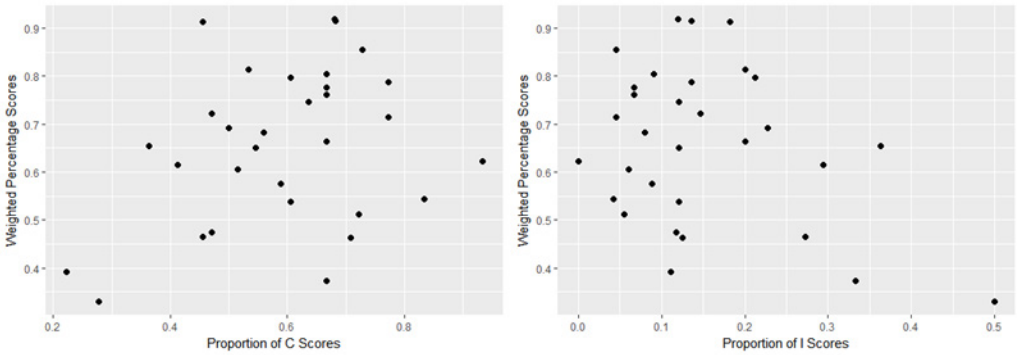
Amount of Information. A related point, also aligned with the *coherence principle*, concerned the density of information within a scene, regardless of relevance to the task. Based on the suggestions of our experts, we targeted multiple information-rich scenes in the study SBTs for investigation. Although, overall, we found that students were successful in handling SBT scenes with large amounts of information and could focus attention on the critical features of the tasks, there was some evidence that some of the information-rich scenes in two of the SBTs caused comprehension issues.

Relating Claim Scores to Students' Weighted Percentage Scores

To understand the overall relationship between students' weighted percentage scores and their proportions of *C* and *I* scores, Pearson correlations were computed for all 31 students across the five SBTs. Our correlation results indicated that there was a statistically significant, moderately positive correlation between the proportion of *C* scores and weighted percentage scores ($r = .40$, $p < .05$), and a statistically significant, moderately negative correlation between the proportion of *I* scores and weighted percentage scores ($r = -.45$, $p < .05$).

Figure 2 shows the scatter plots for these data.

Figure 2. Scatter Plots of Overall Correlation of Comprehension (C) and Issues With Comprehension (I) Claim Scores With Weighted Percentage Scores



As described in the Final Study Design section, the weighted percentage score was computed as the ratio of the sum of the student's weighted item scores to the sum of the weighted

maximum possible score, omitting unreached items. An item's weight was the inverse of the item's *p*+ values, as computed from the 2015 science pilot assessment. Weighted percentage scores are thus only a proxy for NAEP proficiency estimates (plausible values) based on IRT calibration and scaling. However, they are sufficiently meaningful to support the conclusion that, in our exploratory study, students' ability to navigate visual and interactive features of SBTs was significantly associated with students' science achievement, as assessed by the SBTs.

Relating Latency Data to Time to Complete SBTs

As has been alluded to earlier, the SBTs in our study took much longer to complete than had been anticipated prior to the 2015 pilot assessment. These NAEP results were replicated in our study, in which there was only one SBT where the majority of students (six out of seven) completed the SBT within the time limit used in the pilot assessment. For the remaining SBTs, no more than one student (and more often no student) completed the SBT within the pilot assessment time limit.

Except for one SBT, completion rates were much better with the extended time limits used in our study (30 minutes for short SBTs and 45 minutes for long SBTs). There were two SBTs that everyone completed within the allocated time, and two others that at least two-thirds of the students completed on time.

We used an analysis of latency data to shed light on how students who failed to finish their assigned SBT within the extended time limits used in the cognitive labs allocated their time across scenes and whether any of the visual and interactive features we investigated might be associated with their failure to complete the task.

For each scene in each SBT, based on the time stamps of the Camtasia recordings, we coded latency, which we define as the total amount of time a student spent traversing a scene before progressing to a scene not entered previously. Latency was measured in seconds, from the time that a new scene (a scene never previously viewed by the student) was first visible to the time when the student entered another new screen. Time spent backtracking to previously seen scenes was included in a scene's latency measure.

The concluding scene of each SBT was not coded for latency, as the cognitive lab protocol did not direct students to perform any action that could be used as a reliable indicator that the student had finished viewing the scene and completed the SBT.

Overall, we discovered that the scenes on which students spent relatively large amounts of time were characterized by the inclusion of constructed-response items. For example, in one SBT, students spent an average of 30 seconds on the introductory scene, a typical scene that does not contain a scored item. On the other hand, students spent an average of 527 seconds on a scene that contained a three-part constructed-response item.

This pattern was prevalent across all five SBTs. Camtasia recordings of students' behaviors confirmed that students spent most of their time working on scenes with constructed-response items.

On the other hand, we did not find statistically significant correlations between the total time spent and the proportion of *C* scores (r = -.27, p = .15) or *I* scores (r = .07, p = .72).

In summary, we did not find evidence that indicated that the amount of time that students spent on the SBTs was associated with any of the visual or interactive features that we investigated in our cognitive labs. On the other hand, results from our qualitative analyses suggested that the amount of time that students spent on a given SBT was related to the cognitive demand of the scored items. Scenes with the most demanding constructed-response items absorbed the most time.

Relating Students' Contextual Questionnaire Responses to Claim Scores and Performance

After completing the cognitive interviews, we administered a written questionnaire to students inquiring about the following areas:

- Experiences with science learning activities, including students' experiences related to science in general and specifically to the topics and practices explored in their assigned SBTs
- Students' science test-taking experiences
- Students' access to and familiarity with digital technology, such as computers, tablets, and smartphones
- Attitudes toward science, including students' competency beliefs in science, the value they place on science, their enjoyment of science, and their preference for technology-based tests

Items included in the contextual questionnaire were selected from the 2015 NAEP Science Student Questionnaire; 2015 NAEP Computer Access and Familiarity Study Survey[6] (Kitmitto, Bohrnstedt, Park, Bertling, & Almonte, 2018), Competency Beliefs in Science (Activation Lab, 2016), and Values in Science Surveys (Activation Lab, 2017).[7] The full text of the questionnaire items can be found in Appendix B.

We computed correlations between the students' responses to the contextual questionnaire items and their proportions of claim scores (*C* scores and *I* scores) for all 31 students. The nine questionnaire items for which the correlation with either *C* scores or *I* scores was significant at the 0.1 level are shown in Table 4, grouped by topic. In all cases, the item correlations with *C* scores and *I* scores were consistent in that, if one correlation was positive, the other was negative.

Most of the relationships were in the expected direction—that is, positive for *C* scores and negative for *I* scores. Two questionnaire items, however, showed a negative relationship with comprehension (negative for *C* scores and positive for *I* scores): frequency of using a

[6] The Computer Access and Familiarity Study Survey was administered as a special study in conjunction with the 2015 operational NAEP administration.

[7] The Competency Beliefs in Science and Values in Science Surveys were developed by the National Science Foundation-funded Activation Lab and have known psychometric characteristics.

smartphone for science learning in the past school year and degree of interest in a future job that involves using science.[8]

Table 4. Pearson Correlations Between Students' Responses to Contextual Questionnaire Items and Their Proportions of Claim Scores

Item	*C* Claim Score		*I* Claim Score	
	Correlation	*P-Value*	*Correlation*	*P-Value*
Experiences with science learning activities				
Having visited a museum, zoo, or aquarium to learn about science in the past school year	0.20	0.28	-0.41	0.02*
Computer access and familiarity				
Frequency of using a smartphone for science learning in the past school year	-0.24	0.20	0.43	0.02*
Owning a smartphone	0.43	0.02*	-0.12	0.53
Learning at school how to look up the meaning of a word using a computer	0.33	0.07*	-0.39	0.03*
Having a desktop computer at home	0.34	0.06*	-0.37	0.04*
Learning how to troubleshoot problems with a computer at school	0.32	0.08*	-0.26	0.17
Attitudes toward science				
Preference for taking tests on a computer rather than using paper and pencil	0.47	0.01*	-0.31	0.09*
Degree of interest in having a future job that involves using science	-0.41	0.02*	0.39	0.03*
Degree of "liking science"	0.33	0.07*	-0.32	0.08*

C = (C)omprehension; I = (I)ssues with comprehension

*$P \leq 0.1$

In a second analysis, we computed correlations between the students' responses to the contextual questionnaire items and their weighted percentage scores. The seven items for which the correlation with weighted percentage scores was significant at the 0.1 level are shown in Table 5. Once again, most relationships were in the expected direction—that is, items measuring experiences with science learning activities and access to and familiarity

[8] We also collected data on students' socioeconomic status—as measured by participation in the National School Lunch Program—and English language learner status, with the intention of examining how these two demographic variables related to students' uptake of the visual and interactive features of the SBTs in our study. However, the sample that was ultimately recruited included only six students receiving free or reduced-price lunch and only four who reported that English was not the primary language spoken at home. These small numbers precluded the possibility of generating reliable results, so, for this reason, analyses involving these two variables were not conducted.

with computers were positively related to performance, as measured by weighted percentage scores, and students' perception of the difficulty of the SBT "test" was negatively correlated. Less intuitive is the finding that students' experience of learning how to write a computer program or app at school was negatively related to performance. Some of these unexplained relationships with comprehension or performance might disappear with a larger, nationally representative sample.

Table 5. Pearson Correlations Between Students' Contextual Questionnaire Items and Their Weighted Percentage Scores

Item	Weighted Percentage Scores	
	Correlation	*P-Value*
Experiences with science learning activities		
Students' experience of participating in a science fair in the past school year	0.34	0.06*
Students' prior experience with scientific investigations similar to those presented in the scenario-based tasks (SBTs)	0.31	0.09*
Computer access and familiarity		
Own a smartphone	0.36	0.05*
Students' experience of learning at school how to install a new program or app	0.40	0.03*
Students' experience of learning at school how to search for information on the internet	0.35	0.06*
Students' experience of learning at school how to write a computer program or app	-0.32	0.08*
Perception of the test		
Students' perception of the difficulty of the test (i.e., the SBT) compared with most other tests that they took last year in school	-0.33	0.07*

*$P \leq 0.1$

Findings From General Prompt-and-Probe Questions Used Across SBTs

In addition to eliciting students' retrospective think-alouds and probing their understanding of specific visual and interactive features that were targeted for investigation in a given SBT, our cognitive lab protocols included three sets of prompt-and-probe questions that were administered to all students.

Overall Comprehension of SBT/Precision of Wording

At the beginning of the retrospective think-aloud, the cognitive lab administrator paused the playback of the student's performance at the end of the first scene and asked the student: "In your own words, please summarize what the task was about. What problem were the scientists trying to solve and how did they go about solving it?" There were two purposes for this initial prompt/probe cluster. One purpose was as a warm-up to facilitate a student's recall of the SBT.

The second was to gather verbal report data that could be used to estimate how well the student comprehended the basic elements of the investigation laid out in the SBT.

Specifically, we were interested in students' ability to articulate each of the three main structural components of an SBT investigation:

1. Statement of the problem**:** The SBT states the main problem in the first few scenes. The "problem" is the subject of the investigation within the SBT.
2. Measures taken to solve a series of subproblems that, taken together, lead to a solution for the main problem: The SBT contains a series of activities that incrementally target the main problem posed by the SBT.
3. Solution of the climax problem: At some point toward the end of the SBT, a solution to the main problem is reached, building from the actions taken to solve the subproblems.

Based on our consideration of the above, we coded students' responses to the introductory prompt/probe cluster on two dimensions: overall comprehension and precision of wording. Both were coded on a scale of 1–3.

The coding rubric for overall comprehension of the SBT was:

3 = Good task comprehension: The student's response indicated a detailed and accurate understanding of the main components of the task structure.

2 = Some task comprehension: The student's response indicated a less detailed or less accurate understanding of the main components of the task structure.

1 = Little or no task comprehension: The student's response indicated little or no understanding of the main components of the task structure.

The coding rubric for precision of wording was:

3 = Good precision of wording: The student used technically accurate terms, in a manner consistent with their use in the SBT. The student was highly articulate and communicated ideas with detail and precision.

2 = Some precision of wording: The student used some technically accurate terms but also may have used colloquial or less precise terminology in his or her responses. The student was somewhat articulate but may have described at least some of the elements of the SBT investigation in a less detailed and precise manner.

1 = Little or no precision of wording: The student used few or no technically accurate terms. The student was unable to describe the elements of the task structure with detail and precision.

Average overall comprehension scores ranged from 2.0 to 2.67 across the five SBTs, and average precision of wording scores ranged from 1.67 to 2.17. The correlation between overall SBT comprehension and weighted percentage scores, computed across all five SBTs, was moderate and statistically significant (r =.47, p < .05), while the correlation between precision of wording and total weighted percentage scores was low and statistically nonsignificant (r=.11). The latter indicates that the extent to which students clearly articulated their

understanding, using technically accurate words, was not associated with their SBT performance scores.

Prior Instruction

Another prompt/probe cluster for all students was administered at the end of the retrospective think-aloud. This cluster asked whether the student had previously received instruction on science relevant to his or her assigned SBT. We coded students' responses as either "yes" or "no." An independent group *t*-test was performed, with students' weighted percentage scores as the dependent variable and their prior instruction responses as the independent variable. Results showed that, calculated across all students, prior instruction was unrelated to the weighted percentage score (t (29) = -0.99, p = 0.33). This suggests that students' prior instruction—at least in the broad terms captured by the prompt/probe cluster—did not determine their performance on their assigned science SBTs.

Overall Evaluation of the SBTs

The final prompt/probe cluster administered to all students asked: "What did you think of the task as a whole?" The purpose was to elicit students' overall evaluation of their SBT experience. We coded responses on a five-level Likert scale as shown in Table 6.

Table 6. The Five-Level Likert Scale for Overall Evaluation of Scenario-Based Tasks (SBTs)

5	Very positive	Student shows high enthusiasm and approval.
4	Positive	Student shows some enthusiasm and approval.
3	Neutral	Student response is mixed or lacking in positive or negative appraisal.
2	Negative	Student shows little enthusiasm and some disapproval.
1	Very negative	Student shows no enthusiasm and high disapproval.

On average, students were positive about the SBTs (scores for the five SBTs ranged from 3.50 to 4.00). The correlation between students' evaluations of the SBT experience and their weighted percentage scores, computed across all SBTs, was not statistically significant (r = 0.27, p = 0.15), indicating no relationship between students' evaluations of their SBT experience and their task performance.

In Summary

Finally, we wanted to remind readers that our investigation was based on pilot versions of SBTs because NAEP was just beginning to explore the use of SBTs and these were the only versions available when our study began. Based on the results from the 2015 pilot assessment, as well as interim feedback from our study, many of the SBTs were significantly revised, and many of the issues we identified were corrected or became irrelevant. These include issues with students failing to finish the SBTs in the allotted time. Consistent with standard NAEP practice, any SBTs that underwent significant revisions were repiloted before being used operationally.

DESIGN RECOMMENDATIONS FOR SBT DEVELOPMENT

Although our study focused on visual and interactive features in particular, these features are only one part of the multimedia context that must be considered when evaluating cognitive load and the demands placed on students. On the basis of our study, therefore, we offer a range of design recommendations for future development of science SBTs and similarly complex multimedia tasks.

Soft Feedback

Within a task, certain inputs from students can be designed to trigger "soft feedback," which signals to students that their input has been registered by the system. The use of soft feedback is consistent with Mayer's *signaling principle,* which states that key information should be highlighted.

We suggest reviewing a wider range of scenarios in which interactive features are used in order to determine which would benefit from adding soft feedback for student inputs. Soft feedback can be delivered visually but also in an audio format, such as momentary "clicks."

Scene Navigation

The SBTs used in our study were designed with a variety of means for students to navigate from scene to scene. We suggest standardizing the use of the navigation tools across scenes within a given SBT, and ideally across SBTs. This should include standardizing the relationship between the within-task navigation tools and the navigation tools built into the eNAEP toolbar. Furthermore, navigation tools should follow standard conventions of consumer-available digital platforms when possible. If nonstandard navigation tools are necessary or desirable for whatever reason, we suggest providing clear instructions (e.g., by visually signaling or highlighting the button to be used to navigate to the next scene). These suggestions are consistent with Mayer's *signaling principle*.

Screen Layout for Visual Features

Another tactic that can help reduce construct-irrelevant cognitive load is to employ a consistent and readily interpretable layout when positioning visual features within scenes (e.g., use the same screen position for the same or similar visual features that appear across multiple screens within an SBT). Layouts that, for example, place the text instructions for pressing a button on one side of the screen and the button itself on the other side should be avoided. Mayer's *spatial contiguity principle,* which recommends that wording and icon labels be placed next to their graphical referents, applies in this case.

Task Instructions and Textual References to Other SBT Features

The wording of text features is critical for comprehension of SBTs, especially when such text refers to or explains other visual and interactive features. Review procedures for SBTs should include explicit consideration of all text features.

Insufficiently Specified Assessment Items

The pilot SBTs used in the study contained some scored items that our experts considered to be so open ended conceptually that it was doubtful that students would be able to understand the author's intent—and therefore respond appropriately—without further prompting. This conjecture was supported by our latency data analysis, in which we found that students spent the highest average time out of all scenes across SBTs (527 seconds) on one underspecified item that required three extended-constructed responses, all of which we judged to be excessively open ended because of the wide range of plausible responses.

We suggest prompting students further on open-ended items so that they better understand what is being asked of them. Another option is to adjust the rubrics to accommodate a wider range of responses, but this would not impact the amount of time students spent on the underspecified items.

Data Representations

The clarity of data representations, including displays such as tables, graphs, and virtual instrument readings, is critical for understanding their content. This might be especially true in the case of dynamic representations or representations that are not consistent with standard graphical conventions.

In developing future SBTs, we suggest minimizing construct-irrelevant difficulty by evaluating each data representation in light of the specific measurement objective(s) it is intended to assess. Also, if pretest data are collected, such as from cognitive labs, it would be desirable to probe students' presentations of the data displays to confirm clarity.

Time

Our close-in analysis of the visual and interactive features of tasks suggested that the pace with which students progressed though the study SBTs was associated not only with some of the visual and interactional features of the SBTs, but also with the difficultly of the scored items—in particular, items that required constructed responses. It is an open question as to whether more careful design, both in terms of item wording and visual and interactive features associated with the items, could ameliorate the cognitive load and speed students' progress while still assessing the cognitive targets.

Everyday Design Conventions

Task design should be informed by the design conventions that a student is likely to encounter in everyday life outside of the testing environment, given that these conventions play a role in shaping student expectations and, in turn, student performance. Research has shown that the expectations of an interface can structure user interactions, and that violating the conventions of digital interfaces can have short-term effects on task performance (Still & Dark, 2010). It may not always be advantageous or beneficial to align design with user conventions, but we suggest that task designers be parsimonious and purposeful in choosing to violate them. To fully leverage current conventions also would require keeping up to date

with trends in the design of interfaces that students are likely to have experienced going into the assessment.

Avatars/Agents

A pedagogical agent was used in each of the study SBTs to give directions and provide context for the student, and all but one used a visual representation for the agent (i.e., an avatar). We recommend that the design of the avatars be informed by multimedia research. Several of Mayer's principles (e.g., *personalization, voice principle*) refer to properties of such agents and suggest that their depiction be personable and informal. Other principles suggest that the design of the avatars minimize the amount of extraneous information that the student needs to process. In addition, in order for item developers to better understand and control the impact of different instantiations of avatars, we recommend that avatars be treated in a consistent manner across SBTs or that any variation be purposeful.

Extraneous Information

Even though our cognitive lab data did not yield evidence that information identified as extraneous by our experts caused problems for students, we suggest, based on findings from other studies (e.g., Clark & Mayer, 2011; Harp & Mayer, 1998; Mayer, Heiser, & Lonn, 2001), that visual and interactive features that introduce excessive irrelevant information—that is, information not critical to comprehending a scene or performing actions within it—should either be modified or eliminated. Examples include marginally relevant images that take up a disproportionate amount of the visual field as well as animation that serves no clear purpose.

This recommendation follows Mayer's *coherence principle,* which states that unnecessary or extraneous information be excluded so as not to distract students from attending to more critical features.

Amount of Information

A related point, also aligned with the *coherence principle*, concerns the density of information within a scene, regardless of relevance to the task. We recommend that task designers explicitly consider the amount of information presented on each screen in light of students' attentional resources and the potential for cognitive overload.

STUDY LIMITATIONS AND IMPLICATIONS FOR FUTURE RESEARCH

Response process validity studies involving computer delivered SBTs in science and other content areas are in their infancy and can take many forms. Below, we suggest some directions for new studies, given the strengths and limitations of our study and its findings. We begin with mentioning some of the key limitations of our study.

Study Limitations

Foremost among these limitations are the small sample of SBTs (n=5) and the small number of students who were administered each task (six or seven per SBT, for a total of n=31). We included all the Grade 8 science pilot SBTs that were available as of the date of the study, but this small number of SBTs did not create a robust sampling of the range of possible SBTs and the variety of visual and interactive features that can arise within them. Although the universe of possible SBTs is immense and open ended, it would be valuable to develop a methodology for systematically cataloging their visual and interactive features. This would facilitate a more comprehensive investigation of the effects of such features on student response processes, with implications for the validity of SBTs as a whole as well as the validity of the subcomponents of these tasks.

With regard to sampling of students, the sample sizes we used are appropriate for cognitive lab studies, but do not produce findings that are reliably representative of the performance of the possible universe of students, either for the SBTs used in our study or, for that matter, across the universe of possible SBTs. Although, theoretically, increasing the number of students in a study with a design similar to the one we used could yield more representative results, it needs to be recognized that there are practical limits to such a strategy because of the intensive case study costs entailed. Realistically, a different type of study design, with different investigative goals, would be required.

Another lesser, but still notable, limitation of our study was the lack of a method to establish the reliability of the scores we assigned to students' actions and verbal responses—scores that were used as the warrant for assignment of claim scores. Offsetting this limitation was the fact that the claim scores were never used in isolation or in a purely quantitative manner. All conclusions about the conjectures that were the main driving focus of our study also were informed by a close-in review of the details of students' responses. This approach was made possible by the intensive cognitive lab methodology and shows the counterbalancing benefits of working with small sample sizes.

That said, reliability of scores is still an important issue. Our study procedures did involve a calibration procedure in which both of our coders scored data for the first student to be processed for each SBT, and then met to discuss and resolve differences. Due to the limitations of study resources, however, all the remaining cases for a given SBT were only scored by one of our coders. Any future implementation of a study such as ours would be well served by using multiple coders for at least a percentage of cases, so as to be able to compute reliability.

Future Studies

I. Studies With Larger and More Representative Samples

The findings of this study, although preliminary, suggested that students' attitudes toward science, experience with science learning activities, and computer access and familiarity all were associated, in the expected direction, with their comprehension of SBT visual and interactive features and/or performance on SBT items. That is, more favorable attitudes, greater experience, and greater computer access and familiarity were associated with better understanding of, and higher performance on, the SBTs. These associations could be further explored in future studies with larger samples of students who are known to vary with regard to these characteristics.

Another area that deserves further inquiry is the relationship between students' socioeconomic status (SES) and English language learner (ELL) status, and their comprehension of visual and interactive features. Future research should investigate whether students from different key demographic groups show equal facility in comprehending the visual and interactive features of SBTs, controlling for students' previous science exposure and familiarity with technology.

II. Finer Grain Investigations Taking Account of the Science Content and Cognitive Processing

A second line of promising research would be to parse, for each of the successive scenes in an SBT, how the scenes are associated with specific components of an investigation (e.g., identify patterns in data), what explicit problem and processing demands are present, and what visual and interactive features are used.[9] It would then be possible to conceptually model or hypothesize how the visual and interactive features of the scene interact with the information demands of that step in an investigation.

One way to do this would be to produce (a) a detailed process model (Kane & Mislevy, 2017) for each scene of an SBT, based on the investigative component with which the scene is associated, and (b) a conceptual model of the science content domain expertise required by the scene, based on the NAEP science assessment specifications. Systematic attention could then be given to hypothetically modeling ways that—at a given step in an investigation—particular visual and interactive features varied in their importance for students' subsequent performance.

This is not unlike the approach taken in the present study. The difference is that the approach would involve modeling the performance requirements for each SBT scene, categorized by the steps in the NAEP model for conducting scientific investigations. The cognitive lab procedure used in the present study could be modified for use in such a study,

[9] This type of analysis was not possible in the present study because the only scenes for which we could definitively identify the investigative component intended by the test developer were the scenes that contained scored assessment items; in those instances, the item metadata identified the target skill, which, in turn, could be associated with a specific component of an investigation.

although the duration of the retrospective think-aloud, with associated prompts and probing, would need to be extended considerably.

III. Experimental Studies With More Causal Designs and the Controlled Occurrence of Visual and Interactive Features

We suggest that experimental methods be used in future studies to isolate potential causal relationships between specific visual and interactive features and students' comprehension of and performance on SBTs. This would require the creation of alternate versions of tasks that systematically vary a small number of critical features.

Our study was limited by the fact that we had to rely on naturally occurring, rather than systematically engineered, variations in visual and interactive features. For example, when probed about their reaction to the presence of the avatars, none of our students' responses indicated that their comprehension or performance was hindered by the avatars, regardless of the avatars' static or animated form, screen size, or where they appeared on a screen. Some students expressed no interest in the avatar that appeared in their assigned SBT, while others used words such as "creative" or "helpful," indicating their approval of the visual presence of the avatar. One student reported that it felt like it was the avatar that was "talking" to him or her instead of "just the computer," suggesting that the avatar enhanced engagement. This finding is encouraging and consistent with the hypothesis that multimedia presentation of tasks on a computer is engaging, but only by replicating it in a true experiment series could we develop reliable, replicable evidence to inform design principles for avatars in SBTs.

IV. Exploration of Other Student Performance Data

As part of an enhanced program of SBT response process validity studies, we also suggest examining process data to develop a more fine-grained understanding of students' actions. Such process data might include, for example:

- Records of students' cursor movements into and across visual regions of a scene and associated instances of input (e.g., button presses) on objects shown in these regions, regardless of whether the scene includes prompts or probes calling for such input.
- Data from other instrumentation that can track where attention is focused.

The goal of tracking observable process data would be to develop and analyze evidence of how students deploy their attention selectively to visual and interactive features of tasks, and then to study whether this deployment (a) helps explicate students' problem-solving strategies, and (b) associates positively or negatively with students' problem-solving accuracy and time efficiency.

Enhancing the Development and Piloting Process of SBT-Like Tasks

The development and implementation of NAEP involves the coordination of many complex and simultaneous activities. In this context, the creation of the NAEP science SBTs has been a grand learning experience in operationalizing innovative computerized assessment that is

better aligned with student learning in a world increasingly dependent on electronic technologies.

NAEP item development procedures include a series of sequential steps, such as content, universal design, and bias reviews, play testing, usability cognitive labs, pilot testing, and field testing, all of which are designed to support the validity of operational NAEP assessments However, these procedures must be fitted into a "critical path" with stringent timeline requirements. Our study is an example of work that can be conducted outside this critical path to further inform development, particularly development of new or innovative item types.

The results of our study suggest that, with relatively few exceptions, NAEP science SBT development procedures are working to produce well-crafted scenarios, and students in our small sample generally understood the visual and interactive features of SBTs as intended. That said, our study also suggests that better formulated principles, or guidelines, for visual and interactive features should be developed through research of the types suggested above, and that quality control focused on those principles be integrated into the critical path. Such a strategy would likely have caught—earlier in the development process—features that we found to be problematic. The exact means for defining or implementing such principles and quality control strategies remain to be developed, but we suggest that close-in investigations of students' interactions with these types of features could be a component, particularly if these investigations were informed by previous research on multimedia principles, as applied to assessments.

REFERENCES

Activation Lab. (2016, March). *Competency beliefs in science* (version 3.2). Retrieved from http://activationlab.org/tools/

Activation Lab. (2017, March). *Valuing science* (version 3.3). Retrieved from http://activationlab.org/tools/

Baddeley, A. (1998). *Human memo*ry. Boston, MA: Allyn & Bacon.

Carr, P. (2012, June 26–29). *Moving to the next generation of science assessments: Lessons learned from the National Assessment of Educational Progress*. Presentation at the 2012 National Conferences on Student Assessment, Minneapolis, MN.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. C*ognition and Instruction, 8*(4), 293–332.

Clark, R. C., & Mayer, R. E. (2011). *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning* (3rd ed.) New York, NY: John Wiley & Sons.

Ercikan, K., & Pellegrino, J. W. (Eds.). (2017). *Validation of score meaning for the next generation of assessments: The use of response processes.* New York, NY: Routledge.

Harp, S. F., & Mayer, R. E. (1998). How seductive details do their damage: A theory of cognitive interest in science learning. *Journal of Educational Psychology, 90*(3), 414–434.

Kane, M., & Mislevy, R. (2017). Validating score interpretation based in response processes. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes.* New York, NY: Routledge.

Kitmitto, S., Bohrnstedt, G. W., Park, B. J., Bertling, J., & Almonte, D. (2018). *Developing new indices to measure digital technology access and familiarity.* San Mateo, CA: American Institutes for Research.

Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31–48). New York, NY: Cambridge University Press.

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). New York, NY: Cambridge University Press.

Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 43–71). New York, NY: Cambridge University Press.

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology, 93*(1), 187–198.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 43–52.

Mislevy, R. (2008, September 8). *Some implications of expertise research for educational assessment*. Keynote address at the 34th International Association for Educational Assessment (IAEA) Conference, University of Cambridge, Cambridge, England.

National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24783

National Assessment Governing Board (NAGB). (2014). *Science framework for the 2015 National Assessment of Educational Progress*. Washington, DC: U.S. Government Printing Office.

National Center for Education Statistics. (2012). *The Nation's Report Card: Science in action: Hands-on and interactive computer tasks from the 2009 science assessment* (NCES 2012–468). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Shah, P., Mayer, R. E., & Hegarty, M. (1999). Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph construction. *Journal of Educational Psychology, 91*(4), 690–702.

Still, J. D., & Dark, V. J. (2010). Examining working memory load and congruency effects on affordances and conventions. *International Journal of Human-Computer Studies*, *68*(9), 561–571.

Watzman, S., & Re, M. (2012). Visual design principles for usable interfaces: Everything is designed: Why we should think before doing. In J. A. Jacko (Ed.), *The human–computer interaction handbook—Fundamentals, evolving technologies and emerging applications (*3rd ed., pp 317–339). New York, NY: CRC press.

APPENDIX A. SELECTED MAYER'S MULTIMEDIA PRINCIPLES AND IMPLICATIONS FOR NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS SCENARIO-BASED TASKS

Multimedia Design Principle	Definition and Main Point	Implications for the National Assessment of Educational Progress (NAEP) Scenario-Based Tasks (SBTs)
Coherence principle	The coherence principle suggests that students' learning would be hurt when interesting but irrelevant multimedia features, such as words, pictures, sounds, or music, are added to an instructional explanation. Meanwhile, students would learn better from a lesson containing fewer materials that get to the point than from a lengthy and elaborated multimedia lesson (Mayer, 2005).	In the context of science SBTs, Mayer indicates that coherence means deleting extraneous material. The graphics should present the key elements that are central to the scenario and minimize nonessential material that is not relevant to the scenario.
Spatial contiguity principle	When multiple interdependent sources of information are physically or temporally separate from each other, making it impossible for a reader to attend to both simultaneously, trying to mentally integrate separate sources of information increases working memory load and may interfere with comprehension and learning. The spatial contiguity principle suggests that to minimize the cognitive load, multiple sources of information should be integrated into an optimal format. Students learn better when the corresponding words and pictures are presented near rather than far from each other on the same screen or page (Mayer, 2005).	In the context of science SBTs, Mayer suggested placing printed words next to the part of the graphic they refer to. For example, some of the SBTs may contain large text blocks that are separated from the graphics. Key words in the text blocks should be connected to the graphics, either through pointer lines or placement of the text next to the corresponding part of the graphic.
Personalization principle	In multimedia game design, Mayer (2014) suggests that people learn better in multimedia mode when words are in conversational style rather than formal style.	The same principle can apply to the design of SBTs in the assessment context. For instance, the words used in the scenario should directly address the student as "you" and should refer to the narrator as "I" rather than solely using third person. Students should be invited to join the narrator in solving an interesting scientific problem. The wording should be polite, especially when giving feedback and directions.
Segmenting principle	Break a complicated screen into smaller parts (Mayer, 2014).	In the context of SBTs, some of the scenarios fill the screen with too much information at one time. This information should be broken into meaningful parts that can be presented in succession.
Signaling principle	Highlight key information.	Key words in the text can be highlighted using devices such as bold font. Key aspects of the graphics can be highlighted with arrows, motion, framing, or coloring.

Multimedia Design Principle	Definition and Main Point	Implications for the National Assessment of Educational Progress (NAEP) Scenario-Based Tasks (SBTs)
Interactivity principle	Do not provide unconstrained interactivity.	When a scenario requires that students drag and drop an object, provide guidance and cues so it is obvious to the student what action to take. When a scenario has a slider bar, make its usage obvious or, better yet, convert it to a set of discrete values to click. When a scenario has tabs that can be consulted as needed, make them salient and highlight when they have been used or not used.
Modality principle	Present instructional words in spoken form rather than printed form with the boundary conditions in which the material is complex, the presentation is fast paced, and the learners are familiar with the words (Mayer, 2014).	The same principle can apply to the SBTs. However, in some cases, print may be more efficient and may be required for technical words or when it is important for the material to be available on the screen. Printed text may be preferred for students who are not native speakers of English.
Voice principle	Speak in a human voice rather than a machine voice (Mayer, 2014).	Research has shown that students try harder when the voice is a friendly, likable human voice rather than a machine voice. The print-to-speech tool may have to use a machine voice for cost reasons, but any in-scenario voices should be human.

APPENDIX B. STUDENT CONTEXTUAL QUESTIONNAIRE

Section 1.

1. Do you have science class this year/this semester?

   Ⓐ Yes

   Ⓑ No

2. Have you done any scientific investigation in any of your science classes like the task that you were presented here?

   Ⓐ Yes

   Ⓑ No

3. If your answer is Yes, please describe the scientific investigation(s) in a few words.

4. In this school year, have you participated in any of the following activities? Fill in one oval on each line.

	Yes	No
a. Science fair	Ⓐ	Ⓑ
b. Science club	Ⓐ	Ⓑ
c. Science competition	Ⓐ	Ⓑ

5. In this school year, have you visited a museum, zoo, or aquarium to learn about science?

   Ⓐ Yes

   Ⓑ No

6. How hard was this test compared with most other tests you have taken this year in school?

   Ⓐ Easier than other tests

   Ⓑ About as hard as other tests

   Ⓒ Harder than other tests

   Ⓓ Much harder than other tests

Section 2.

In this section, you will answer questions about different kinds of computers, tablets, and smartphones.

This is an example of a desktop computer.

7. Do you have a desktop computer at home?

   Ⓐ Yes

   Ⓑ No

8. In this school year, how often have you used a desktop computer **for science learning**? Count all the times you did this at home, at school, in an afterschool program, or anywhere else.

   Ⓐ Never or hardly ever

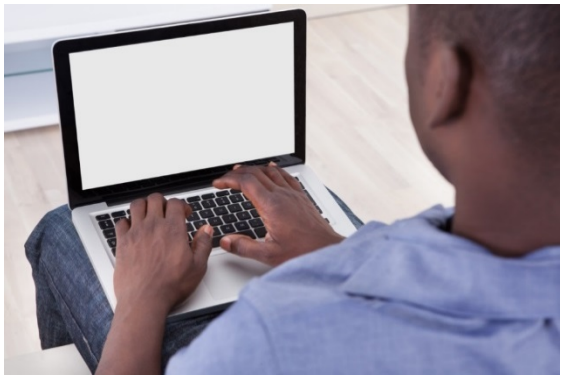
   Ⓑ Once every few weeks

   Ⓒ About once a week

   Ⓓ Two or three times a week

   Ⓔ Every day or almost every day

This is an example of a laptop computer.



9. Do you have a laptop computer at home?

   Ⓐ Yes

   Ⓑ No

10. In this school year, how often have you used a laptop computer **for science learning**? Count all the times you did this at home, at school, in an afterschool program, or anywhere else.

    Ⓐ Never or hardly ever

    Ⓑ Once every few weeks

    Ⓒ About once a week

    Ⓓ Two or three times a week

    Ⓔ Every day or almost every day

This is an example of a smartphone. A smartphone is any phone that is able to connect to the internet. Besides making phone calls and taking pictures, smartphones allow you to do many of the same things as a desktop or laptop computer.

11. Do you own a smartphone?

    Ⓐ Yes

    Ⓑ No

12. In this school year, how often have you used a smartphone **for science learning**? Count all the times you did this at home, at school, in an afterschool program, or anywhere else.

    Ⓐ Never or hardly ever

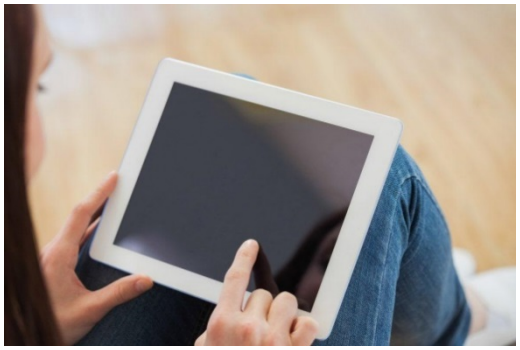
    Ⓑ Once every few weeks

    Ⓒ About once a week

    Ⓓ Two or three times a week

    Ⓔ Every day or almost every day

This is an example of a tablet. A tablet is bigger than a smartphone. It allows you to do many of the same things as a smartphone and a laptop, but it does not make phone calls. If you use a tablet at school, you might use it to do things such as reading books or practicing math problems.



13. Do you have a tablet at home?

    Ⓐ Yes

    Ⓑ No

14. In this school year, how often have you used it **for science learning**? Count all the times you did this at home, at school, in an afterschool program, or anywhere else.

    Ⓐ Never or hardly ever

    Ⓑ Once every few weeks

    Ⓒ About once a week

    Ⓓ Two or three times a week

    Ⓔ Every day or almost every day

15. **At home**, do you have **Wi-Fi** or some other **internet connection** you can use?

    Ⓐ Yes

    Ⓑ No

16. Were you taught any of the following **at school**? Fill in **one** oval on each line

	Yes	No
a. How to type on a computer keyboard using the correct fingers	Ⓐ	Ⓑ
b. How to write sentences and paragraphs using a computer	Ⓐ	Ⓑ
c. How to edit text using a computer	Ⓐ	Ⓑ
d. How to search for information on the Internet	Ⓐ	Ⓑ
e. How to use a tablet	Ⓐ	Ⓑ
f. How to draw a picture using a computer	Ⓐ	Ⓑ
g. How to look up the meaning of a word using a computer	Ⓐ	Ⓑ
h. How to create a spreadsheet using a computer	Ⓐ	Ⓑ
i. How to create a presentation using a computer	Ⓐ	Ⓑ
j. How to run simulations using a computer	Ⓐ	Ⓑ
k. How to write a computer program or app	Ⓐ	Ⓑ
l. How to create a graph or chart using a computer	Ⓐ	Ⓑ
m. How to maintain a website or blog	Ⓐ	Ⓑ
n. How to install new software or apps	Ⓐ	Ⓑ
o. How to troubleshoot problems with a computer	Ⓐ	Ⓑ

17. Which best describes the way you type on the computer keyboard?

Ⓐ I don't know how to type using a computer keyboard.

Ⓑ I can type with one or two fingers, but I have to search for where the letter keys are.

Ⓒ I can type with one or two fingers, and I know where most of the letter keys are.

Ⓓ I can type with all ten fingers when I look at the keyboard.

Ⓔ I can type with all ten fingers without looking at the keyboard.

18. Compared with other students in your English language arts class, how fast do you type on a computer keyboard?

Ⓐ I am slower than most students.

Ⓑ I type about the same speed as others.

Ⓒ I am faster than most students.

Ⓓ I don't know.

19. Would you rather take a test at school using paper and pencil or a computer?
    - Ⓐ Paper and pencil
    - Ⓑ Computer

20. How often do you feel you can understand what the teacher talks about in science class?
    - Ⓐ Never or hardly ever
    - Ⓑ Sometimes
    - Ⓒ Often
    - Ⓓ Always or almost always

21. How often do you feel you can do a good job on your science tests?
    - Ⓐ Never or hardly ever
    - Ⓑ Sometimes
    - Ⓒ Often
    - Ⓓ Always or almost always

22. How often do you feel you can do a good job on your science assignments?
    - Ⓐ Never or hardly ever
    - Ⓑ Sometimes
    - Ⓒ Often
    - Ⓓ Always or almost always

23. For each item, please circle only one answer that best represents your opinion or belief.

Item ID Number	Prompt	Response Options and Coding
V01	Knowing science is important for:	4=all jobs 3=most jobs 2=a few jobs 1=no jobs
V02	Knowing science helps me understand how the world works:	4=all the time 3=most of the time 2=sometimes 1=never
V03	Thinking like a scientist will help me do well in:	4=all my classes 3=most of my classes 2=a few classes 1=none of my classes
V04	I think scientists are the most important people in the world.	4=YES! 3=yes 2=no 1=NO!
V05	I think science is more important than anything else.	4=YES! 3=yes 2=no 1=NO!
V06	Science makes the world a better place to live.	4=YES! 3=yes 2=no 1=NO!
V07	Knowing science is important for being a good citizen.	4=YES! 3=yes 2=no 1=NO!
V08	I think science ideas are valuable.	4=YES! 3=yes 2=no 1=NO!

24. Please indicate how much you DISAGREE or AGREE with the following statements about science. Fill in **one** oval on each line.

	Strongly disagree	Disagree	Agree	Strongly agree
a. I do science-related activities that are not for schoolwork.	Ⓐ	Ⓑ	Ⓒ	Ⓓ
b. I like science.	Ⓐ	Ⓑ	Ⓒ	Ⓓ
c. Science is one of my favorite subjects.	Ⓐ	Ⓑ	Ⓒ	Ⓓ
d. I take science only because I have to.	Ⓐ	Ⓑ	Ⓒ	Ⓓ
e. I need to do well in science to get the job I want.	Ⓐ	Ⓑ	Ⓒ	Ⓓ
f. I would like a job that involves using science.	Ⓐ	Ⓑ	Ⓒ	Ⓓ