Innovations in Science Map, Assessment & Report Technologies

I-SMART

Designing, Developing, and Evaluating Innovative Science Assessments: Evidence from the I-SMART Project

Meagan Karvonen
Lindsay Ruhter
Michelle Shipman
Russell Swinburne Romine
Gail Tiemann

The authors wish to thank the many ATLAS and CAST staff members who have contributed to this project. Additionally, we thank the many classroom teachers who offered their time and expertise to develop and review items and testlets. Finally, we would like to thank the I-SMART project governance board, including current and former state education agency representatives and the project technical advisors listed below, for their ongoing support and guidance. We acknowledge them all for their contributions.

I-SMART Project Governance Board

State Education Agency Representatives

Ann Hermann, Maryland State Department of Education
Michael Plummer, Maryland State Department of Education
Shaun Bates, Missouri Department of Elementary and Secondary Education
Lisa Sireno, Missouri Department of Elementary and Secondary Education
John Boczany, New Jersey Department of Education
Kimberly Murray, New Jersey Department of Education
Kristen Desalvatore, New York State Education Department
Jacqueline Harnett, New York State Education Department
Vanessa Mercado, New York State Education Department
Kurt Johnson, Oklahoma State Department of Education
Todd Loftin, Oklahoma State Department of Education

Former State Education Agency Representatives

Marci Torchon, Maryland
Elizabeth Celentano, New Jersey
Christie Stephenson, Oklahoma

Project Advisors

Karen Erickson, University of North Carolina
Neal Kingston, University of Kansas
Cara Laitusis, ETS
Jim Pellegrino, University of Illinois-Chicago
Michael Wehmeyer, University of Kansas
Phoebe Winter, Consultant

I-SMART Personnel

Research Team

Meagan Karvonen, Principal Investigator, ATLAS, University of Kansas
Russell Swinburne Romine, Co-Principal Investigator, ATLAS, University of Kansas
Gail Tiemann, Co-Principal Investigator, ATLAS, University of Kansas
Jose Blackorby, Co-Principal Investigator, CAST, Inc.
Robert Dolan, Principal Research Scientist, CAST, Inc.
Brooke Nash, Associate Director of Psychometrics, ATLAS, University of Kansas
Sue Buchard, Senior Technical Advisor, ATLAS, University of Kansas

Test Development Team

Michelle Shipman, ATLAS, University of Kansas
Lindsay Ruhter, ATLAS, University of Kansas
Mitch McCann, ATLAS, University of Kansas
Alica Thomas, ATLAS, University of Kansas
Leah Youngren, ATLAS, University of Kansas

Former Test Development Team Members

Lori Anderson
Christine Mars

Preferred Citation: Karvonen, M., Ruhter, L., Shipman, M., Swinburne Romine, R., & Tiemann, G. (2020). *Designing, developing, and evaluating innovative science assessments*. Accessible Teaching, Learning, and Assessment Systems (ATLAS), University of Kansas.

All authors contributed equally to this publication.

The project described in this report was developed under a grant from the U.S. Department of Education. However, the content does not necessarily represent the policy of the U.S. Department of Education, and you should not assume endorsement by the Federal government.

Table of Contents

Introduction

This report describes the process of conceptualizing, designing, developing, and evaluating Innovations in Science Map, Assessment, and Report Technologies (I-SMART) assessments.

The overarching goal of this part of the I-SMART project was to try out new ideas in assessment design that might improve the way the field could measure complex Next Generation Science Standards (NGSS)–aligned performance expectations.

As the only NGSS-aligned operational alternate assessment system in place when I-SMART was funded, the Dynamic Learning Maps® (DLM®) Alternate Assessment System was the starting point for I-SMART design. Thus, some DLM terminology transferred to the I-SMART project. **Essential Elements** (EEs) are the grade-level expectations for students with the most significant cognitive disabilities. Fine-grained **learning map** models include **nodes** that are knowledge, skills, and understandings students master as they develop competence in the science domain. Some nodes are aligned with the EEs. Items are aligned to nodes and grouped together with an engagement activity into **testlets**. Nodes before and after the EE are grouped into **linkage levels** and provide multiple points of access to the content so that all students have the opportunity to demonstrate mastery of content aligned to the EEs.

I-SMART testlets extend beyond the DLM design in several ways:

1. They are designed for a broader population of examinees, including the I-SMART secondary population (i.e., students with or without disabilities who are struggling to meet grade-level expectations). Some testlets had differentiated text complexity to be more appropriate to the secondary population.
2. We evaluated testlet design features that are not common in summative assessments but may be appropriate for more closely aligning assessments to science instruction approaches and engaging learning activities.

This report summarizes the design of the series of I-SMART testlets that were fully piloted, including their purpose, conceptual underpinnings, structure, and development processes. For further details about cognitive labs, see *Evaluating Innovative Science Assessments: Evidence from I-SMART Cognitive Labs*. For a description of the pilot study, see *Evaluating Innovative Science Assessments: Evidence from the I-SMART Pilot Study*. Additionally, information about the I-SMART map neighborhoods can be found in the *Developing and Evaluating Learning Map Models in Science: Evidence from the I-SMART Project.*

Potential Purposes and Use of the Assessments

The purpose of I-SMART assessments is to maximize science achievement and progress across grades for students with the most significant cognitive disabilities and for students with and without disabilities who are consistently not successful with grade-level content. Assessments are designed to measure student learning of rigorous science standards as represented by the continua of knowledge, skills, and understandings in learning map models. Assessments also are designed to maximize student engagement and minimize accessibility barriers. Results are intended to provide timely, accurate, and useful formative feedback about student performance that supports ongoing teaching and learning processes. I-SMART testlets are intended to be used formatively and thus embedded with instruction throughout the year.

The I-SMART project scope was limited to prototype testlets for a sample of EEs. In theory, a fully developed pool of science testlets would span a broad range of EEs, leveraging learning map neighborhoods to support assessments at different levels of complexity. In a fully built I-SMART system, a dashboard would provide teachers with recommendations of content to instruct on and assess next, based on the student's instructional plan or results of prior assessments. The dashboard would also provide examples of instructional resources based on the content being taught. Teachers could also manually choose testlets for the student in the system based on their knowledge of the student's goals and current level of understanding. We prioritized formative uses over summative for the I-SMART project and thus do not make claims that results should be used for summative purposes. However, lessons learned from the I-SMART project could inform the development of future summative science assessments.

Conceptual Framework

The conceptual framework for the I-SMART assessment builds on our prior work integrating evidence-centered design (ECD) and Universal Design for Learning (UDL; Bechard et al., 2019). Basic ingredients of ECD include domain analysis, domain modeling, and a conceptual assessment framework (Mislevy et al., 2003). The learning maps and research narratives that were developed for I-SMART (Swinburne Romine et al., 2018) support these ECD components. For instance, the ECD process includes descriptions of the connections between the student's knowledge, skills, and understandings and assessment results to provide evidence that supports inferences that will be made from those results. The assessment tasks are designed to gather evidence to support these inferences.

In addition to the ECD framework, I-SMART expands on the application of the UDL principles previously used in the DLM assessment system to make content accessible to students with significant cognitive disabilities. The UDL Guidelines (CAST, 2018) describe how to make teaching and learning accessible and engaging via three principles (i.e., multiple means of engagement, representation, and action and expression), which are further described through nine guidelines.

The I-SMART assessments use the same online test delivery platform as the DLM system. The platform incorporates UDL options that have been used in the DLM system, as well as new features that were developed for this project. The online delivery system provides means for customizing the test experience to address specific student needs that were identified in a Personal Needs and Preferences Profile.

Literature Review Results

I-SMART staff conducted a literature review for the purpose of reviewing and extending the DLM project's prior conceptualization of testlet design. I-SMART staff queried common academic search databases, paying special attention to newer literature (i.e., published after DLM testlets were first designed) on focused topics such as students with disabilities, including those with significant cognitive disabilities, students performing below grade level, science assessment and instruction, and the concepts of UDL, cognition, motivation, and engagement. Results contributed to the conceptual framework used to guide development of new testlet innovations and prototypes.

Promoting Interest and Engagement

I-SMART staff reviewed literature that could provide guidance on effective strategies for promoting student interest and engagement in the testlets. Key to this literature was a four-phase model of interest, including triggered situational interest, maintained situational interest, emerging individual interest, and well-developed individual interest (Hidi & Renninger, 2006). The phases suggested how interest could be triggered during an assessment experience.

Support for introducing choice in support of student interest and engagement came from experimental psychology and literature review studies. The literature supported judicious use of choice to promote intended outcomes and prevent unintended consequences (Iyengar & Lepper, 2000; Patall et al., 2008; Patall et al., 2014). Self-regulation literature specific to students with disabilities also supported use of choice in promoting access to the general curriculum (Korinek & deFur, 2016).

The literature presented theories for use of multimedia to support interest and engagement (Mayer, 2003). Evidence-based strategies for multimedia use included (1) restricting extraneous use of words, sounds, and pictures; (2) using conversational style rather than formal style; and (3) using conceptually relevant information (Mayer et al., 2001). Additionally, the gamification of learning literature supported use of narrative (Abdul Jabbar & Felicia, 2015; Dickey, 2005; Kiili, 2005; Vasalou et al., 2017) and personally relevant topics to maintain interest and engagement in a multimedia experience (Ainley & Ainley, 2011).

Cognitive Load Considerations

A key aim of the I-SMART project was to consider test design innovations that would enhance interest and engagement without introducing new accessibility barriers for students. Cognitive load theory provided some guidance to meet this goal. Cognitive load theory indicates that working memory load determines the effectiveness of learning. For example, processing of very complex, high-element interactivity information can easily exceed working memory capacity and produce comprehension difficulties, especially for novice learners (Pollock et al., 2002).

Literature suggested that processing constraints experienced by students with disabilities could be managed by splitting complex tasks into smaller segments, increasing content knowledge (e.g., chunking, deep processing, problem schemata), decreasing information load (e.g., external memory aids, brief and simple instructions, multiple sensory channels), and using external representations (Solaz-Portolés & Sanjosé López, 2009). Additionally, literature supported presenting information in small segments and without redundant information (Mayer & Moreno, 2003), with relevant diagrams (Carlson et al., 2003).

Accessible Item Design Considerations

Accessibility is conceptualized as the sum of interactions between test features and individual test-taker characteristics (Beddow et al., 2013). To enhance accessibility and remove barriers, the literature supports enhancements to assessment items (Cohen et al., 2013; Dickenson et al., 2013). Enhancements can offset deficits in certain areas of cognition, including working memory, sustained attention, executive function, and cognitive effort (Cohen et al., 2013).

Literature-supported item enhancements included bolding or underlining text, boxing items within reading passages, using graphic organizers, using relevant pictures, modifying graphics (enlarged or simplified), use of white-space, and cuing (Cohen et al., 2013).

Science Instruction Considerations for Test Design

Limited literature was found concerning science instruction for students with significant cognitive disabilities that reflected the expectations in multidimensional science standards like the NGSS standards. I-SMART project staff did glean some implications for testlet design, including supports for self-regulation and providing performance feedback. Although the overall literature review provided insight into student engagement and cognitive load, there was little evidence specific to NGSS-aligned science instruction and assessment for students with significant cognitive disabilities.

Summary of Literature-Based Guidance for Test Design

The literature suggested that enhancements could be made to test design to increase interest and engagement while maintaining accessibility. Additionally, strategies could be used to manage any increased cognitive load that could result from item enhancements.

Key design considerations included making test design enhancements relevant to students and aligned to the content conceptually, as well as taking great care to keep the test free from extraneous words, pictures, or media.

Design Input from Project Partners

Literature review results were reviewed during the October 2017 in-person project governance meeting held in Baltimore, Maryland. The Project Governance Board includes roughly equal numbers of state education agency staff from I-SMART partner states and advisors with expertise in several related areas (e.g., NGSS, large-scale assessment, psychometrics, and students with significant cognitive disabilities).

Members of the Project Governance Board noted that some testlet design ideas and enhancements may help some students while at the same time being less ideal for others. Members encouraged test designers to provide options for the presentation of sensory information for specific types of students. For example, students with autism may need media and audio separately instead of concurrently as some literature suggested (Mayer, 1997).

Members also suggested that media (video, pictures) limit the number of concepts or practices demonstrated and that design emphasis should be placed on sustained engagement. Additionally, members recommended that the assessment design team consider the time demands for students and educators given the potential for expanding engagement and lengthening testlets beyond what was already familiar in the DLM science assessments. Members thought that the testlets should gather the needed data as efficiently as possible.

Finally, members recommended that design decisions should be included in the testlet specifications, including a description of the nature of the items; the cognitive demand; where the testlet features are grounded in the literature, testlet type, or associated familiar contexts; and alignment to NGSS/EEs.

Overall Assessment Design

Assessment Targets

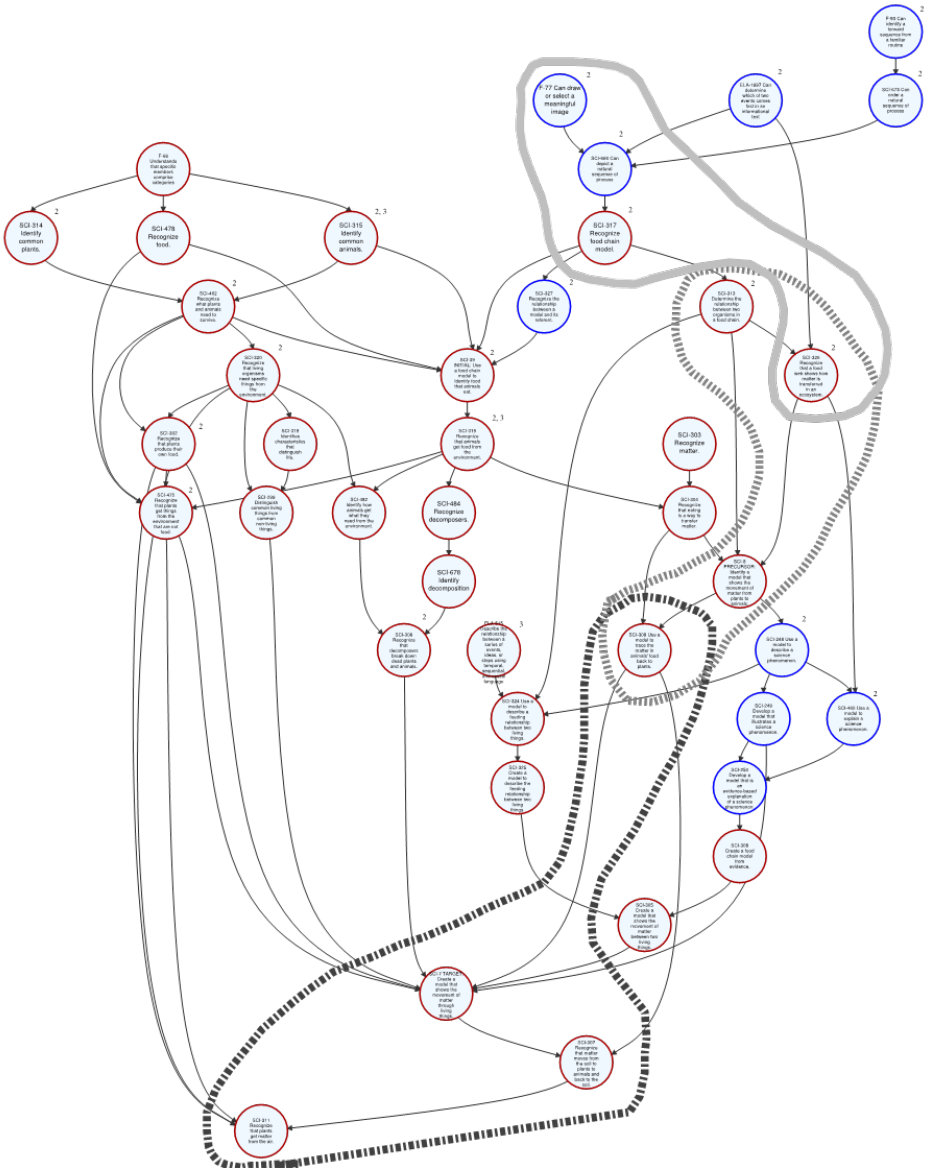
Assessment targets are nodes selected from within learning map neighborhoods. Eleven neighborhoods that had been developed earlier in the I-SMART project were used in this process.[1] Each learning map describes pathways of development toward grade-level expectations.[2] These learning map structures support design variations in assessment content complexity when they are used to select linkage levels.

Testlets measure EEs at four levels of complexity, known as linkage levels. The four linkage levels are Initial, Distal (high school grade band only), Precursor, and Target. Each linkage level is a group of four assessment targets or nodes. The set of linkage levels for each EE span the learning map neighborhood from beginning to end, thereby providing access points for students with a wide range of levels of understanding to optimize cognitive accessibility. Nodes were selected for assessment based primarily on content considerations (i.e., that the progression of linkage levels reflected increasingly complex disciplinary core ideas [DCI] and science and engineering practices [SEP] content, and that each linkage level aligned with the DCI and SEP for the EE). To support psychometric modeling, the nodes in each linkage level included at least one internal direct connection and the adjacent linkage levels overlapped by one node, as illustrated in Figure 1.

[1] See *Developing and Evaluating Learning Map Models in Science: Evidence from the I-SMART Project* for more information about the 11 learning map model neighborhoods.

[2] For the primary population (students eligible for DLM assessments), the grade-level expectation is the EE for the chronologically appropriate grade. For exploration with the secondary population (struggling learners), we used the same maps but identified assessment targets based on instructional match: content matches based on teacher report of skills students had mastered and not yet mastered, regardless of enrolled grade.

Figure 1
Sample Neighborhood Map with Nodes Selected for Assessment at Three Linkage Levels



The NGSS are multidimensional, addressing the DCI, SEP, and the crosscutting concepts. The I-SMART assessments are multidimensional and assess the DCI and SEP for each EE. Each assessment was a blend of unidimensional and multidimensional items tested together in the context of a science narrative.

Testlet Structure

In I-SMART assessments, items are grouped together into testlets. Each testlet has 12–16 items that address a single EE and a single linkage level for the EE. Three to four items were written to each node within a linkage level. Each testlet holistically measures elements of both a DCI and a SEP.

I-SMART testlets were reconceptualized to expand on the established practices of ECD and UDL from the DLM system. Additional UDL options were incorporated at varying degrees of complexity across linkage levels.

Science Narrative

The science narrative is an important component of most I-SMART testlets. The science narrative provides a theoretical situation for students to explore during the testlet. The narrative is designed to promote student engagement, activate prior knowledge, and provide appropriate background knowledge for the student to avoid bias in testing outside of the assessed concepts. Science narratives are designed to be based in circumstances familiar to students or that reflect typical classroom experiences. First, the student is given a context and evidence, usually through a narrative description of a student engaging in science inquiry, and the student is charged with interpreting or making meaning of the tested concept. Items ask students to make connections between the SEP and the DCI.

Essential Element Concept Maps

The EE concept maps (EECMs) are task templates that outline the relationship between testlet elements, such as concepts, nodes, and student questions, using principals of ECD for learning. Originally designed for use in the DLM assessment system, the EECMs provide item writers with information to help them make sense of the intricate knowledge, skills, and understandings from the learning map model. An EECM includes information about an EE and provides a learning map neighborhood mini-map to show the connection between the Target, Precursor, Distal (high school grade band only) and Initial linkage levels. In general, mini-maps depict the critical knowledge, skills, and understandings, and their order of development, representing the essential junctures or stages of learning toward the mastery of a specific grade-level academic target. The EECMs also include key terms and ideas, descriptions of the nodes and examples of student mastery of the nodes, prerequisite and requisite skills, common misconceptions that can be used when creating distractor response options, common questions to ask, and possible accessibility barriers for students (Bechard et al., 2019).

The EECMs are designed to ensure that the testlets are aligned with the cognitive processes represented by the map nodes (Bechard et al., 2019). The EECMs allow item writers to create testlets that are effective, consistent, and accessible and offer a framework that easily allows for modifications to the item-writing process. Item writers used EECMs to connect all aspects of linkage levels for each EE. The EECMs provided guidance for the item writers in their selection of the phenomenon to anchor the testlet and in the creation of a “wonder question” (a question posed to students to reevaluate at the end of the testlets). The use of EECMs also ensured that specifications guided the writing of content to the same EE and linkage level, thus making testlets exchangeable.

EECM Design Revisions

Developing the I-SMART EECMs was a collaborative effort among the I-SMART research and assessment development teams with input from the Project Governance Board. I-SMART assessments use a similar test design model as DLM alternate assessments that includes the alignment of items to linkage level nodes in the learning map model, grouping of items into short testlets that share a common engagement activity, and the continued application of principles of UDL. To support increased student engagement, measuring multiple linkage level nodes within

a single testlet, and developing testlets that are representative of scientific inquiry processes, a reconceptualized EECM was needed. Using DLM testlet development models as a starting point, the I-SMART testlet design framework was expanded to make changes to the structure of testlets. These changes led to an increase in student interest and engagement by supporting student decision-making and self-regulation. These are skills that are required for SEPs and support student choice-making as a part of the assessment experience. The UDL guidelines developed by CAST influenced the EECM redesign. An iterative process was used by the I-SMART research and assessment development teams during the redesign process. The teams collaborated about the most essential components of the principles of ECD and a format that would yield succinct and easily understood information for item writers. The Project Governance Board was asked to provide feedback about the approach of the EECM redesign work, address specific design questions, and provide suggestions throughout the EECM reconceptualization process.

New features were added to the I-SMART EECMs beyond the template used in prior work for the DLM system. The purpose of the new features is to make more explicit connections between item design elements, EEs, and learning map nodes in addition to providing guidance on applying the principles of UDL. Enhancements to EECMs included (1) extended information related to the DCIs and SEPs for each EE, (2) visual representations of small areas of the learning map to provide context for each linkage level, (3) examples of how one might observe the cognitive skill specified by each learning map node, and (4) UDL options to enhance student engagement, self-regulation, and comprehension. Six EECMs were developed to support I-SMART item writing, and 11 EECMs were created by the end of the project.

Final EECM Design

The I-SMART EECMs were redesigned to support changes in the test development framework. The redesigned EECM begins with a title page naming the EECM, the science domain and topics, and connections to the NGSS. See Appendix A for an EECM example. The title page is followed by three sections that each contain information about one linkage level. Each linkage-level section contains assessment targets (nodes) aligned to both the DCI and the SEP for the EE. The nodes provide a visual means of identifying the relationships between skills. The node descriptions give a detailed overview of the skill or skills each node assesses, whereas the node observation describes what the behavior looks like for a student who is achieving the skill represented by each node and what evidence of mastery might look like for students.

The redesigned EECM also includes (1) vocabulary that describes terms needed to demonstrate understanding of the science content, (2) example questions to elicit evidence of student understanding, (3) progression information that describes the increase in complexity in both DCI- and SEP-related skills in each assessed linkage level, and (4) a section about how to select and frame an appropriate, engaging, and accurate science phenomenon to be used as the context for each testlet. Additional information describes the DCI and SEP components at each linkage level. The EECM includes potential UDL options that are specific to each linkage level and are focused on applications of principles of UDL to the specific science content for the EE. These include options for recruiting interest, language and symbols, comprehension, self-regulation, and multiple means of expression and communication. The interconnection of UDL applications within the EECM provides item writers with an accessible, comprehensive set of test specifications to support effective item construction. As an applied tool, the EECM leverages the value of a theory-grounded, intentional design process in visual format that

supports nonprofessional item writers as they engage in creating high-quality items and develop their knowledge of the learning map models.

Populating I-SMART EECMs

The assessment design team used excerpts from learning map model research narratives to populate sections of the EECMs. The narratives are based on extensive reviews of cognitive and developmental empirical research, common instructional practices, and other curricular information. Thus, the EECMs include research-based information about the EE, the nodes, and the connections among nodes within each linkage level. Information from the NGSS was also used to inform many sections of the EECM. For example, information from the NGSS was used to populate the header of each EECM, which describes the three dimensions (DCI, SEP, and cross-cutting concept) and appropriate vocabulary and concepts associated with each science standard. Essential questions, misconceptions, progression information, and linkage levels were developed using learning map neighborhood narratives that detail the content, SEP, and cross-cutting concepts associated with each EE. The wonder question, phenomenon, and example questions to ask were developed by the research and assessment design teams. The UDL Guidelines (CAST, 2018) provided examples of UDL options from which to select that served the content and the linkage level assessed. Nodes assessed at each linkage level were included, as well as descriptions and observations, which were created with the learning map neighborhoods. A mini-map, which includes untested and tested nodes for the EE, was included. An excerpt from the learning map neighborhood narrative provides additional context for how the depth, breadth, and complexity develop over the combined linkage levels for the EE.

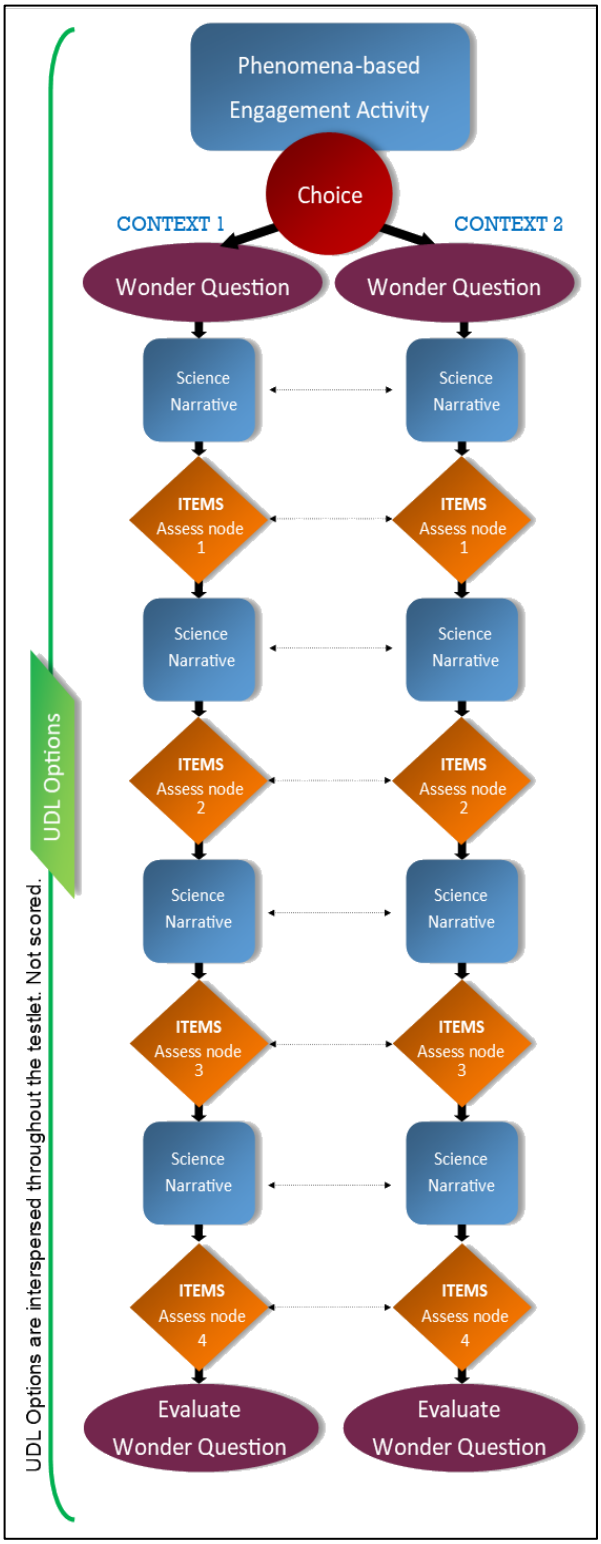
Prototype Testlets

Prototype Testlet Structure

We designed prototype testlets to be evaluated through cognitive labs before deciding on final testlet design for the rest of the I-SMART project. The prototype testlets were intentionally designed to include many UDL options (Figure 2). The

Figure 2

Prototype Testlet Structure



prototype testlets include phenomenon-based engagement activities, wonder questions, and opportunities for self-evaluation. Testlets contain a science narrative centered around a phenomenon that provides context for the science concepts to be assessed. The wonder question introduces students to the overarching line of inquiry contained in the testlet. Items ask students to make connections between the SEP and the DCI. In Initial, Distal (high school grade band only), and Precursor, the student chooses from two related contexts, construct-relevant and construct-irrelevant, within the phenomenon. For the Target testlets, an extended narrative was incorporated to enrich the context of the testlet and increase student engagement. The Target testlets were also adapted for the secondary population and contain higher-level vocabulary and science terms. Figure 2 illustrates the ordered structure of the prototype testlets at the Precursor linkage level.

Prototype testlets represented multidimensional science content at elementary, middle school, and high school grade bands at Initial, Distal (high school only), Precursor, and Target linkage levels. To keep the size of the cognitive lab study within scope of the project timelines, prototype testlets were not created at all combinations of linkage levels and grade bands. Table 1 describes the content of the prototype testlets by grade band. All testlets contained three to four scored items per node, with four nodes per testlet.

Table 1
Prototype Testlet Content by Grade Band

Grade band	Essential Element	Essential Element description	Linkage levels available	Number of nodes
Elementary	LS2-1	Create a model that shows the movement of matter (e.g., plant growth, eating, composting) through living things.	Initial Target	4
Middle school	LS2-2	Use models of food chains/webs to identify producers and consumers in aquatic and terrestrial ecosystems.	Precursor Target	4
High school	LS2-2	Use a graphical representation to explain the dependence of an animal population on other organisms for food and their environment for shelter.	Initial Target	4

Prototype testlets were developed for cognitive labs based on two templates: choice-based and elaborated science narrative. Choice-based testlets were offered to students at the Initial and Precursor linkage levels. Elaborated science narrative testlets were offered to students at the

Target linkage level. Testlet options based on the UDL framework were intended to support student interest and engagement.

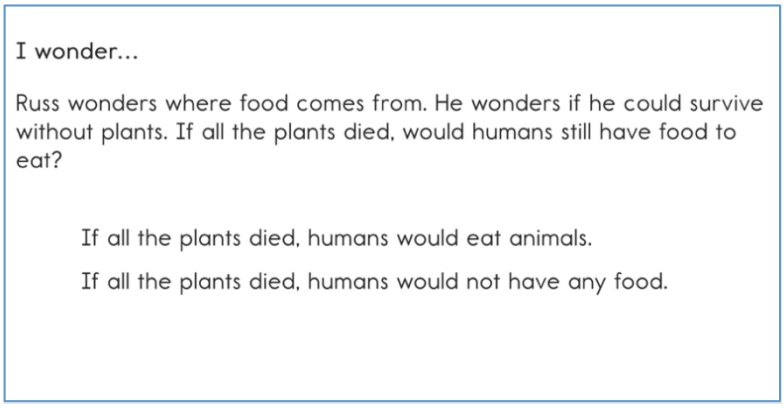
UDL Features and Item Types

Choice Options. Choice options were presented to students at the Initial and Precursor linkage levels. Student choice, corresponding to the UDL guideline of recruiting interest, allows for students to pick a context of interest and potentially increase student engagement. The choice-based testlets had two variations. In the first variation, construct-relevant, the choice was related to the construct, such as the choice between two different animals or ecosystems. For the second variation, character-related (or construct-irrelevant), the choice was related to a character. Once chosen, the remainder of the testlet corresponded with the selected choice. Both variations were tested during cognitive labs.

Elaborated Science Narrative. Elaborated science narrative testlets contained a slightly more complex narrative than the choice-based testlets. Science narrative testlets aligned with the more advanced skills and knowledge needed at the Target linkage level. These testlets provided students with opportunities to pause and think about the topic to engage curiosity and natural interest.

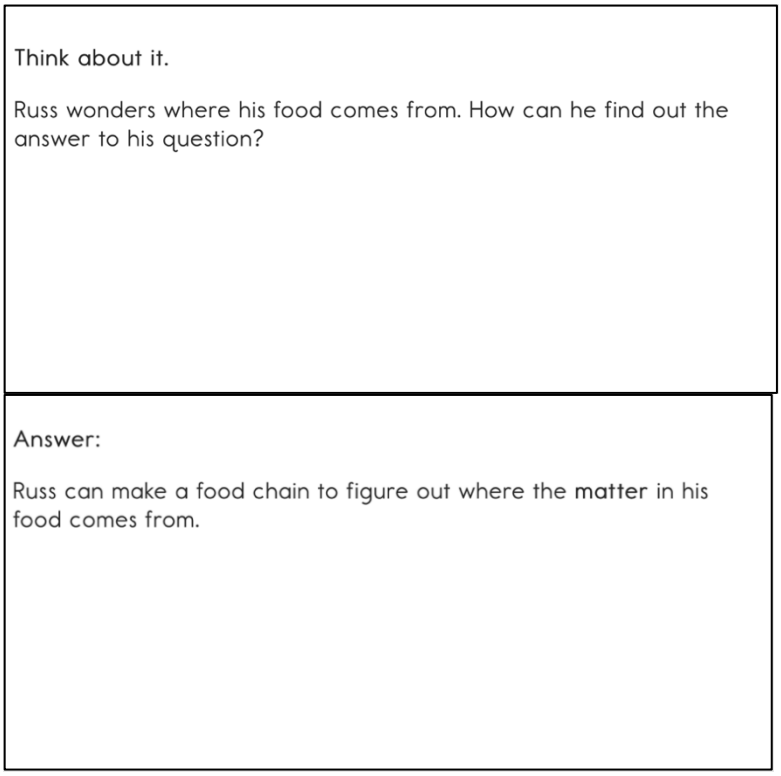
I Wonder. At the Precursor and Target linkage levels, a wonder question asked students to reflect on important science concepts based on a science phenomenon introduced in the science narrative. The I Wonder question presented two, unscored answer options and appeared at the beginning and the end of the testlet so students could reevaluate their original answer. The question is designed to increase both students' interest and capacity for progress monitoring as they reflect on their understandings after progressing through a testlet. An example of a wonder question is displayed in Figure 3.

Figure 3
Example of I Wonder



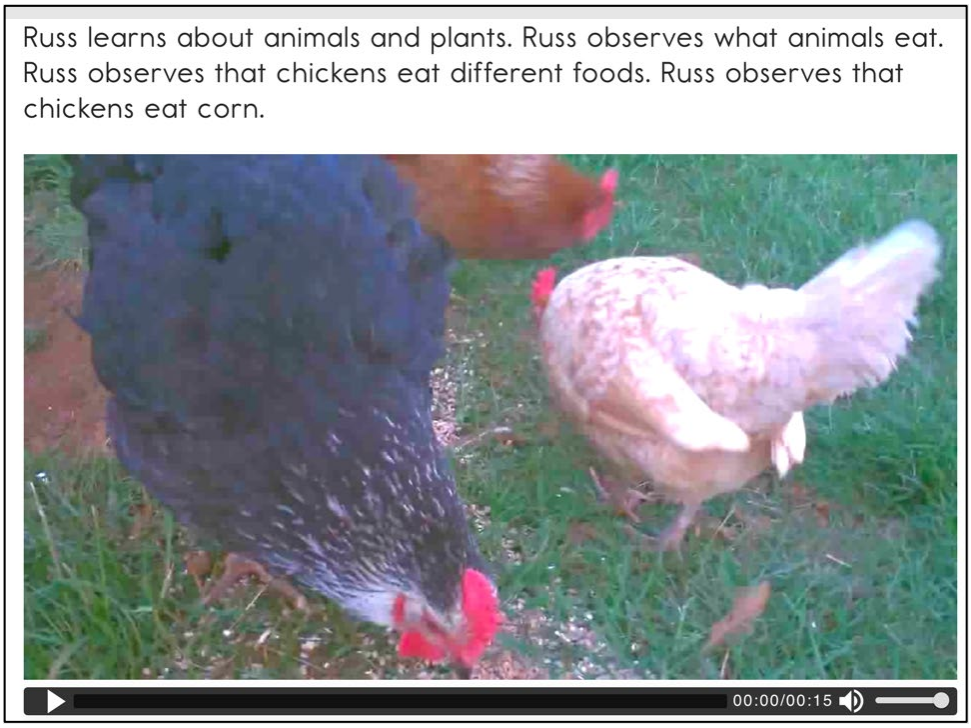
Think About It. Think About It questions were included in Precursor and Target linkage level testlets as unscored items embedded in the testlet that asked students, "What should you do next?" or "How would you find this answer?" The feature was designed to support students' planning and strategy development (UDL guideline of executive function) and increase engagement as an active learner. The Think About It feature posed a question on the first screen without answer options. On the next screen, an answer was presented. An example Think About It feature can be found in Figure 4.

Figure 4
Think About It Screens



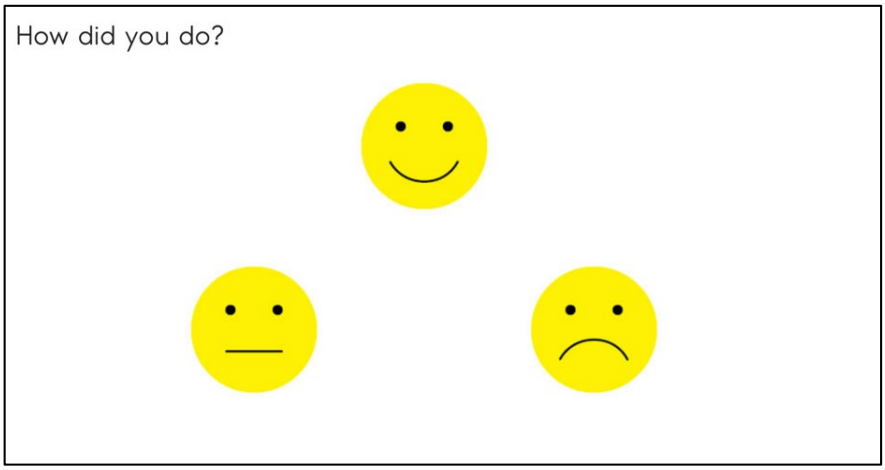
Use of Video. Videos were available in Precursor and Target linkage level testlets. Videos were included as options for multiple means of expression and support in decoding information and increasing student engagement. Comprehending the content of the video was not required to answer items, though the videos were directly related to the narrative presented in the testlet. An example of a video is shown in Figure 5.

Figure 5
Use of Video Example



Self-Assessment. A self-assessment item was presented at the end of the testlet. The unscored self-assessment item offered emoticon images in a familiar layout for students to indicate how they performed. This unscored item allowed students at the Initial linkage level to practice the skill of reflecting on their performance, but at a lower level of complexity than required for the I Wonder at the Precursor and Target linkage levels. The self-assessment item is displayed in Figure 6.

Figure 6
Self-Assessment Item



Item Types. Prototype testlets primarily relied on single-select multiple-choice item types for scored items. Drag-and-drop items were used in the computer-delivered testlets at the Precursor and Target linkage levels at both middle school and high school grade bands.

Cognitive Labs

Twenty-five cognitive labs were conducted in three I-SMART partner states during 2018 and 2019. The purpose of the cognitive labs was to explore the prototype testlet features with students and to collect response process evidence. Response process evidence supports the extent to which students interact with testlet content as intended and whether the proposed item formats introduced construct-irrelevant variance through increased response process demands. A full description of the cognitive lab methods and results can be found in the report, *Evaluating Innovative Science Assessments: Evidence From I-SMART Cognitive Labs.*

Cognitive labs explored the following research questions (RQs):

1. How do students interact with the features of innovative item types and with innovative testlets?
2. How much time is required to complete a testlet?
3. Do students' responses represent the science performance expectations the items were designed to measure?
4. What are students' and teachers' perceptions of students' experiences with the new testlets?

Cognitive labs were conducted on nine prototype testlets across three linkage levels in three groups: (1) choice-based testlets based on construct-relevant narrative options, (2) choice-based testlets based on preferred characters, and (3) elaborated science narrative testlets.

Cognitive Lab Results Overview

Results of cognitive labs are summarized below.

RQ 1: How do students interact with the features of the innovative item types and with innovative testlets?

Choice Options. Neither variant option, construct-relevant nor character-related, was clearly preferred by students; thus, the construct-relevant option was chosen for further development by item writers and for subsequent cognitive lab sessions.

I Wonder. The wonder question was a novel feature that some middle school students found unfamiliar and thus struggled with. Students at the high school level had less difficulty; however, most administrators provided unintentional extra prompts that there was a question on the screen to answer.

Use of Video. Usability concerns surfaced from the lab sessions, including trouble with playing the video (it required scrolling) and a delay in the video page loading. Across middle school and high school levels, seven students indicated that they generally liked the video.

Item Navigation. At the middle school level, students encountered little difficulty with the drag-and-drop items. However, one item was displayed as selected response but appeared very similar to a drag-and-drop item, which caused many students to have difficulty responding.

RQ 2: How much time is required to complete a testlet?

The prototype testlet design presented more items than had been used before in DLM science testlets; thus, the total amount of time needed for each student to complete the lab session was documented. Each testlet contained between 14 and 17 items. Testing times ranged from 11 minutes and 47 seconds to 29 minutes and 28 seconds, which were within acceptable ranges, even with the inauthentic nature of the lab-based testing experience. Based on cognitive lab testing times, the testlet specifications were retained for the I-SMART pilot.

RQ 3: Do students' responses represent the science performance expectations the items were designed to measure?

Related to testlet content, comments from students based on think aloud and retrospective probes supported that students were generally interpreting content as intended. Some guessing and use of process of elimination is expected from students, but many were able to articulate the intended response processes from the item specifications. Students at the Initial linkage level were also able to answer some questions correctly and some without construct-irrelevant behaviors, especially with familiar content. Teachers of students at the Initial linkage level were concerned that some of the testlet content was too difficult.

RQ 4: What are students' and teachers' perceptions of students' experience with the new testlets?

I-SMART staff interviewed students (where possible) and teachers about their experiences with the testlets using a semi-structured interview protocol.

In general, students and teachers had positive opinions concerning testlet content and use of media. Teachers noted some areas for improvement, which led to the implementation of several navigation and layout updates. Some concerns about accessibility of concepts were also shared with I-SMART staff for additional consideration.

Updates to Testlet Design From Cognitive Lab Results

Based on cognitive lab results, the research and assessment design teams made several design decisions before developing testlets for the pilot phase. They chose to move forward with only the construct-relevant student choice option. Additionally, the acceptable range of testlet completion times indicated no change was needed to the general design specifications.

At the Initial linkage level, scripted prompts or statements were added for test administrators to reduce confusion around how to engage students. In addition, some shared reading strategies were included, such as, "Sally learns about seals. What is your favorite animal?" Shared reading strategies were included to increase student engagement and relevance for students taking Initial linkage level testlets. With the addition of shared reading strategies, science stories were reduced in length to ensure the testlet remained at an appropriate level of cognitive demand.

Additionally, the design team decided that text complexity and vocabulary levels for the extended-narrative Target linkage level testlets should be increased to better align with the needs of the secondary population of students.

Pilot Testlet Development

Item writers developed testlets during the I-SMART item-writing event in Baltimore, Maryland, in July 2018. After the initial development work, the testlets were reviewed by internal subject matter experts and prepared for an external panel review.

Item Writers

Working with state partners and Project Governance Board members, the assessment design team recruited practicing educators to serve as item writers. Item writers were selected from within I-SMART partner states based on (1) areas of expertise, including general education science, special education science, and special education; (2) subjects taught in the last 3 years, including in life, physical, and Earth and space science; and (3) years of experience.

In sum, 12 educators were selected as item writers. Selected special educators had 4–22 years of service, and general education educators' experience ranged from 6 to 30 years. Seven item writers were female and five were male. Item writers also represented all five I-SMART state partners, including Maryland (three item writers), Missouri (two item writers), New Jersey (three item writers), New York (two item writers), and Oklahoma (two item writers). Eight item writers identified as White, three as Black or African American, and one as having multiple ethnicities.

Purpose

The purpose of the onsite event was to (1) provide item writers with an understanding of the science learning map neighborhoods, (2) train item writers to use resources that incorporated all aspects of innovative testlet design, and (3) produce high-quality testlets.

Item-Writer Training

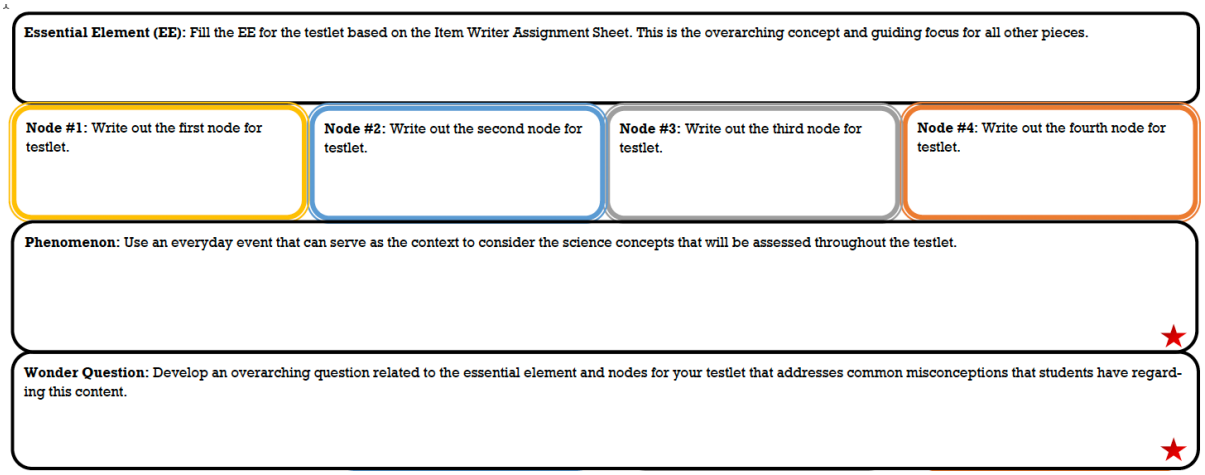
The assessment design team provided asynchronous online training prior to the event and additional onsite training. The trainings included an overview of the I-SMART project, examples of UDL features, innovative design features, and strategies for writing engaging science narratives based on a science phenomenon to provide context for items. In addition, item writers were trained in key concepts of writing items that measure the intended multidimensional construct (DCI and SEP), accessibility considerations for all students (i.e., accessible text and graphics), and awareness of possible causes of bias and sensitivity issues (i.e., fair representation, free of controversial issues, free of stereotypes).

Item-Writing Resources and Procedures

Many aspects of the I-SMART item-writing event were influenced by approaches previously used by the DLM assessment system. The event was structured to allow item writers to construct their own knowledge about key concepts for producing high-quality testlets. Item writers were placed together in pairs based on their expertise with grade-band science content or student population. This structure promoted collaboration throughout the item-writing process. Item-writing pairs worked together to understand the construct and brainstorm the science narrative and science phenomenon for the context. As each item writer drafted the innovative UDL features and multidimensional items, they completed informal feedback checkpoints, followed by a structured peer review of the testlet using the EECM specifications as the measure. See the I-SMART Peer Review Checklist in Appendix B.

The assessment design team designed and developed innovative item-writing resources that displayed the key information for the EE and linkage level. Using the EECM specifications, multiple items measuring the same EE and linkage level can be written to the same specifications, thus building efficiency and accuracy into the item-writing process and supporting the development of well-aligned items. Item writers used other resources to draft testlets following the guidelines for the innovative testlet features. Item writers used the UDL Engagement Testlet Features resource to provide guidance on features that could be included in testlets at each linkage level. See the UDL Engagement Testlet Features resource in Appendix C. Item writers used the I-SMART Storyboard Organizer partially shown in Figure 7 to begin building the foundations of the testlet. Once item writers finished drafting the testlet, they entered the testlet narrative, items, and graphics into a I-SMART Content Brainstorm Template, which are technology-based testlet templates. See an example of an I-SMART Target Storyboard Organizer and an I-SMART Target Computer-Delivered Content Brainstorm Template in Appendix D.

Figure 7
I-SMART Storyboard Organizer



Post–Item Writing Steps

The testlets received an editorial review, including a content alignment check, and accessible graphics were developed. Next, testlets were entered into the content management system and prepared to be externally reviewed by educators.

Testlet-Development Outcome

Item writers and the assessment design team developed 40 testlets—34 testlets for the primary population and six testlets for the secondary population. Among the 40 testlets, 623 items were developed. Of those, 538 were scored items and 85 were unscored items. Table 2 lists the count of testlets written by grade band, EE, and target population.

Table 2
*Count of Testlets Developed by Grade Band, Essential Element, and Target Population (*N = *40)*

Grade band and Essential Elements	Primary population	Secondary population
Elementary		
EE.5.LS2-1	5	1
EE.5.PS1-3	5	1
Middle school		
EE.MS.LS2-2	5	1
EE.MS.PS1-2	5	1
High school		
EE.HS.ESS3-3	7	1
EE.HS.LS2-2	7	1

At the beginning of the choice-based testlets, students pick a context of interest. Because that context carries forward into the remainder of the testlet, choice-based testlets were written as a single testlet and treated as two separate testlets for the purpose of external review. The 40 unique testlets included 14 choice-based testlets, thus a total of 54 testlets were externally reviewed.

Testlet External Review

Many aspects of the I-SMART external review event were influenced by approaches previously used by the DLM assessment system. The purpose of the I-SMART external review event was to evaluate items and testlets for content, accessibility, and bias and sensitivity. Panelists completed an online advance training course prior to a face-to-face meeting in Baltimore, Maryland, in September 2018. The event consisted of onsite training that included a practice activity, panelists individually rating items and testlets for the criteria they were assigned, and consensus discussions and recommendations when there was a discrepancy in the criteria ratings for items or testlets. After discussion, a group consensus rating and recommendation was submitted.

Two tables of panelists, assigned to elementary/middle school and middle school/high school grade bands, reviewed the content criteria. One table of panelists reviewed the accessibility criteria, and one table of panelists reviewed the bias and sensitivity criteria. The assessment design team members were the room facilitators and delivered the training. Two additional test development coordinators, a research project manager, and a co-principal investigator were the table facilitators.

Review Criteria

The I-SMART External Review Criteria contain specific criteria to evaluate content, accessibility, and bias and sensitivity of the items and testlets. The DLM External Review Criteria was the starting point for creating the I-SMART External Review Criteria. Additional criteria were incorporated into each review type to evaluate the increased use of technology-enhanced items and UDL options. See the I-SMART External Review Criteria in Appendix E.

Content panelists examined the alignment of the items to the construct of the assessed nodes. The vocabulary and graphics within the items were evaluated for accuracy. Panelists ensured items had only one correct response option and distractors that were not misleading. Panelists examined if the testlets were instructionally relevant. Panelists reviewed the Cognitive Process Dimension of each item. The DLM Science Cognitive Process Dimensions (DLM Consortium, 2018), was the base for creating the I-SMART Cognitive Process Dimensions. The purpose and instructions were updated and noted with asterisks. See the I-SMART Cognitive Process Dimensions in Appendix E.

Accessibility panelists considered if the text provided an appropriate level of challenge, was aligned with grade-level content standards, and used clear and easily comprehensible language and graphics. Panelists reviewing for accessibility considered whether the testlets presented barriers to students who take the I-SMART assessments. Panelists examined if the testlets were instructionally relevant.

Bias and sensitivity panelists examined items for bias, establishing that an assessment item did not require knowledge outside of the targeted content, establishing that items did not use stereotypes and also used people-first language, and establishing that there was representation of diversity in ethnicity, gender, disability, and family composition. Panelists examined testlets for sensitivity, determining them to be free of controversial or disturbing content due to issues including culture, gender, religion, ethnicity, and socioeconomic status.

Panelists

External review panelists were recruited from I-SMART partner states based on their professional experience and their student population expertise or content expertise. Demographic characteristics for external review panelists included gender, current grade-level teaching, content area expertise, years teaching, and education. Most external reviewers had earned a master's degree, and reviewers represented a range of teaching backgrounds, including all grade bands, and both general and special education experience. Item writers were not eligible to serve as external reviewers. Panelist demographic characteristics are summarized in Table 3.

Table 3
*Demographic Characteristics of External Review Panelists (*N = *18)*

Characteristic	*n*
Gender	
Male	4
Female	14
Grade band	
Elementary	8
Middle school	7
High school	6
Multiple grade bands	5
Content area expertise	
General education	11
Special education	6
Both general education and special education	1
Teaching experience (in years)	
1–5	3
6–10	1
11+	14
Highest degree obtained	
Bachelor's	2
Master's	13
Specialist	2
Doctorate	1
National board certified	1

Training (Advance/Onsite)

Panelists completed advanced training prior to attending the I-SMART external review event. The advance training included a general overview of the I-SMART external review process, an example specific to the criteria a panelist was assigned to, and the procedures for completing ratings. Panelists completed a quiz and practice activity after finishing the advance training.

The onsite training included an overview of the advance training, guided practice with table facilitators, and a review of available resource materials.

Rating Process

The DLM External Review Process Diagram and the DLM Guide to External Review of Testlets were the foundation for the creation of the-SMART External Review Process Diagram and the I-SMART Guide to External Review of Testlets. See the I-SMART External Review Process Diagram and the I-SMART Guide to External Review of Testlets in Appendix F. To begin the rating process, the assessment design team assigned items and testlets to panelists for independent review. Panelists rated each assessment item and testlet as accept, revise, or reject. Panelists used the criteria for their assigned panel type to determine if the items and testlets met minimum standards for field-testing readiness and indicated their recommendations using the following rating scale:

- Accept: The item or testlet is within acceptable limits for field testing.
- Revise: The task or testlet violates one or more of the review criteria; however, the task or testlet has potential merits and can be acceptable for field testing after revisions to address the criteria.
- Reject: The content of the task or testlet is fundamentally flawed; revision would not bring the task or testlet to acceptable limits.

If there were any independent revise or reject ratings, the facilitators led a discussion with the panelists at the table. After the panelists reached consensus about the items or testlets with revise or reject ratings, the table facilitator recorded the panelists' final decision and recommendation.

Remote External Review

Due to the complexity and large quantity of testlets, the external review process took longer than expected at the onsite event. Therefore, additional remote external review was necessary after the onsite event. Two additional panelists, who serve regularly as content and special education experts for DLM, completed the additional reviews. The remote panelists were qualified for all three types of panel review. Their expertise is included in Table 3. Remote panelists completed the online advance training course and received the same resources as the onsite panelists prior to rating testlets. Each remote external review panelist completed the review for all three panel types.

External Review Results

The external review results include the onsite and remote panel ratings: 5.61% (*n* = 35) of the 623 items and 3.70% (*n* = 2) of the 54 testlets received accept ratings across all three panel types. Table 4 displays the panel ratings for items and testlets per panel type.

Table 4
Item and Testlet Ratings by Panel Type (Item N = *623, Testlet* N = *54)*

Panel type and rating	Items *n*	Items %	Testlets *n*	Testlets %
Content				
Accept	425	68.22	24	44.44
Revise	168	26.97	29	53.70
Reject	30	4.82	1	1.85
Accessibility				
Accept	362	58.11	21	38.89
Revise	243	39.00	32	59.26
Reject	18	2.89	1	1.85
Bias and sensitivity				
Accept	541	86.84	29	53.70
Revise	65	10.43	25	46.30
Reject	17	2.73	0	0.00

Evaluation of the Panel-Review Process

Panelists evaluated the external review process, with 100% rating the online advance training, guide to external review, and discussion with other panelists as “effective” or “very effective.” All panelists (100%) “agreed” or “strongly agreed” that the staff members were knowledgeable about the academic content, that the experience proved a valuable professional development opportunity, and that they would participate in future events.

Post-Panel Review Process

The assessment design team processed the external review data and used trends in the data to revise items and testlets. The revision recommendations were used to maximize the quality, instructional relevance, and accuracy of the items and testlets. Many panelists recommended adjusting vocabulary to be clearer and more accessible to all students. Panelists also recommended revising graphics to be more accessible and support student understanding. Additionally, panelists suggested revisions in the science narrative to promote accuracy and reduce complexity of the text.

Based on their analysis of panel recommendations, the assessment design team accepted 487 (78.17%) items and 34 (85.19%) testlets. The team revised 136 (21.83%) items and 6 (14.81%) testlets. No items or testlets were rejected. Most panelist recommendations were concise and explicit, though some marked testlets to be revised when they wanted changes beyond the limits of I-SMART test design, such as including multiple images and tables in the same item. Comments that fell outside the realm of assessment design, such as technology constraints, and recommendations that would be inappropriate for the assessed linkage level (e.g., technology enhanced items at the Initial linkage level) did not factor into team decisions.

Test Design for Pilot Study

Testlets were written as intact units and put together on forms after external review. To support the project’s research design, forms included two testlets presented in order from lower to higher linkage levels within an EE. The design plan included 24 testlets, sampling at least two DCIs in each of three grade bands. For a description of the pilot study, see *Evaluating Innovative Science Assessments: Evidence from the I-SMART Pilot Study*.

Conclusion

The test development objectives for I-SMART provided opportunities to design and evaluate new science assessment features. To begin the design phase, assessment design team members and research team members selected assessment targets for six alternate content standards that each linked to an NGSS performance expectation. Based on the assessment targets, the project produced 11 new EECMs. Item writers, along with assessment design team members, then used the EECMs to develop 54 testlets with 623 items. These items were externally reviewed, and all items and testlets were either accepted or revised.

The design and development phases of I-SMART generated many innovations and lessons learned. Cognitive labs with students generally supported the assessment design, including inclusion of UDL-based features. Responses to items and interview prompts suggested that students were interpreting content as intended. Additionally, students and teachers offered positive comments and useful feedback on the media, testlet content, and assessment design.

Feedback from the I-SMART item writers provided preliminary evidence that the EECMs, advance and onsite training, and design of the event generally produced the desired outcomes. Assessment design team members did find that it was challenging for item writers to be creative in their use and placement of UDL options because of the novelty of the approach. However, providing prototype examples and quality training around the purpose and goals of the UDL options was helpful to writers. Additionally, through the design and development process, research and assessment design members saw ongoing value in the use of UDL guidelines to systematically self-assess the extent to which UDL can be incorporated into the many components of a large-scale assessment system.

Similarly, the external review process proved successful and worked largely as intended, with panelists participating in high-quality, effective discussions. External reviewers were challenged by the novelty of the UDL options and by the large amount of content that needed to be reviewed during a single-day event. While most reviewer suggestions were concise and explicit, other comments were outside of the control of test designers. Overall, panel recommendations were useful to the assessment design team, even though the number of specific revisions enacted was smaller than the number of suggestions.

In sum, I-SMART's innovations expanded on prior evidence that students with and without disabilities can demonstrate what they know and can do when the principles of UDL and ECD are used to develop assessments that maximize accessibility. The iterative processes of I-SMART, along with collaboration among all I-SMART personnel, Project Governance Board members, and educators, led to an innovative design for assessing students in multidimensional content in science with far-reaching potential for future assessment developers.

References

Abdul Jabbar, A. I., & Felicia, P. (2015). Gameplay engagement and learning in game-based learning: A systematic review. *Review of Educational Research, 85*(4), 740–779. https://doi.org/10.3102/0034654315577210

Ainley, M., & Ainley, J. (2011). Student engagement with science in early adolescence: The contribution of enjoyment to students' continuing interest in learning about science. *Contemporary Educational Psychology, 36*(1), 4–12. https://doi.org/10.1016/j.cedpsych.2010.08.001

Bechard, S., Clark, A., Swinburne Romine, R., Karvonen, M., Kingston, N., & Erickson, K. (2019). Use of evidence-centered design to develop learning maps-based assessments. *International Journal of Testing, 19*(2), 188–205. https://doi.org/10.1080/15305058.2018.1543310

Beddow, P. A., Elliott, S. N., & Kettler, R. J. (2013). Test accessibility: Item reviews and lessons learned from four state assessments. *Education Research International*, *2013*. https://doi.org/10.1155/2013/952704

Carlson, R., Chandler, P., & Sweller, J. (2003). Learning and understanding science instructional material. *Journal of Educational Psychology, 95*(3), 629–640. https://doi.org/10.1037/0022-0663.95.3.629

CAST. (2018). Universal Design for Learning Guidelines version 2.2. Retrieved from http://udlguidelines.cast.org

Cohen, D. J., Danielson, L., Stoica, W., Wothke, W., & Zhang, J. (2013). Test development: Item modifications. In M. L. Thurlow, S. S. Lazarus, & S. Bechard (Eds.), *Lessons learned in federally funded projects that can improve the instruction and assessment of low performing students with disabilities* (pp. 205–246). National Center on Educational Outcomes, University of Minnesota.

Dickenson, T. S., Gilmore, J. A., Price, K. J., & Bennett, H. L. (2013). Investigation of science inquiry items for use on an alternate assessment based on modified achievement standards using cognitive lab methodology. *The Journal of Special Education, 47*(2), 108–120. https://doi.org/10.1177/0022466911414720

Dickey, M. D. (2005). Engaging by design: How engagement strategies in popular computer and video games can inform instructional design. *Educational Technology Research and Development, 53*, 67-83. https://doi.org/10.1007/BF02504866

Dynamic Learning Maps Consortium. (2018). DLM Science Cognitive Process Dimensions. Accessible Teaching, Learning, and Assessment Systems, University of Kansas.

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist, 41*(2), 111–127. https://doi.org/10.1207/s15326985ep4102_4

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology, 79*(6), 995–1006. https://doi.org/10.1037/0022-3514.79.6.995

Kiili, K. (2005). Digital game-based learning: Towards an experiential gaming model. *The Internet and Higher Education, 8*(1), 13–24. https://doi.org/10.1016/j.iheduc.2004.12.001

Korinek, L., & deFur, S. H. (2016). Supporting Student Self-Regulation to Access the General Education Curriculum. *TEACHING Exceptional Children*, *48*(5), 232–242. https://doi.org/10.1177/0040059915626134

Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions? *Educational Psychologist, 32*, 1–19. https://doi.org/10.1207/s15326985ep3201_1

Mayer, R. E. (2003). The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and Instruction, 13*(2), 125–139. https://doi.org/10.1016/S0959-4752(02)00016-6

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. https://doi.org/10.1037/0022-0663.93.1.187

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 43–52. https://doi.org/10.1207/S15326985EP3801_6

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Patall, E. A., Cooper, H., & Robinson, J. C. (2008). The effects of choice on intrinsic motivation and related outcomes: a meta-analysis of research findings. *Psychological bulletin*, *134*(2), 270–300. https://doi.org/10.1037/0033-2909.134.2.270

Patall, E. A., Sylvester, B. J., & Han, C. W. (2014). The role of competence in the effects of choice on motivation. *Journal of Experimental Social Psychology, 50*, 27–44. https://doi.org/10.1016/j.jesp.2013.09.002

Pollock, E., Chandler, P., & Sweller, J. (2002). Assimilating complex information. *Learning and Instruction, 12*(1), 61–86. https://doi.org/10.1016/S0959-4752(01)00016-0h

Solaz-Portolés, J. J., & Sanjosé-López, V. (2009). Working memory in science problem solving: A review of research. *Revista Mexicana de Psicologia, 26*, 79–90.

Swinburne Romine, R., Andersen, L., Schuster, J., & Karvonen, M. (2018). *Developing and Evaluating Learning Map Models in Science: Evidence from the I-SMART project*. Accessible Teaching, Learning, and Assessment Systems (ATLAS), University of Kansas.

Vasalou, A., Khaled, R., Holmes, W., & Gooch, D. (2017). Digital games-based learning for children with dyslexia: A social constructivist perspective on engagement and learning during group game-play. *Computers & Education, 114*, 175–192. https://doi.org/10.1016/j.compedu.2017.06.009

Appendix A: Essential Element Concept Map Example

Essential Questions for the Target Level
- Does the student understand that the food for most animals can be traced back to plants?
- Does the student understand that matter cycles between the air and soil and among plants and animals?
- Does the student understand that matter that is not food is changed by plants into matter that is food?
- How can the movement of matter through living things be described in a model?

Target Level Name	Target Level Description
EE.5.LS2-1.T	Create food-chain models and use food-chain models to trace matter from the environment to plants, through living things, from animals' food to plants, and from the soil to plants to animals and back to the soil.

Vocabulary	
Concepts	movement of matter in an ecosystem
Words	food chain, plant, soil

Misconceptions

(SCI-309) The student does not understand that the arrow shows the direction of matter movement.
(SCI-7) The student does not recognize indirect relationships in a food chain/web containing more than two organisms.
(SCI-7) The student views the relationships in food webs as simple cause and effect relationships (eating and growing) rather than of the movement of matter within an ecosystem.
(SCI-307) The student thinks dead things decay naturally without the action of decomposers.
(SCI-307) The student thinks that dead things disappear.
(SCI-311) The student does not understand that air is matter (i.e., believes air does not have weight or take up space).
(SCI-311) The student believes that plants get matter (i.e., food) from soil or fertilizer (i.e., plant food provided by people).

Progression Information	Phenomenon	Wonder Question	Science and Engineering Practices
Food-chain models are used to trace matter in animals' food back to plants (LS2.A & Developing and Using Models)-> Create a model that shows how matter moves through living things (LS2.A & Developing and Using Models) -> Understand that matter moves from soil to plants to animals and back to soil (LS2.B) -> Recognize that plants get matter from air (LS2.B) ->	General mechanism: Plants get matter mainly from air and water. Plants get nutrients from soil (i.e., minerals). Matter moves through ecosystems as plants make their own food, plants are eaten by animals, animals eat other animals, and dead things are broken down by decomposers to make matter available to plants again. Not all matter is passed to the next level; some matter leaves at each level (e.g., excretion, breathing). Examples of systems include specific organisms or ecosystems. Example Phenomena: In a forest, oak trees grow. Oak trees make their own food with matter from air and water. Oak trees take in air through openings in their leaves. Oak trees also get nutrients from the soil. Squirrels eat leaves and acorns from the oak tree. Owls eat squirrels. Dead animals and leaves are broken down by worms and become part of the soil.	The wonder question is presented at the beginning and revisited at the end of the testlet. The wonder question connects to a research-based misconception that can be resolved through inquiry activities in the testlet. Example: What would happen to animals if all the plants died? Two answer options are presented, one of which is a common misconception. Example: AO1 - Animals that eat other animals would survive. (misconception) AO2 - No animals would survive. Over the course of the testlet, students should gain information that will help them reevaluate the wonder question at the end of the testlet. The flow of items should build to items closely related to the wonder question answer.	*Developing and Using Models* Students create models (e.g., food chain/web) that describe the movement of matter through living things. They use the models to trace the matter in animals' food back to plants. Food chain/web models use arrows to show the direction that matter moves between two living things. Animals eat organisms that they are directly connected to in the chain/web. Organisms depend on organisms that are directly connected in the chain/web as well as those that are indirectly connected because the same matter moves across the entire food chain.

The contents of this document were developed under a grant from the Department of Education. However, those contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government.

UDL Options	
Principle	**Guideline: Description**
Representation	Language & Symbols: Vocabulary support (i.e., definitions of terms from earlier nodes in the map, multiple representations). Language & Symbols: Use of video and/or images to support decoding. Language & Symbols: Language in science narrative is concise and appropriately complex given grade level and linkage level. Items provide student clear directives and expectations to demonstrate knowledge and understanding. Comprehension: Science narrative provides background knowledge, big ideas, and relationships Information may be accessed through physical scale models (e.g., tactile displays) and/or computer-generated models. Represent relationships with diagrams representing only the most relevant information.
Engagement	Recruiting interest: Phenomenon is a common, high-interest situation that a student might experience, makes connections to the real world. Pedagogically relevant, age-appropriate contexts for testlets optimize relevance, value, and authenticity. Self-regulation: Wonder question follows engagement activity and ties to typical misconception. Return to wonder question at the end of the testlet. Self-regulation: Model reflective practice (i.e., think aloud) question at end of testlet such as "Did the testlet show what I know?" or "How did I like it?" Items asking students to reflect on performance develop self-assessment and reflection.
Action & Expression	Expression & Communication: Variety of item response types included, such as multiple-choice, drag and drop, and multiple-select multiple choice. Executive Functions: Data is presented in tables or graphs to facilitate organization and managing of information. Executive Functions: Questions embedded in science narrative inquiry activity to support strategy development, such as "What should you do next?" or "How would you find this answer?"

Target linkage level nodes

Nodes should be presented in the testlet in an order that creates a logical flow in inquiry activity, which may differ from the order that they occur in the map.

Nodes (order from map)	**Description**	**Observation & Example Questions to Ask**
SCI-309 Use a model to trace matter in animals' food to plants.	Linking node Integrated Node 4 items Use a model to trace the matter in animals' food back to plants.	The student is presented with a simple food web (e.g., grass ->rabbit -> fox). The student identifies that the matter in the fox's food came from grass. Example Questions: What does the model show about how the [organism] gets matter? Which model shows how [organism] gets matter?
SCI-7 Create a model that shows the movement of matter through living things.	Integrated Node 3 items Create a model that shows the movement of matter (e.g. plant growth, eating, composting) through (three or more) living things.	The student is shown a partially complete food-chain model (e.g., one organism or arrow is missing). The student is asked to fill in the missing item based on the description of the feeding relationships from an engagement activity story. Example Questions: Which food chain shows how matter moves? Put the plants and animals in the correct box to show how matter moves [drag and drop item]. What goes between [organism1] and [organism2] to show how matter moves [AOs are types of arrows]?
SCI-307 Recognize that matter moves from the soil to plants to animals and back to the soil.	DCI Node 3 items Recognize that matter moves from the soil to plants to animals and back to the soil.	When shown an example of a cycle food web (e.g., grass ->rabbit ->fox ->worm) the student identifies that the web shows that matter moves from grass to rabbit to fox to worm to soil. The student identifies that plants get nutrients from the soil, but not matter. [Note: Confusing food and nutrients is a misconception. Nutrients for plants are like vitamins for people.] Example Questions: What does the model show about how the [organism] gets matter? What does [character's] food chain show about matter?
SCI-311 Recognize that plants get matter from the air.	DCI Node 3 items Recognize that plants get matter from the air (i.e., carbon dioxide).	When asked, "How does a plant get material it needs to grow?", the student indicates that plants get matter (carbon dioxide) from the air. For example, when asked, "How does a tree get material it needs to grow?", the student indicates that trees take in air through their leaves to get the material they need to grow. Example Questions: What helps a [plant] get matter? How does [plant] get matter to grow? How does a [plant type] take in the material it needs to grow? What is the material that [plant type] uses to grow?

Accessible Teaching, Learning, & Assessment Systems. (2019, December). *Essential Element Concept Map: EE.5.LS2-1.T.* Lawrence, KS: University of Kansas. Retrieved from https://ismart.works/resources

*I-SMART utilized the DLM Essential Element Concept Map originally developed for the DLM alternate assessment system.

Appendix B: Peer Review Checklist

I-SMART Peer Review Checklist

Use the I-SMART Peer Review Checklist while reviewing your partner's testlet. Use the comment feature in the Testlet Brainstorm Template PowerPoint to leave feedback. Provide detail about the issues and specific revisions.*

Checks on Universal Design for Learning (UDL)

- **Teacher Administered Testlets:** Testlet utilizes a shared reading approach throughout the science narrative to engage students.*
- **Computer Delivered Testlets:** The phenomenon provides the context for the Wonder Question. The Wonder Question includes potential content misconceptions students may have about the Disciplinary Core Idea (DCI).*
    - The wonder question asks students to make a prediction at the beginning and re-evaluate that prediction at the end of the testlet.*
- Incorporated **UDL options** are appropriate for the linkage level and increase engagement and access to the content (i.e., make testlet more engaging and allow student to demonstrate what they know).*

Checks on Alignment and Accessibility

- Items **align to nodes** by only assessing the concept(s) and/or practices(s) listed in the node to reduce cognitive demand.*
- The **phenomenon** chosen is a familiar context and leads to a pedagogically productive instructional activity.*
- Science **narrative screens and items** progress logically throughout the testlet.*
- The **science narrative** provides the student an opportunity to engage with the science concept and the SEP to answer items.*
- **Language complexity** is appropriate based on the grade band, EECM, linkage level, and target audience.*
    - The **language** in the testlet maintains the link to grade-level content without introducing unnecessary, confusing, or distracting verbiage.
    - The **word choices** for the stem and answer options in all items are logical, concise, and avoid cueing the answer to different items.
    - **Content specific vocabulary** provides an appropriate level of challenge while maintaining a link to grade-level content.*
- Testlet uses simple **sentence structure**, especially for students in the I-SMART primary population, even when the concept tested is challenging.
- **Graphics and images** in the testlet are logical, appropriate, and do not introduce extraneous information.

*I-SMART utilized the peer review checklist originally developed for the DLM alternate assessment system. The peer review criteria developed specifically for I-SMART are noted by an asterisk.

Checks on Bias and Sensitivity

- Concepts assessed do not depend on a particular **sensory capability**.*
- Items are free of **bias** toward a particular subgroup of students and do not include
- **stereotypes**.
- Items do not require **background knowledge** outside of the construct being tested.*
- Items offer a fair **representation of diversity** in ethnicity, gender, disability, and family composition.
- The testlet avoids content that could potentially cause an **extreme emotional reaction**.
- Avoid sensitive content in science narrative and items unless necessary to assess the node.

*I-SMART utilized the peer review checklist originally developed for the DLM alternate assessment system. The peer review criteria developed specifically for I-SMART are noted by an asterisk.

Appendix C: UDL Engagement Testlet Features

UDL Engagement Testlet Features

Testlets may contain the following methods of recruiting interest:

Linkage Level	Methods of recruiting interest
Initial	Choice of context Picture response cards Shared reading
Distal	Choice of context Videos, self-reflection (e.g., Think About It, How Did You Do?)
Precursor	Choice of context Wonder Question Videos, self-reflection (e.g., Think About It, How Did You Do?)
Target	Wonder Question Increased text complexity Videos, self-reflection (e.g., Think About It, How Did You Do?)

The contents of this document were developed under a grant from the Department of Education. However, those contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government.

Appendix D: Storyboard Organizer and Target Computer-Delivered Content Brainstorm Template

Innovations in Science Map, Assessment & Report Technologies

I-SMART

I-SMART Target Storyboard Organizer

Use this storyboard to begin formulating the context for the testlet. Review each testlet feature below to connect how each feature interacts and informs overall testlet design. Consider how the progression of specific features in the testlet will flow within the context of the phenomenon chosen. Consider how Universal Design for Learning options will be utilized in individual tasks that will assess the content in each node. Items may build with increasing complexity throughout each node assessed and throughout the testlet. Collaborate with your partner at each step that is outlined in the I-SMART Testlet Diagram (designated by a red star). Consider the cyclical nature of the testlet development process and your personal work preferences to begin building the foundation of the testlet in the storyboard before entering the content into the I-SMART Content Brainstorm Template.

Essential Element (EE): Fill the EE for the testlet based on the Item Writer Assignment Sheet. This is the overarching concept and guiding focus for all other pieces.

Node #1: Write out the first node for testlet.	**Node #2:** Write out the second node for testlet.	**Node #3:** Write out the third node for testlet.	**Node #4**: Write out the fourth node for testlet.

Phenomenon: Use an everyday event that can serve as the context to consider the science concepts that will be assessed throughout the testlet.

Wonder Question: Develop an overarching question related to the essential element and nodes for your testlet that addresses common misconceptions that students have regarding this content.

Narrative #1:	**Narrative #2:**	**Narrative #3:**	**Narrative #4:**
Item #1:	**Item #1:**	**Item #1:**	**Item #1:**
Item #2:	**Item #2:**	**Item #2:**	**Item #2:**
Item #3:	**Item #3:**	**Item #3:**	**Item #3:**
Item #4:	**UDL Options:** Think About It, Vocabulary Support, Video Engagement		
Item #5 (Optional):			
Item #6 (Optional):			

Evaluation: Return to the Wonder Question and Consider the successful understanding/attainment of the skills outlined in that question.

I-SMART Target Computer-Delivered Content Brainstorm Template

Item Writer Name:

Testlet Name:

Essential Element:

Linkage Level:

Node(s):

Screen #:

PHENOMENA-BASED ENGAGEMENT ACTIVITY SCREEN

Text:

Graphics Request (brief image request description):

Screen #:

WONDER QUESTION

Text:

Answer Options:

A.

B.

Graphics Request (image and image name from Image Library, or brief image request description):

Key (Enter the letter and text of the correct answer option. Reminder: Wonder Questions not scored):

Screen #:

SCIENCE NARRATIVE SCREEN

Text:

Graphics/Video Request (brief image request description):

Screen #:

TASK

Text:

Answer Options:

A.

B.

C.

Graphics Request (brief image request description):

Node:

Key (Enter the letter and text of the correct answer option.):

Screen #:

THINK ABOUT IT SCREEN

Text:

Graphics Request (brief image request description):

Screen #:

THINK ABOUT IT ANSWER

Text:

Graphics Request (image and image name from Image Library, or brief image request description):

Screen #:

EVALUATE WONDER QUESTION

Text:

Answer Options:

A.

B.

Graphics Request (brief image request description):

Key (Enter the letter and text of the correct answer option.
Reminder: Wonder Questions not scored):

Screen #:

HOW DO YOU THINK YOU DID?

Text:

Answer Options:

A.

B.

C.

Graphics Request (brief image request description):

Appendix E: External Review Criteria and Cognitive Process Dimensions

External Review Criteria

Items and testlets use the Innovations in Science Map, Assessment & Report Technologies (I-SMART) External Review Criteria for evaluation. There are three external review panel types: content, accessibility, and bias and sensitivity. Each external review panel has criteria to consider for items and other criteria to consider for testlets. The I-SMART External Review Advance Training Moodle course and I-SMART External Review Practice Activity provide training on the criteria.

Criteria for Content Panels

Items

1. The item assesses the content of the targeted node.
2. The level of Cognitive Process Dimension (CPD) is appropriate for the node. CPD is listed as “Cognitive Category” in Content Builder.
3. The content of the item is technically correct (text, images, and graphics).*
4. Item answer options should contain only one correct answer option (the key) unless the item is a multiple-choice multiple-select item, which has two-three correct answer options. The distractors are incorrect and include typical student misconceptions where appropriate. Nothing in the item cues the correct answer.*
5. The item type is logical and appropriate for the targeted content.
6. Images and graphics (if used) contribute to the quality of the item.

Testlets and Science Narratives

1. The testlet as a whole is instructionally relevant and is grade-level appropriate.
2. The context of the science narrative*
    a. Is grade-level appropriate and pedagogically relevant*
    b. Enables students’ use of the Science & Engineering Practice (SEP) to gain information about the Disciplinary Core Idea (DCI).*
3. The science narrative and items progress in a logical order.*
4. The Wonder Questions appropriately introduce the targeted science concepts, frame the science narrative, and allow the student to self-evaluate understanding of the science concepts assessed.*

Criteria for Accessibility Panels

Items

1. The text within the item provides an appropriate level of challenge and maintains a link to grade-level content without introducing unnecessary, confusing, or distracting verbiage.
2. The text of the items require inferences only when appropriate for the assessed SEP and prior knowledge to comprehend the targeted content.*
3. Images and graphics are clear and do not introduce confusion.
    a. Images and graphics contain only the basic necessary elements.*
    b. Items that reference graphs should limit information to what is necessary to respond to the item.*
    c. Graphics can be presented in tactile form and described in alternate text.
4. The item type is accessible and allows for an accurate response.*

Testlets and Science Narratives

1. The testlet is instructionally relevant and grade-level appropriate.
    a. The text within the testlet provides an appropriate level of challenge and maintains a link to grade-level content without introducing unnecessary, confusing, or distracting verbiage.*
    b. Images and graphics are clear and do not introduce confusion. Graphics can be presented in tactile form and described in alternate text.*
2. The testlet does not introduce barriers for students with
    a. Limited working memory
    b. Limited implicit understandings of others' emotions and intentions
3. UDL options enhance access and support learner performance, without introducing possible barriers to student success (e.g. cognitive load, distraction from the targeted content). Options support the intended learning objectives to be measured.*
    a. For language & symbols: supports understanding of vocabulary from earlier nodes in the map or decoding.*
    b. For comprehension: provides appropriate background knowledge, big ideas, or relationships (e.g., science narrative).*
    c. For recruiting interest: features a common, high interest situation that a student might experience and makes connections to the real world (e.g., context of the science narrative).*
    d. For executive function: supports student planning and strategy development in science problem solving (e.g., Think About It question and answer).*
    e. For self-regulation: asks students to reflect on performance to develop self-assessment and reflection (e.g., How Did You Do?).*
4. Flexible options still support the intended learning objectives that need to be measured without introducing additional barriers to accessibility (e.g., cognitive load, relevancy, does not distract from the targeted content).*

Criteria for Bias & Sensitivity Panels

Items

1. Item does not require prior knowledge outside the bounds of the targeted content.
2. Language used does not prevent or advantage any group from demonstrating what they know about the targeted content.
3. Where applicable, there is a fair representation of diversity in race, ethnicity, gender, disability, and family composition.
4. Stereotypes are avoided. Appropriate labels are used for groups of people. People-first language is used for individuals with disabilities.
5. Item does not focus on material that is likely to cause an extreme emotional response.
    a. Avoids sensitive content when possible without limiting access to the node.*

Testlets and Science Narratives

1. Testlet is free of content that is controversial, disturbing, or likely to cause an extreme emotional response due to issues of culture, region, gender, religion, ethnicity, socio-economic status, occupation, or current events.
    a. Avoids sensitive content when possible without limiting access to the node.*
2. The language in the testlet neither prevents nor disadvantages any regional or cultural group from demonstrating what they know about the targeted content.
    a. People-first language is used for individuals with disabilities. Populations are not depicted in a stereotypical manner.

Innovations in Science Map, Assessment & Report Technologies

I-SMART

Taxonomy for Cognitive Process Dimensions

Purpose

The purpose of this document is to explain the definitions and examples of the cognitive process dimension (CPD) levels used in Innovations in Science Map, Assessment & Report Technologies (I-SMART) science testlets. Definitions in this document explain the CPD levels and the examples demonstrate how to apply the CPD levels to I-SMART science content.

Cognitive Process Dimensions

This table summarizes the definitions and examples for each listed CPD level. The CPD column contains the name of each CPD level. The Definitions column presents the definition of the CPD level. The Science-Specific Definitions column defines the CPD level specific to science. The Science Examples column provides examples of tasks for each CPD level.

CPD[1]	Definitions[3]	Science-Specific Definitions[2]	Science Examples[3]
Remember	Retrieve relevant knowledge from long-term memory in a novel context	Identify, recall, list, recognize, or label	When given a diagram of the digestive system, the student can answer the question, "Which is the stomach?" When asked, "Which organs are in the digestive system?" the student lists the major organs (e.g., mouth, esophagus, stomach, intestines). When provided a model of an organ system, the student can label the organs in the model.
Understand	Construct meaning from instructional messages, including oral, written, and graphic communication	Describe or explain in your own words, retell, or summarize	When shown a line of data in a table or a point on a graph, the student can recognize the piece of data as a representation of two related variables. For example, when shown a data table with the characteristics of steel, the student recognizes that the line in the table provides the characteristics for steel. When asked, "What does the table tell us about steel?" the student identifies that the table tells us that steel is magnetic because they can describe what is in the table. When the feeding relationships in a food chain are presented in a science story, the students can describe those feeding relationships.
Apply	Carry out or use a procedure in a given situation.	Predict an outcome using several pieces of information or concepts; use information in a new context	When presented with data about the effectiveness of multiple safety devices, the student can predict the outcome of using a safety device that is included in the data.

CPD[1]	Definitions[3]	Science-Specific Definitions[2]	Science Examples[3]
Analyze	Break material into its constituent parts and determine how the parts relate to one another and to an overall structure or purpose	Infer, interpret, compare, contrast; understand how components relate to each other and to the process as a whole.	When shown two or three lines of data in a table or points on a graph, the student can determine the maximum or minimum value or if values increase or decrease over time (e.g., applies a procedure to science data). When shown a data table or graph, the student uses the information in the table or graph to answer a science question. For example, when presented with data from an experiment that is intended to determine the answer to a science question (e.g., if a device provides improved protection in a collision), the student can determine the answer to the question (e.g., make an inference). Often this type of task involves making meaning of a maximum value, minimum value, or trend.
Evaluate	Make judgement based on criteria and standards	Determine/critique relative value; determine merit or if a hypothesis is supported by data	When shown a data table or graph, the student determines which data are evidence of meeting criteria for success. For example, when presented with data from multiple devices, the student can determine which devices best meet criteria, such as which device is the most effective safety device.
Create	Put elements together to form a coherent or functional whole; reorganize elements into a new pattern or structure	Create something new using/combining disparate sources of information, such as developing a hypothesis, designing an experiment, or creating a model.	Given a description of organisms and their food, the student organizes them into a food chain.

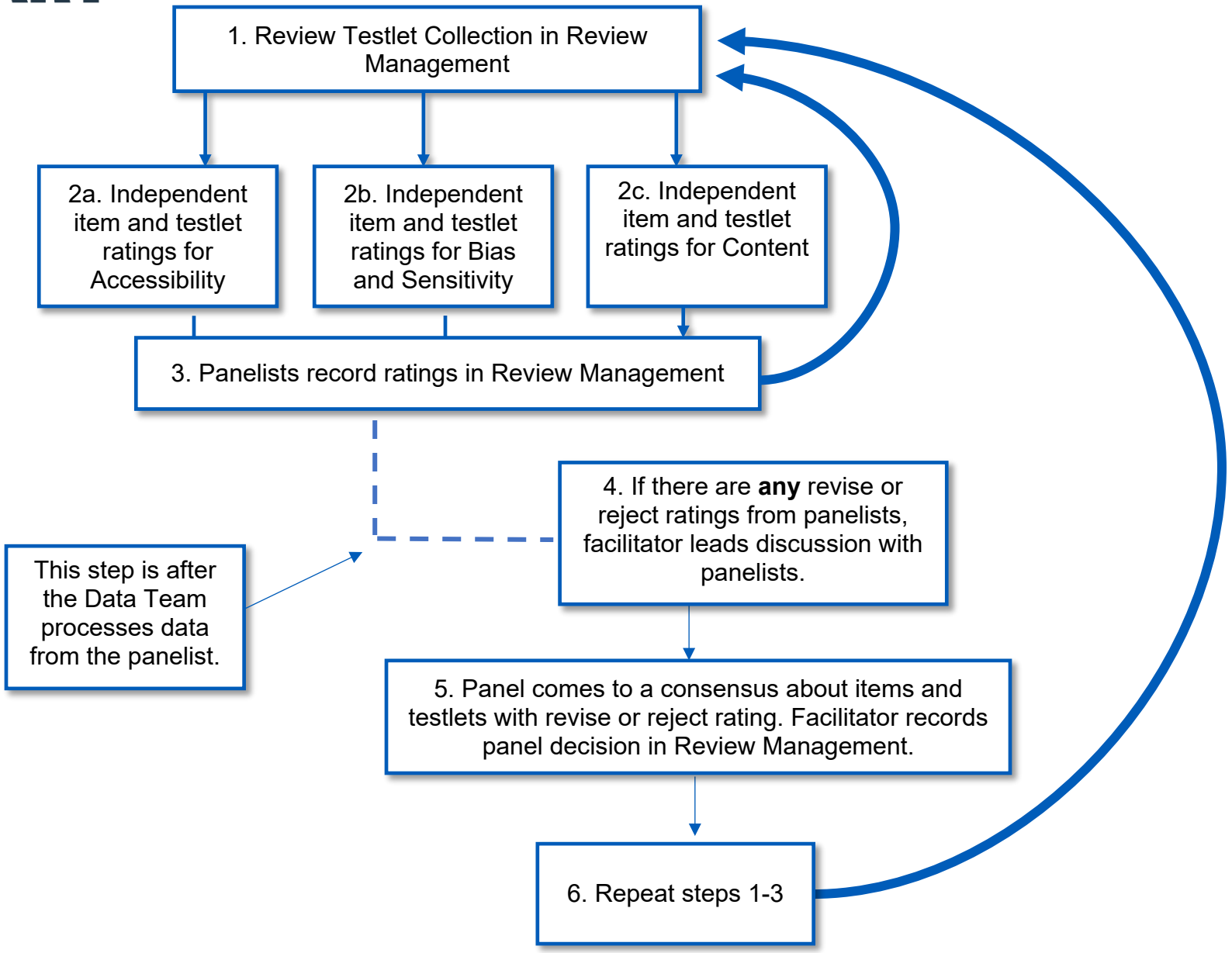
References

[1] Anderson, L. W., Krathwohl, D. R., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., ... & Wittrock, M. (2001). A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy.

[2] Crowe, A., Dirks, C., & Wenderoth, M. P. (2008). Biology in bloom: Implementing Bloom's taxonomy to enhance student learning in biology. *CBE-Life Sciences Education*, *7*(4), 368-381. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2592046/

Appendix F: External Review Process Diagram and Guide to External Review of Testlets

I-SMART External Review



*I-SMART utilized the external review process diagram originally developed for the DLM alternate assessment system.

GUIDE TO EXTERNAL REVIEW OF TESTLETS

VERSION 1.0

AUGUST 2018

*I-SMART utilized the DLM Guide to External Review of Testlets originally developed for the DLM alternate assessment system.

SERVICE DESK

For questions regarding the Kite® system or for additional assistance, please contact the Kite Service Desk at 1-855-277-9751 (toll-free) or email kite_support@ku.edu.

The Kite Service Desk is open Monday through Friday from 7:00 a.m. to 5:00 p.m. Central Time.

The Kite Service Desk provides support for a variety of situations, including problems with review management and login issues.

When contacting the Kite Service Desk, provide as much detail as possible about the issues encountered and the system on which it occurred. Please include

- error messages
- operating system and browser information
- information about network configuration

EXTERNAL REVIEW GUIDE

CONTENTS

INTRODUCTION

PURPOSE OF REVIEW

The purpose of external review is to evaluate testlets developed for the Innovations in Science Map, Assessment & Report Technologies (I-SMART).Using specific criteria, panelists will decide whether to recommend that the content be accepted, revised, or rejected. Feedback from external reviewers is used to make final decisions about items and testlets before they used with students.

OVERVIEW OF REVIEW PROCESS

Members of three distinct review panels conduct external reviews: **Content, Accessibility,** and **Bias and Sensitivity.** Panelists will be assigned to one type of review panel and use the criteria for that panel to conduct their reviews. Each time panelists receive a review assignment, they will evaluate **items** that are grouped together into **testlets**. Testlets are groups of items, bundled with instructions and engagement activities. Some are designed for the student to see directly, while others are designed to guide the teacher through the process of administering a testlet outside the system. I-SMART staff will review the ratings panelists provide for each item and testlet and make a final decision about the testlet.

REVIEW COLLECTIONS AND ASSIGNMENTS

Each review assignment will consist of one collection. Each collection consists of up to 10 testlets. Panelists will review multiple collections.

TRAINING

Before reviewing testlets, panelists must first complete the I-SMART External Review Advance Training Moodle course. The modules in the course include detailed instructions on the review process. Panelists must also complete a quiz after each module. This training must be completed and panelists must score 80% or higher on each quiz before being able to access the practice activity.

PANEL ASSIGNMENT

Based on a panelist's experience and expertise, they are assigned specific criteria to focus on during the external review process of the items and testlets. The three panel types are content, accessibility, and bias and sensitivity. Although panelists are assigned to a panel, they will externally review items and testlets independently. There is no discussion with other panelists during the independent rating times. Following the independent rating times, panelist review the revise and reject items and testlets as a group.

PANELIST RESPONSIBILITIES

The primary responsibility of an external review panelist is to review testlets using established standards and guidelines, which are outlined in the remaining sections of this guide.

Remember that being an external review panelist involves reviewing items and testlets that will be part of a secure testing system. Panelists will be required to complete a **test security and confidentiality statement** before reviewing secure items and testlets. Panelists must maintain the security of materials at all times. In order to maintain the security of these materials

1. Do not make paper or electronic copies of any of the contents from the items or the testlets.
2. Do not use any non-essential electronic devices (i.e., cell phones, iPads, tablets, cameras, etc.) within the proximity of the external review. Do not take screenshots or photographs of the content.
3. Do not talk about the specifics of I-SMART items or testlets with others outside of panel discussions. Direct all questions or concerns to the table facilitator.

Panelists may feel free to talk with others about the I-SMART project in general. Specific features of the system, such as texts, items, and testlets, should **not** be discussed outside of the meeting room.

REVIEW CRITERIA AND DECISIONS

GENERAL REVIEW DECISIONS

For each item and testlet, a panelist should consider this question: **Does this item or testlet meet minimum standards for acceptability based on my panel's criteria?** The I-SMART External Review Criteria document lists each panel's criteria. Acceptability is defined as meeting the minimum standards of the criteria. Panelists will make one of three decisions based on their answer to this question: **accept**, **revise**, or **reject**. The definition of each decision is summarized below.

Decision	Definition
Accept	Item/testlet is within acceptable limits. It may not be perfect, but it is acceptable.
Revise	Item/testlet violates one or more criteria. It has some potential merit and can be acceptable after revisions.
Reject	Item/testlet is ineligible for delivery to students. No revision could bring this item/testlet to within acceptable limits.

Remember that judgments about items are made separately from judgments about testlets because different criteria are used for items and testlets. It is possible to recommend revisions or rejections to items without automatically having to recommend revision or rejection to the testlet.

RECORDING COMMENTS

If a panelist recommends an item or testlet be accepted, no further information is needed. However, if a panelist recommends revision or rejection, they must provide an explanation.

- If a panelist recommends **revision**, the comment needs to **identify the criterion/criteria by number** AND **propose a solution**.
- If a panelist recommends **rejection**, the comment just needs to **identify the criterion/criteria by number**.

It is helpful for panelist comments to be constructive and concise but detailed enough that I-SMART staff will be able to understand the solution. For example, "*too wordy*" is not enough information. "*Text in the question needs simplification. Replace 'supermarket' with 'store'*" is a helpful recommendation.

COMPLETING REVIEWS IN CONTENT BUILDER

We recommend that installing Firefox (52 or higher) ESR on the computer a panelist will use to complete external review. If a panelist does not have it, download and install from this page: https://www.mozilla.org/en-US/firefox/organizations/all/

1. Paste the Firefox link into the current web browser.
2. Select English.
3. Download Firefox ESR for the computer.
4. The download should begin automatically.
5. When the download is complete, open Firefox ESR to make sure it is functioning appropriately.

WHAT IS CONTENT BUILDER?

Content Builder is proprietary software used to build testlets. Testlets built in Content Builder are then placed on forms and uploaded to the Kite® system for delivery to students. Content Builder also houses Review Management, which is the feature used to complete the external review of testlets.

ACCESSING CONTENT BUILDER

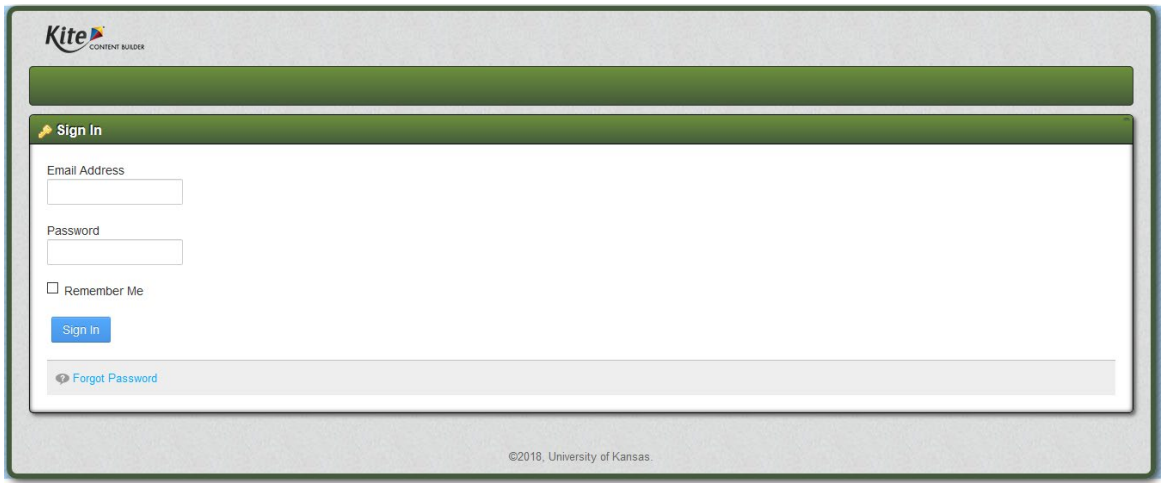
If a panelist is a new user, they should have received an email from Kite Service Desk (kite_support@ku.edu) regarding login credentials for Content Builder. Panelists will be using the Content Builder platform when evaluating items and testlets.

- Panelists will be provided a link to login, username, and temporary password. For security purposes, they will be prompted to update the password upon first login.
- **IMPORTANT NOTE:** If panelists do not log in to activate their account within 72 hours of receiving this email, their account will be deactivated, and they will need to contact kite_support@ku.edu for a second activation email. If panelists have not received the email from kite_support@ku.edu, they will need to check their spam or junk mail folder.

If you have been an external reviewer for ATLAS previously and used the same email address in the recruitment survey in Qualtrics, you will not receive a notification from kite_support@ku.edu, and your login information should remain the same.

To access **Content Builder,** follow these steps:

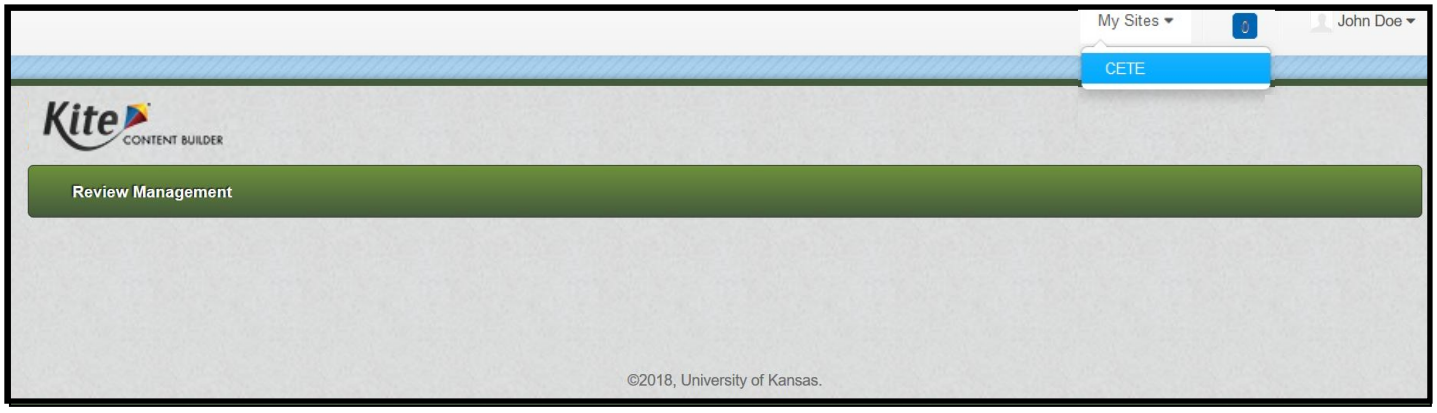
1. Access Content Builder by visiting https://cb.kiteaai.org/
2. The login screen will be the first one to come up.



3. Enter the email address associated with the Content Builder account in the **Email Address** box**.**
4. Enter the password associated with the Content Builder account in the **Password** box**.**
5. To have information automatically associated with the Content Builder login, select the box labeled **Remember Me.** This will eliminate the process of logging in each time in the future.
6. Click **Sign In**.

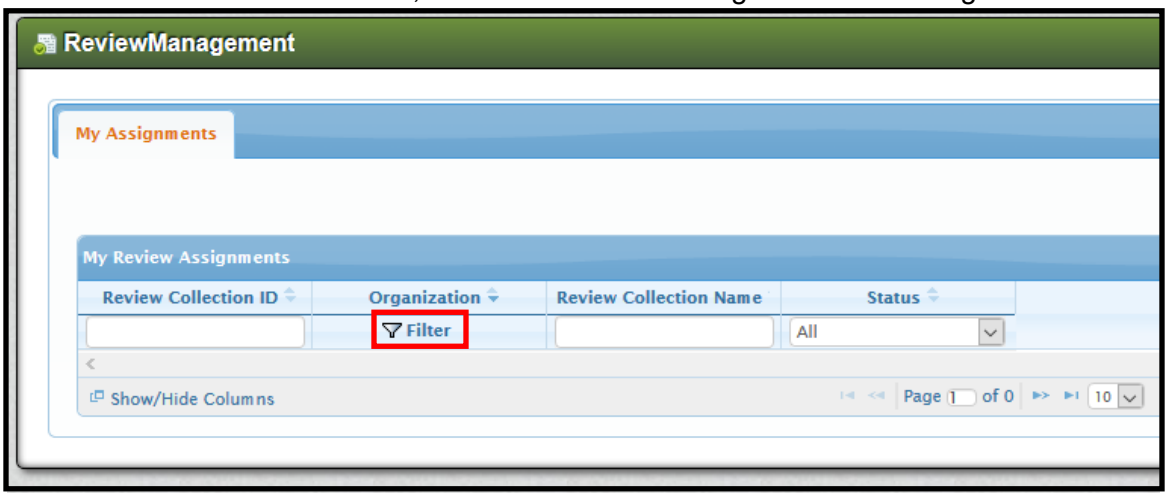
ACCESSING ASSIGNMENTS

After logging in, panelists will be redirected to Content Builder's main page.

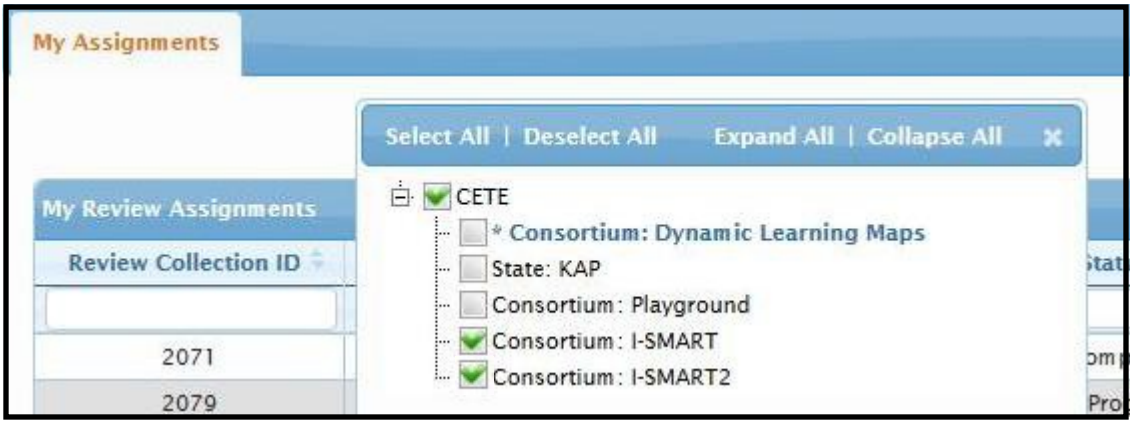


1. Click **Review Management.** If Review Management does not show up in the green bar to the left, click the **My Sites** dropdown at the top right of the screen and select **CETE** from the menu. Then click **Review Management**.
2. Panelists will be redirected to the **Welcome Page** for external reviewers.

3. The assignment list will display one or more collections. Each collection contains up to 10 testlets. Click the **first row** in the **My Review Assignments** tab to start reviewing a collection.
4. If there are no collections listed, click Filter under the Organization heading.



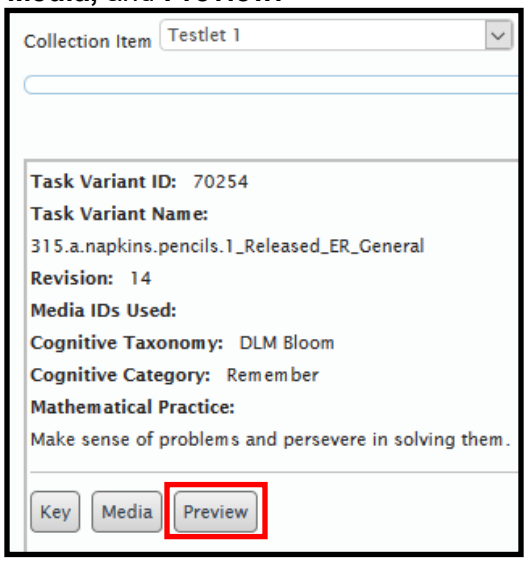
5. A box will pop-up. Select CETE and another checklist will populate. Select Consortium : I-SMART and Consortium : I-SMART2. Click Apply. This will populate the assignment list of collections.



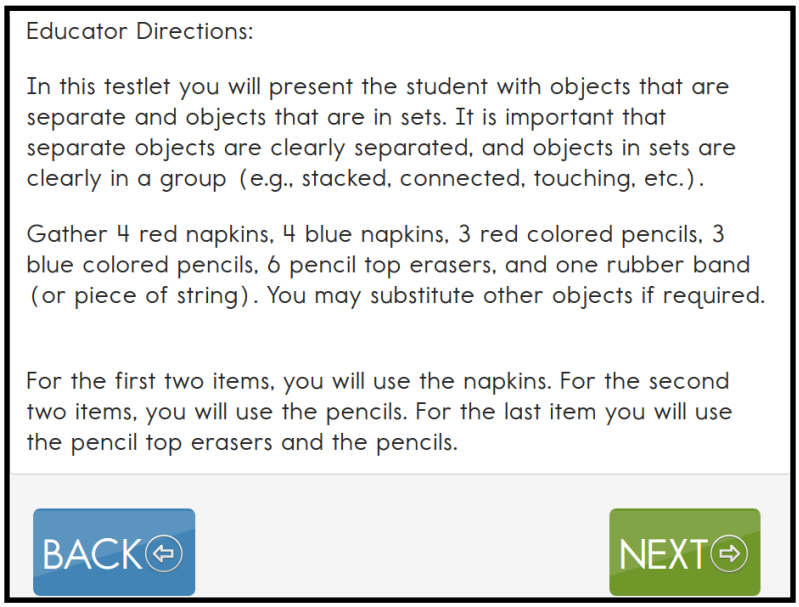
REVIEW ITEMS

Panelists will be redirected to the **Review** tab, where they will see the first item within the first testlet for review.

Between the metadata table and the limited visual of the item are three buttons: **Key, Media,** and **Preview.**



All panels (content, accessibility, and bias and sensitivity) will preview the testlet prior to providing a rating for the items and the testlet. Click **Preview** to view the testlet as it approximately[1] appears to students or educators**.**



Panelists may need to scroll to view the entire screen. The **Back** and **Next** buttons allow for navigation forward and backward in the testlet to evaluate all items before returning to evaluate the items individually. The **Enter** key on the keyboard can be used to advance through the testlet. When panelists are unable to advance the testlet any further by clicking the **Next** button, they have reached the end of the testlet. When ready

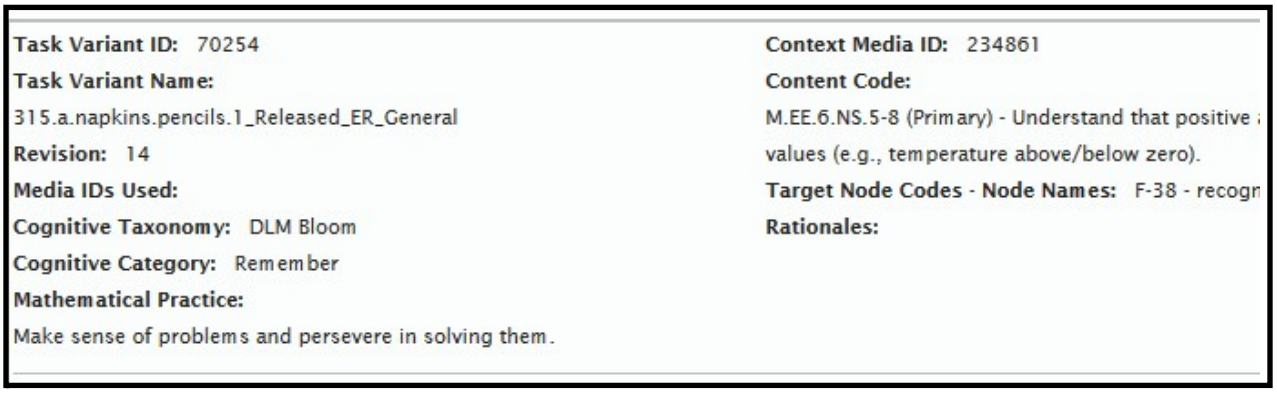
[1] Technology-enhanced items, such as those used for matching or sorting, will not be fully functional in the preview. The contents should still be visible.

to return and evaluate an item, close the **Preview** window by clicking the x at the top right of the preview screen.

At the top of the screen, information about the item will be displayed. Each panel type requires different information in order for the panelist to make decisions about the specific panel criteria. The table below indicates which pieces of information each panel type will reference. Refer to the next screenshot for how this information is displayed in Content Builder.

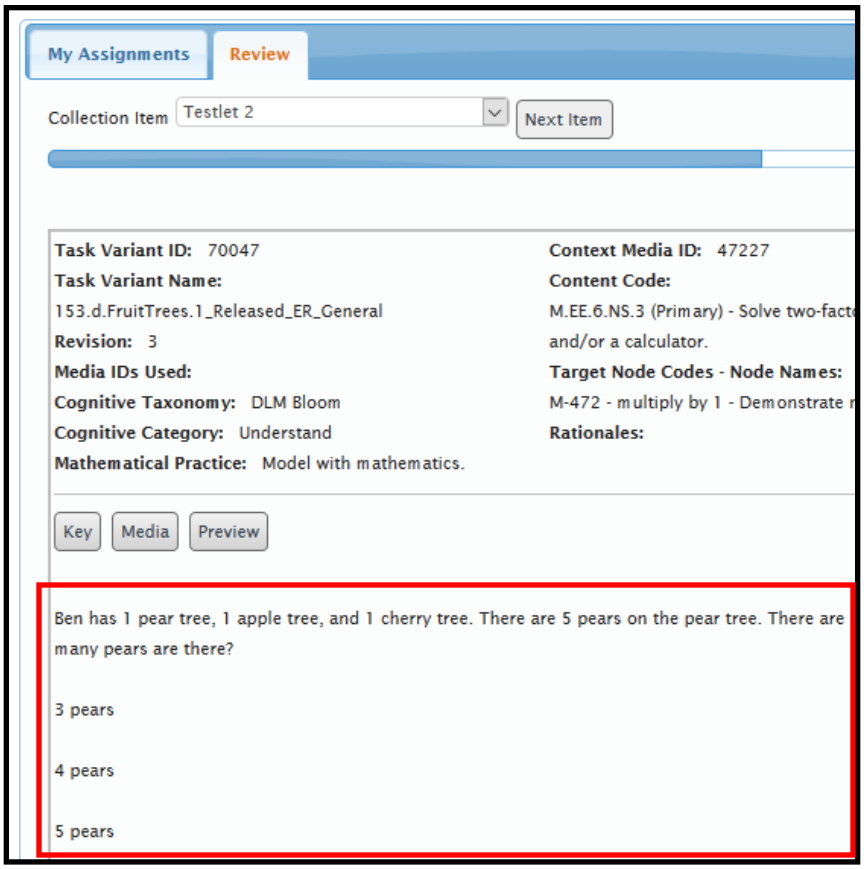
Panel Type	Useful Information
Content	Preview, Target Node Codes and Node Names, Cognitive Process Domain (Cognitive Category), Key
Accessibility	Preview, Target Node Codes and Node Names
Bias & Sensitivity	Preview, Target Node Codes and Node Names

The screenshot below shows how the metadata table is displayed in Content Builder.



The labels in Content Builder are not always the same as the I-SMART system's terminology. **Target Node** lists the specific node that the item assesses. **Cognitive Category** indicates the Cognitive Process Dimension. This screen contains the **Context Media ID** and **Media IDs Used,** but they are unrelated to the external review process.

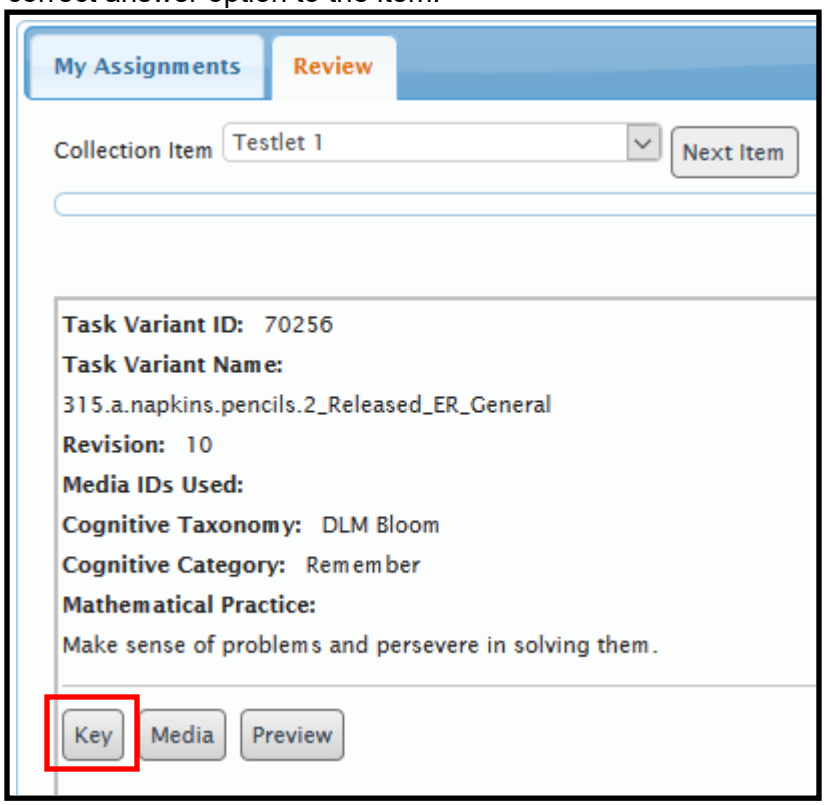
Below the metadata table, panelists will see a limited visual of the item.



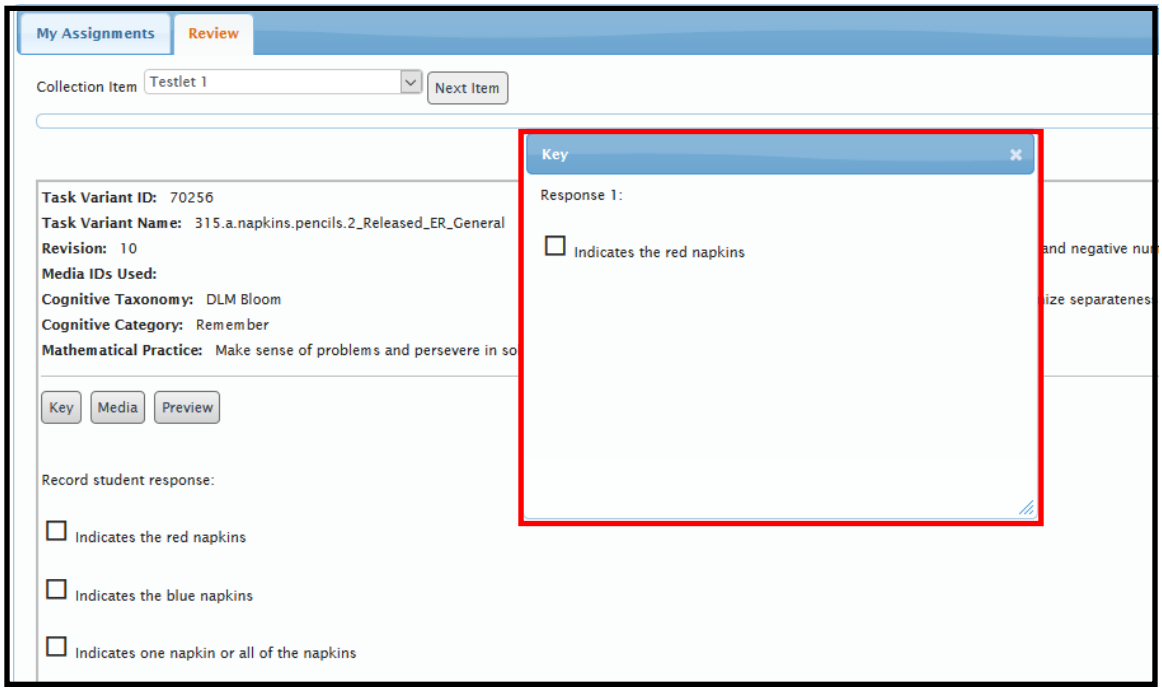
A Special Note About Teacher-Administered Items

For some teacher-administered items, the Educator Directions or stem appears on a page before the answer options. To external reviewers, it may appear as though the item has answer options and no stem. It also may look like several items in a testlet are identical. While the answer options may be identical, the Educator Directions or stem will be different across tasks. Use the Preview option to review the testlet and make sure panelists see the difference in the items being asked within the testlet.

Reviewers on **CONTENT** panels will also need to click the **Key** button to look at the correct answer option to the item.

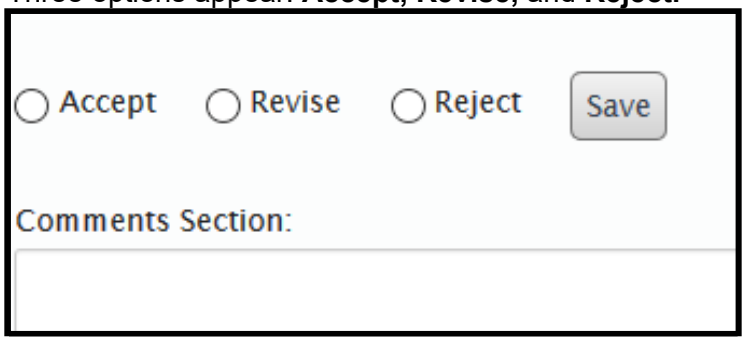


A pop-up box will appear and display the answer key.



After panelists have previewed the testlet and reviewed the information needed for their type of panel by referencing the I-SMART External Review Criteria, make a decision about the item using the assigned panel's criteria.

Three options appear: **Accept, Revise,** and **Reject.**



If the item is acceptable as-is, select **Accept** and then click **Save**. Leaving comments when rating an item or testlet as **Accept** is not required. If the item needs revision or should be rejected, select either **Revise** or **Reject** accordingly**.** Comments explaining the decision are required when either of these options are selected.

Example comments for a *Revise* rating for an *item*:

Comments should be specific and informative. Use the I-SMART External Review Criteria to locate the **Item** criteria for the assigned panel type. If an item does not meet a criterion, provide the number of the criterion as well as a short summary statement of the issue or concern. In addition, provide a specific suggestion on how to edit the item so it meets the listed criterion.

- Example comment for **Content** review: *4. There are two correct answers here. Bears eat fish, but they also eat "people" food from a picnic (i.e., fruit). Suggest changing "fruit" to "car."*
- Example comment for **Accessibility** review: *2. Graphics are very busy. Too many checkers shown for students to count. Suggest changing to include fewer checkers in the graphic or groups of checkers.*
- Example comment for **Bias & Sensitivity** review: *1. Item requires students to know details about how many points are scored for a touchdown to answer the question. Suggest changing item so information about the number of points is provided.*

Example comments for a *Reject* rating for an *item*:
Comments should be specific and informative. Use the I-SMART External Review Criteria to locate the **Item** criteria for the assigned review. If an item does not meet a criterion, provide the number of the criterion as well as a short summary statement of the issue or concern. A suggestion on how to edit the item do not need to be provided.

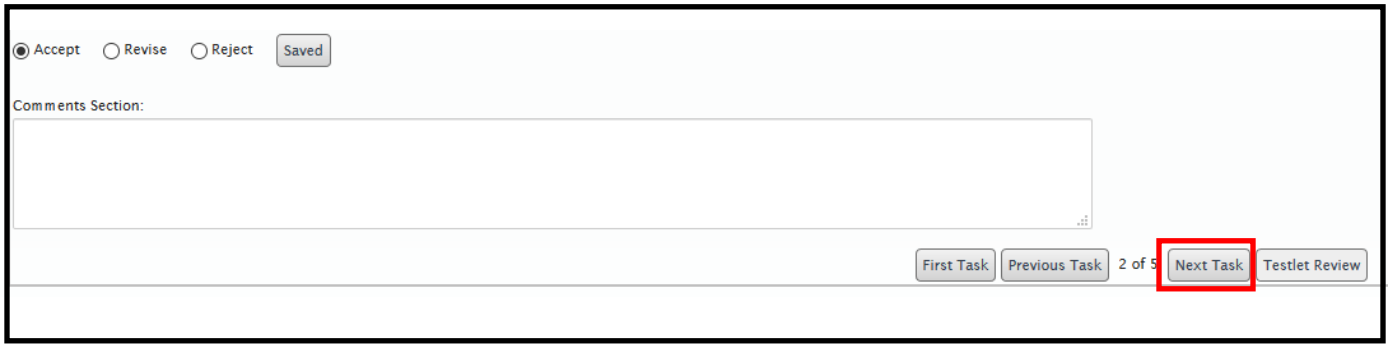
- Example comment for **Content** review: *4. This item has no correct answer options and the question (stem) is illogical.*
- Example comment for **Accessibility** review: *1. The language used in this item and answer options is confusing and more appropriate for a high school item rather than this third grade item.*

- Example comment for **Bias & Sensitivity** review: *5. The context of the testlet is offensive and may cause a severe emotional response in some students.*

After providing comments, click **Save.**

Each testlet is unique and contains multiple tasks. After panelists **Save** their rating for an **item**, they will notice four navigation buttons at the bottom of the screen: **First Task, Previous Task, Next Task,** and **Testlet Review.**

Click **Next Task** to view the next task.



Progress through the remaining tasks in the testlet. After panelists have reviewed all the tasks and provided ratings, they will provide an overall rating for the testlet.

REVIEW TESTLET

After panelists complete reviews for all items, click **Testlet Review**, and the review page will appear. If the **Testlet Review** button is not active, make sure panelists have provided a rating on each item and have clicked **Save**.

The testlet review screen does not contain any new information. Panelists may preview the entire testlet again if they wish. Otherwise, they need to record their rating and comments (if Revise or Reject rating) for the testlet as a whole and click **Save.**

Example comments for a *Revise* rating for the *Testlet*:
Comments should be specific and informative. Use the I-SMART External Review Criteria and locate the **Testlet** criteria for the assigned review panel. If a testlet does not meet a criterion, provide the number of the criterion as well as a short summary statement of the issue or concern. In addition, provide a suggestion on how to edit the testlet so it meets the listed criterion.

- Example comment for **Content** review: *2. Embedded items precede the slide with the relevant information. Suggest that they be placed after the relevant slide instead.*
- Example comment for **Accessibility** review: *1. Graphics are hard to distinguish as grouped for all items. Suggest putting a box around each image to help students differentiate them.*
- Example comment for **Bias & Sensitivity** review: *1. All items in testlet require student to have prior knowledge of baking measurements. Suggest that conversions that are not testing the content be provided to the student in the item.*

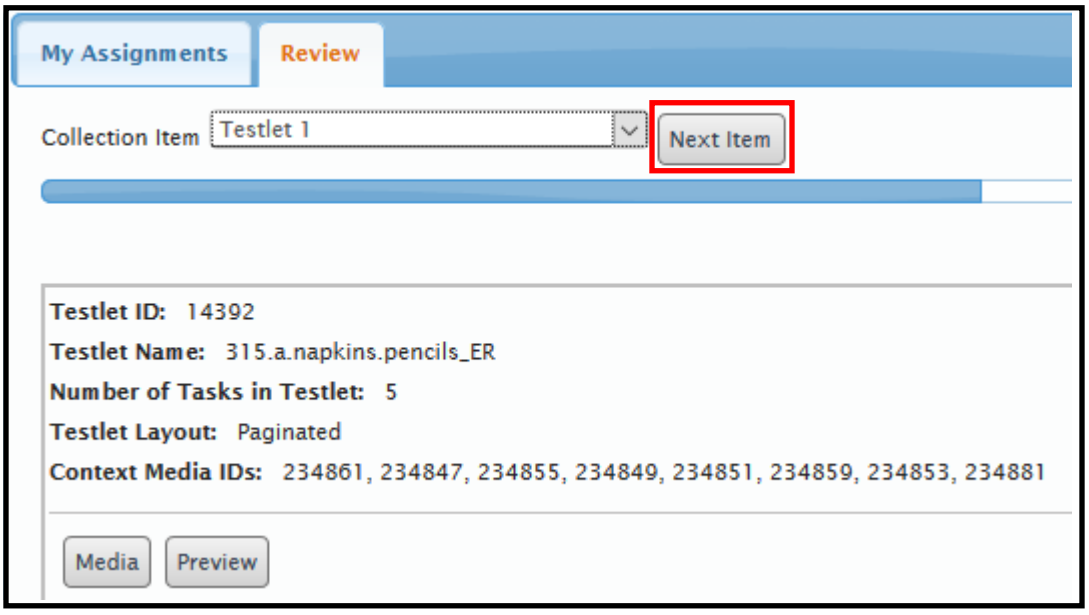
Example comments for a *Reject* rating for the *Testlet*:
Comments should be specific and informative. Use the External Review Criteria sheet and locate the **Testlet** criteria for the assigned review. If an item does not meet a criterion, provide the number of the criterion as well as a short summary statement of the issue or concern. Panelists do not need to provide a suggestion on how to edit the testlet.

- Example comment for **Content** review: *1. The testlet is not grade-level appropriate. Students are not expected to understand simple events until sixth grade.*
- Example comment for **Accessibility** review: *1. The testlet is complex and introduces barriers for students with limited working memory.*
- Example comment for **Bias & Sensitivity** review: *1. The testlet has a few controversial images. Because the images are tied to the context, the testlet should be rejected.*

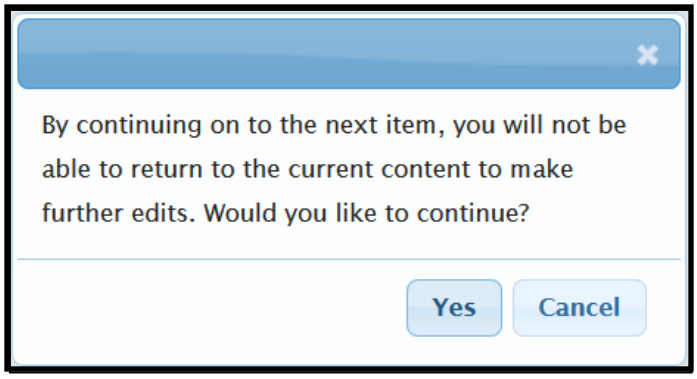
After providing comments, click **Save.**

GO TO THE NEXT TESTLET

After panelists complete the **Testlet Review,** click the **Next Item** button just above the progress bar next to the dropdown menu. This will move panelists to the next testlet within the collection.

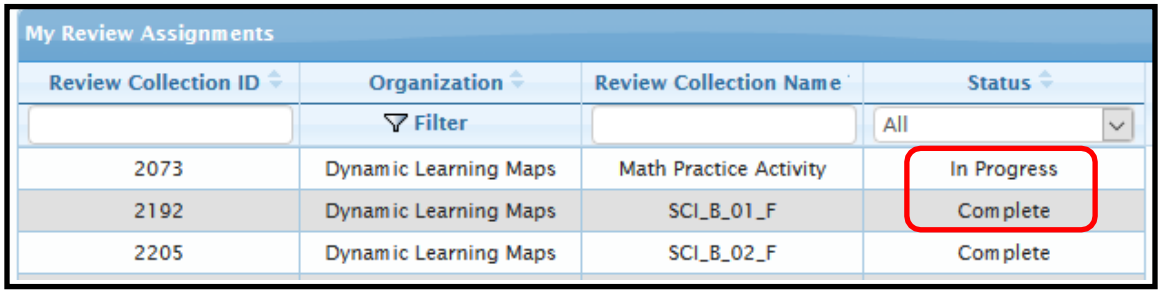


Note that within an assignment, panelists will review testlets one by one. In the Content Builder system, a Collection Item is a testlet. Within a testlet, panelists will review the items first and then the testlet. Each item has to be reviewed in order. The system will let panelists revisit previously rated items within a testlet, but will not allow them to skip ahead or revisit items and testlets already completed. The system will ask panelists to confirm that they are finished with that collection item (testlet).



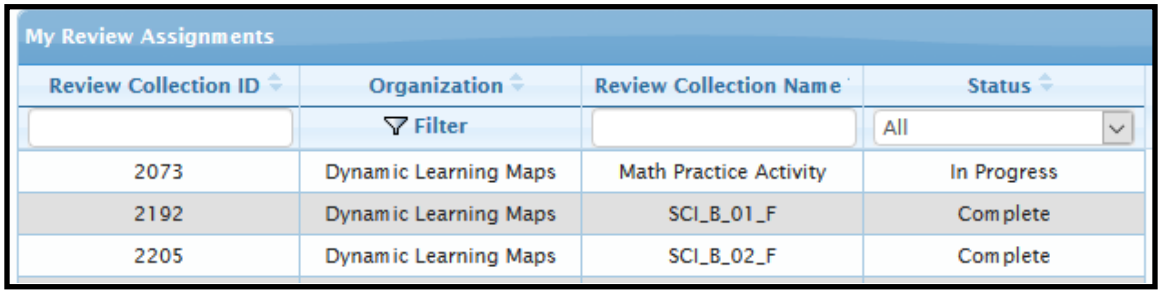
COMPLETE THE ASSIGNMENT

An assignment is complete when all collection items (testlets) have been reviewed and rated. To view the current progress, click the **My Assignments** tab. When one or more collection items have been reviewed, but not all, the **Status** tab in the assignments table will read **In Progress,** indicating the need for additional review work**.** When all testlets in the collection have been rated, the **Status** column in the assignments table will read **Complete.**



SAVING AND RETURNING LATER

Panelists may log out of Content Builder at any time without losing the work they have saved. When they log back in, they will locate the collection that is still marked **In Progress.** Click the **first row** in the **My Review Assignments** tab to start reviewing a collection.



HINT: Panelists may see an inactivity timer pop-up when working in Content Builder. This is an intended security feature. Click **Extend** in the top banner to continue working in Content Builder.

GLOSSARY

This glossary compiles definitions relevant to the external review process for the I-SMART project.

Assignment	A group of testlets assigned to panelists for review.
Collection	A group of testlets
Collection Item	Testlet
Cognitive Process Dimension	Assigned to each item based on the cognitive processes used to complete the item. Gives an overall measure of an item's complexity.
Embedded items	Items presented within a reading of a text. Embedded items focus on measuring cognitive processes as they occur during reading. In order to reduce cognitive load for students, these items are presented during a reading of a text near the relevant informational content from the text, rather than at the conclusion of the text.
Engagement activity	An activity that precedes a testlet that describes a scenario, taps prior knowledge or experience, and/or introduces the concept being addressed.
Stem	Question or direction given to the student
Testlet	Testlets are groups of items, bundled with instructions and engagement activities.