Evaluating Innovative Science Assessments: Evidence from I-SMART Cognitive Labs

Gail Tiemann
Meagan Karvonen

The authors wish to thank current and former ATLAS staff who contributed to this project, including Russell Swinburne Romine, Michelle Shipman, Lindsay Ruhter, Lori Andersen, Emily Bertels Kaufman, Jessie Lancaster, Mari Langas, Christine Mars, Jonathan Schuster, Sheila Wells-Moreaux, and Courtney Woodke.

We also thank the Maryland State Department of Education, the Missouri Department of Elementary and Secondary Education, and the Oklahoma State Department of Education for their assistance with recruiting study participants. Additionally, we thank the classroom teachers and students who volunteered to participate.

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
Marci Torchon, 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
Elizabeth Celentano, 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
Christie Stephenson, Oklahoma State Department of Education
Todd Loftin, Oklahoma State Department of Education

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 Partners

Cast, Inc
BYC Consulting

Preferred Citation: Tiemann, G., & Karvonen, M. (2019). *Evaluating innovative assessments: Evidence from I-SMART cognitive labs*. Accessible Teaching, Learning, and Assessment Systems (ATLAS), University of Kansas.

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

Executive Summary

Cognitive labs for the Innovations in Science Map, Assessment, and Report Technologies (I-SMART) project were conducted with 25 students and their teachers to gather evidence of student response process on short, inquiry-focused science tests. Testlets covered three alternate content standards at elementary, middle, and high school grade bands and included new Universal Design for Learning (UDL)-based features designed to support student engagement, interest, and overall accessibility. Included populations were students with significant cognitive disabilities eligible to take alternate assessments and students with and without disabilities performing significantly below grade level but not eligible for alternate assessments. Sources of data included student think aloud and retrospective verbalizations, lab elapsed time, student and teacher interviews, and teacher surveys. Evidence from results suggested that, in general, the UDL-based features did engage students appropriately without adding unresolvable barriers and that testing time was acceptable. Students and teachers liked the testlet content, especially use of media, and offered practical suggestions for improvement. Comments and student test-taking experiences generally supported intended interpretations of testlet content.

Introduction

The I-SMART project is a federally funded enhanced assessment grant project with an overall goal to deliver innovative assessments in science and score reports that improve the utility of information about student performance for students with significant cognitive disabilities and students who are performing significantly below grade level.

A key goal of I-SMART is to design, develop, and evaluate assessments that incorporate science disciplinary content and science and engineering practices in highly engaging, universally designed, technology-delivered formats. An objective of this goal is to develop prototype item types and to test the prototypes with students via cognitive labs using techniques adapted from previous research on the Dynamic Learning Maps (DLM) Alternate Assessment System (Karvonen et al., 2020). To this end, cognitive lab sessions were conducted in partner states in 2018. The primary purpose of the cognitive labs was to gather response process evidence to evaluate whether students interacted as intended with the testlets or whether the item format and response process demands introduced construct-irrelevant variance (American Educational Research Association et al., 2014). Findings from the cognitive labs were used to inform revisions to item design prior to further development of testlets for the I-SMART pilot. The 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?

Prototype Testlets

As part of the project, prototype testlets were developed at three grade bands: elementary, middle school, and high school. Prototypes were written to three different levels of complexity, known as linkage levels. Linkage levels represent small collections of learning map model nodes, with nodes representing individual knowledge, skills, and understandings. Testlet complexity increases as one moves from Initial, to Precursor, to Target linkage levels, with the Target level most closely aligning with the Essential Element. Due to time and resource constraints, we did not develop or test all the combinations of grade band and linkage levels for this study.

The prototype testlets contained three to four scored items per node, with four nodes per testlet. The majority of the items were standard multiple choice; a few testlets contained items that required students to drag and drop an answer choice into a category. All testlets were structured around a rich science narrative that followed an inquiry process and a specific science phenomenon. Items asked students to make connections between the science and engineering practice and the disciplinary core idea.

The prototype testlets also incorporated options based on UDL, including student-determined choice of context, phenomenon-based engagement activities, wonder questions, and opportunities for self-evaluation.

A full description of the I-SMART testlet features and the processes used to development them can be found in the report, *Designing, Developing, and Evaluating Innovative Science Assessments: Evidence from the I-SMART Project.*

Elementary

The testlet prototypes for elementary (grades 3–5) measured Life Science Essential Element LS2-1 at two linkage levels, Initial and Target. The Essential Element was:

> Create a model that shows the movement of matter (e.g., plant growth, eating, composting) through living things.

Elementary testlets at the Initial and Target linkage levels both measured four nodes, with three to four scored items per node.

Middle School

The testlet prototypes at middle school (grades 6–8) measured Life Science Essential Element LS2-2 at two linkage levels, Precursor and Target. At middle school, this Essential Element was:

> Use models of food chains/webs to identify producers and consumers in aquatic and terrestrial ecosystems.

Middle school testlets at the Precursor and Target linkage levels both measured four nodes, with three to four scored items per node.

High School

The testlet prototypes at high school (grades 9–12) measured Life Science Essential Element LS2-2 at two linkage levels, Initial and Target. At high school, this Essential Element was:

> Use a graphical representation to explain the dependence of an animal population on other organisms for food and their environment for shelter.

High school testlets at the Initial and Target linkage levels both measured four nodes, with three to four scored items per node.

Universal Design for Learning Options—All Grades and Levels

In order to support student engagement and interest, prototype testlets were developed with new features based on UDL. To try out these features with students, testlets were created based on two testlet templates. The first template was offered to students at the Initial and Precursor linkage levels. In this template, at the beginning of the testlet, students were presented with an unscored item that probed their choice of testlet topics. Once students made a choice, the narrative for the remainder of the testlet was related to that choice. This template had two variations. In the first variation, students were presented with a choice item that was rooted in the construct itself. For example, the student could choose that their testlet would be about polar bears or seals. In the second variation, the choice options were not related to the construct and instead related to characters such as Tim or Sally. By testing both variations during lab sessions, we explored student preferences for one variant or the other, or neither.

The second and final testlet template was offered to students at the Target linkage level and consisted of a slightly more elaborated science narrative. The second template had no variants, and no topic choices were offered. However, this template offered students the opportunity to pause and think more deeply about the topic as a way to engage natural interest and curiosity. The more elaborated science narrative aligned with the slightly more complex skills required at the Target linkage level.

The following lists describe the UDL options that were included in the testlet prototype templates. The options are described further in the results section.

Testlet Template 1 (Initial and Precursor, except where noted)

- Phenomenon-Based Science Narrative
- Choice Option (construct-relevant or construct-neutral)
- I Wonder (Precursor linkage level only)
- Use of Video (Precursor linkage level only)
- Self-Assessment

Testlet Template 2 (Target linkage level only)

- Elaborated, Phenomenon-Based Science Narrative
- I Wonder
- Think About It
- Use of Video
- Self-Assessment

In sum, nine prototype testlets were developed for cognitive labs. Table 1 lists the testlets by linkage level and grade band. Testlets based on the construct-related choice template are labeled A. Testlets based the character-related choice template are labeled B. Testlets based on the elaborated science narrative template are labeled C.

Table 1
Prototype Templates by Linkage Level and Grade Band

Linkage level	LS2-1 Elementary	LS2-2 Middle school	LS2-2 High school	Total
Initial	Testlet A1 Testlet B1		Testlet A3 Testlet B3	4
Precursor		Testlet A2 Testlet B2		2
Target	Testlet C1	Testlet C2	Testlet C3	3
Total	3	3	3	9

Sample

The I-SMART project serves three populations of students. The populations are described in Table 2.

Table 2
I-SMART Student Populations

Group name	Description
Group 1	Students with significant cognitive disabilities (SCD) who are eligible to take alternate assessments based on alternate achievement standards (AA-AAS)
Group 2	Students with disabilities who perform significantly below grade level but are not eligible take AA-AAS
Group 3	Students without disabilities who perform significantly below grade level

For Group 1 (students eligible for DLM assessments), the grade-level expectation is the Essential Element for the chronologically appropriate grade. For exploration with Groups 2 and 3 (struggling learners), we used the same learning map neighborhoods 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.

Cognitive Lab Eligibility

Students in Group 1 were identified by their enrollment in the DLM alternate science assessment in the academic year of the cognitive lab session. Students in Groups 2 and 3 were

identified at the local level using teacher judgment about students who met the following suggested inclusion criteria:

1. The student had a documented disability but was not eligible for DLM assessments (Group 2 only).
2. The student received instruction in the state's general education science standards for the student's chronologically age-appropriate grade level. More specifically, the student received instruction in the content reflected in the nodes that the prototype testlets were designed to measure. To assist with interpretation of this criterion, we provided teachers with a bulleted list of science concepts from the elementary Initial level to the high school Target level.
3. The student did not meet grade-level expectations during science instruction and local assessment, despite having effective instruction and appropriate accommodations. More specifically, the student showed evidence of struggle or nonmastery of the content reflected in the nodes the testlets were designed to measure. This evidence was based on teacher judgement and the checklist noted above.
4. Finally, if the student completed a large-scale science assessment in recent years, the student scored at the lowest achievement level.

Other Inclusion and Exclusion Criteria

Students from all groups could represent any grade level within the appropriate grade band as long as the students had received instruction in the testlet content previously. For example, if a state typically tested students in science at eighth grade, seventh-grade students could participate in cognitive labs if they had received instruction.

Due to resource limitations for lab sessions, at the local level, we requested that students eligible for Precursor- and Target-level testlets be able to interact with the online platform using a standard mouse on a laptop or desktop computer, or on an iPad. We also requested that they be able to use speech to communicate in English. For Initial-level testlets, we requested that students be able to consistently communicate an answer through any response mode when asked a question; communication could be speech-based or not speech-based. Finally, for all testlets, we requested that students who were blind and students who had hearing impairments not addressed with hearing aids or a personal amplification device not be recruited at the local level.

To maximize the heterogeneity of the sample, we requested that districts recruit a variety of students representing different genders, races, and ethnicities when possible.

Intended Sample Size

The student sample size that was planned per research condition (population by grade band by testlet) is summarized in Table 3. During recruitment, we over-recruited by one student per condition to guard against attrition between the time of parent consent and data collection. For students in Groups 2 and 3, we balanced recruitment between the two groups but did not have a specific quota, nor did we collect data that distinguished the groups (i.e., whether the student had a documented disability).

Students in Group 1 were assigned to all testlet types (A, B, C) as they reflected the grade-level expectations for students who take alternate assessments. Students who did not take alternate assessments (Groups 2 and 3) were assigned to the Target (C)-level testlets because that linkage level was potentially an instructional match for students struggling in science. The content of testlets A and B were very similar, allowing us to pool information across those administrations for all analyses except for examination of systematic differences between the choice types.

Table 3
Maximum Intended Number of Students Per Grade Band, Target Population, and Prototype Testlet

Group	LS2-1 Initial and Target elementary	LS2-2 Precursor and Target middle school	LS2-2 Initial and Target high school	Total
Group 1	Testlet A1–3 Testlet B1–3 Testlet C1–2	Testlet A2–3 Testlet B2–3 Testlet C2–2	Testlet A3–3 Testlet B3–3 Testlet C3–2	24
Groups 2 and 3	Testlet C1–4	Testlet C2–4	Testlet C3–4	12
Total	12	12	12	36

Recruitment

To recruit volunteer teachers and students, I-SMART staff worked with state partners to disseminate information about the cognitive lab sessions to districts within their state. State education agencies contacted districts directly or provided contact information to I-SMART for further follow-up. After initial contact, I-SMART staff shared information with districts about lab procedures, student eligibility criteria, and onsite needs. Informational letters for teachers and parents/guardians, as well as informed consent documents for both teachers and parents/guardians, were also shared. Once information was disseminated, lab sessions were either arranged by the district or by working directly with teachers who had indicated an interest in participating.

Recruitment was ongoing throughout the 2017–2018 and 2018–2019 academic years. After each cognitive lab session, I-SMART staff targeted recruitment toward remaining student sample needs in each grade band and linkage level.

Procedures

Facilitator Training

Each staff member involved in cognitive labs received training on session procedures, including general facilitation techniques, specific session protocols, data recording processes, and

student assent procedures. Staff who had prior experience with the target populations and with facilitation techniques conducted the sessions.

Testlet Assignment

Prior to the cognitive lab sessions, I-SMART staff worked with teachers to determine the testlet linkage level most appropriate for the student participants. In cases in which the teachers were unsure of the best-fitting testlet, staff reviewed the First Contact survey results and followed testlet assignment guidelines currently in use in the DLM science assessment program. Additionally, teachers provided information on any student accessibility needs. For example, if the student used an accessibility option during instruction or during typical assessment practice, the accessibility option was enabled for the lab session. However, due to resource limitations, text-to-speech tools were not available. When needed or preferred by students, either I-SMART staff or the student's teacher provided human read aloud support.

General Procedures

The lab sessions were held in quiet locations in the students' school buildings and testlets were administered to students individually. One staff member facilitated the session while another staff member recorded observations on a written form. For students in Group 1, teachers were present and assisted with administration where appropriate.

Data Sources

For students who were able, staff encouraged students to think aloud as they responded to the testlets. Prior to beginning, students practiced thinking aloud to orient to the activity. In some cases, it was less cognitively demanding for students to answer retrospective probes directly after answering an item than it was to think aloud while answering the item. Facilitators used professional judgement as to when to probe for further detail.

To capture facial expressions and verbal comments, staff videotaped and/or audiotaped students as they responded to the prototype testlets. Video was only collected with parent or guardian permission. For computer-based testlets, student interactions with the computer were recorded via screen-capture software. Additionally, staff completed observation forms that noted the student's item response, construct-irrelevant behaviors, quotes, levels of engagement, and notes for follow-up questions.

After each session, where possible, staff interviewed students about their experience with the testlets. Students were asked about their general likes and dislikes, thoughts about testlet length, and suggestions for improvement.

Staff also interviewed teachers. Teachers were asked if the cognitive lab session was typical of the student's experience during assessment or instruction. Where appropriate, teachers were asked if students fully understood the choice-based items. Additionally, teachers were asked if they had covered the testlet content during the year and if they had suggestions for improvement.

Finally, teachers were asked to complete a survey for each student that probed the extent to which students had received instruction related to the content of the assessment and the extent

to which students had mastered the content. Survey results supported interpretations of student responses relative to the performance expectations the testlets were designed to measure.

Results

I-SMART staff completed 25 cognitive lab sessions in three partner states and four different schools during 2018 and 2019. The number of sessions completed by grade level and population is described in Table 4. While the number of total sessions was lower than intended (25 completed of the intended 36) and Groups 2 and 3 did not participate at the elementary or high school levels, overall, the completed sessions were well-distributed across the grade bands.

Table 4
Completed Sessions

Group	Elementary	Middle school	High school	Total
Group 1	Initial - 6 Target - 0	Precursor - 2 Target - 2	Initial - 5 Target - 6	21
Groups 2 and 3	Target - 0	Target - 4	Target - 0	4
Total	6	8	11	25

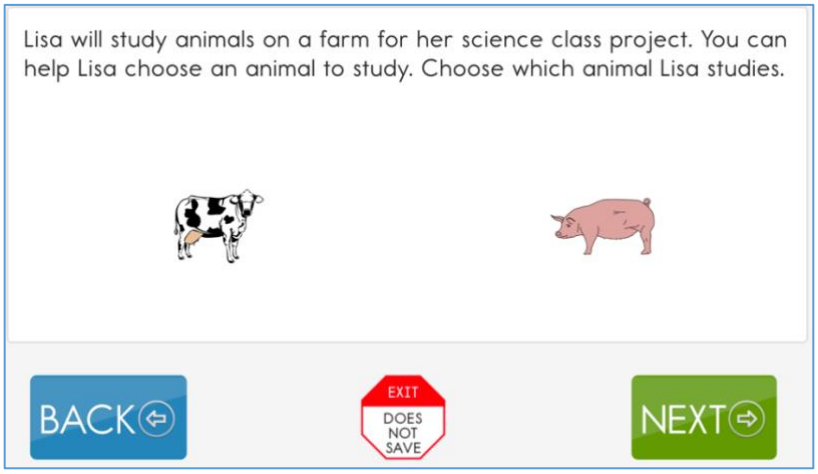
I-SMART staff reviewed sources of observation data including video and audio tapes, observation forms, and transcripts of interviews. Results along with descriptions of specific data analysis methods are described with each research question below.

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

Several innovations were included in the testlet prototypes, as described above. Student experiences with the features were observed and noted from the data.

Choice Options. In the first testlet template, toward the beginning of the testlet, students were presented with an unscored item that allowed them to select their choice of topic for the remainder of the testlet. Testlets from this template were presented at the elementary and high school Initial level and at the middle school Precursor level. An example of a Choice Option is shown in Figure 1.

Figure 1
Example of a Choice Option



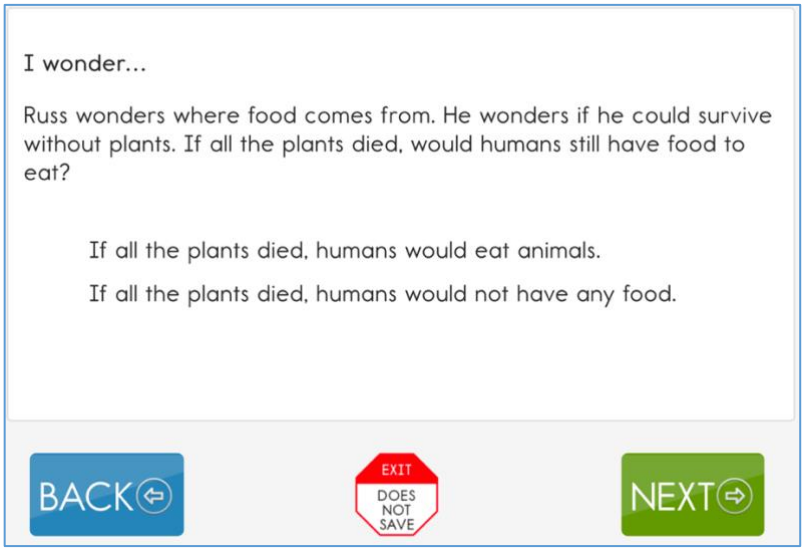
While only two students were tested at the Precursor linkage level, both easily made an intentional choice when presented with the choice item. The student who was presented with the construct-relevant choice selected the "pig" because he really liked pigs. The other student was presented with a choice of character and chose based on the color of the character's shirt, green. Additionally, the students' teacher did indicate that she frequently offered choices during typical instruction.

At the Initial linkage level, across the elementary and high school grade bands, eight of 11 students made intentional decisions of choice option. In all cases, teachers assisted with interpreting the student's choice and whether the student made an intentional decision or chose randomly or based on answer position (e.g., always left, always center). During follow-up interviews, some teachers of students at the Initial linkage level commented that choices were frequently offered during typical instruction; however, not all teachers were asked this question.

I Wonder. The unscored I Wonder question was presented at the beginning and revisited at the end of the Target-level testlets. Designed to support student engagement, this UDL feature was connected to a research-based misconception that could be resolved through the inquiry activities in the testlet.

Eight students at the middle school grade band encountered the question during lab sessions, two with the Precursor-level testlet and six with the Target-level testlet. Four of the middle school students were from Group 1, and four students were from Groups 2 or 3. Six high school students from Group 1 also experienced the I Wonder question, all at the Target level. The question offered only two answer options. An example of a wonder question is displayed in Figure 2.

Figure 2
Example of I Wonder



At the middle school level, all eight students engaged with the feature by reading the screen. At the first presentation of the item, two students failed to mark a response. One student commented during the interview that she did not know how to mark an answer on this screen. Additionally, one middle school teacher commented that the addition of on-screen instructions could help the students better know what to do.

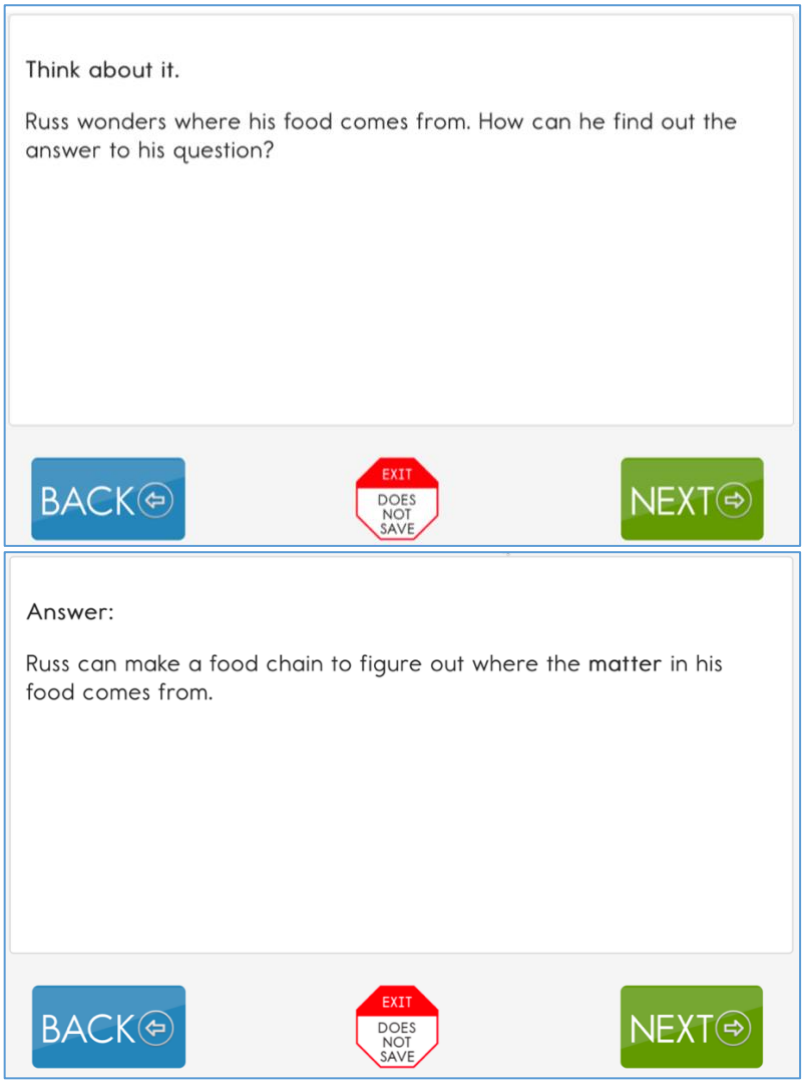
Four middle school students were able to explain the concept correctly. One correct representative response included the following statement: "Well, the chicken wouldn't have crops to eat and then if they're gone then humans wouldn't have anything to eat."

From the first instance to the second instance, two of the eight middle school students changed from the misconception response to the correct response. One student changed from a correct answer to a misconception. Three students retained their misconception-based answer. One of these three students marked the incorrect answer but explained the concept correctly during think aloud. Two students answered the I Wonder question correctly both times.

At the high school level, one student changed from the misconception response to the correct response. This student stated that they changed their answer to the second question based on the information in the testlet. Two students changed from a correct answer to a misconception. Zero students retained their misconception-based answer. Three students answered the I Wonder question correctly both times. Of these three, two stated that they knew the correct answer to the second question based on information in the testlet.

Think About It. Another UDL feature, called Think About It, was intended to provide support for executive function and was presented two times in Target-level testlets. This feature spanned two screens; a question was posed on the first screen (without answer options), and the answer was presented on the next screen. Six students encountered this feature at the middle school level. An example Think About It feature can be found in Figure 3.

Figure 3
Think About It Screens



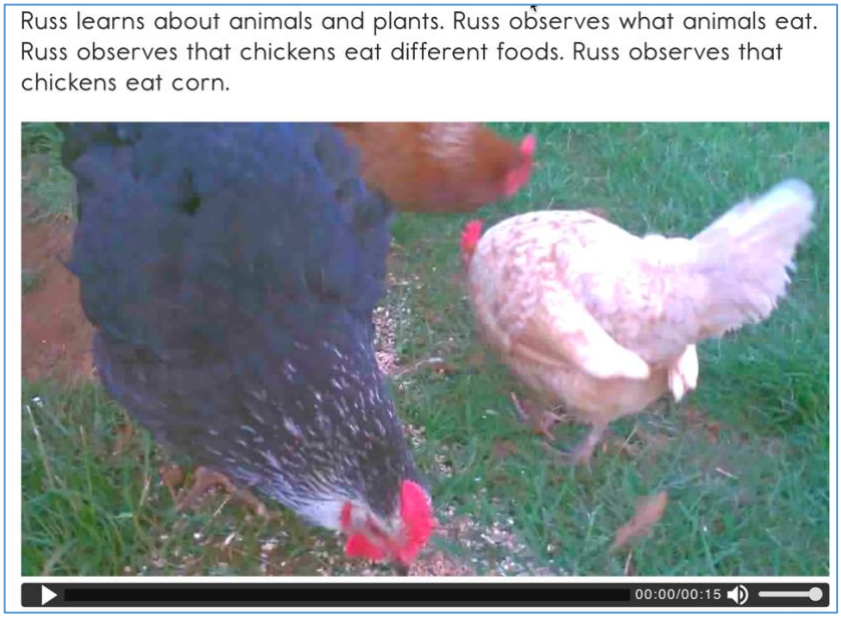
Six middle school students and six high school students experienced this feature during labs. At the middle school level, two students read the first Think About It screen and moved on. Two students offered an answer out loud. One offered an answer after the facilitator probed further. One student looked for an answer choice on the screen and was confused that the screen asked a question but did not offer a place to respond. On the second Think About It, five students paused and thought about the question and/or offered answers out loud and one student read the screen and moved on.

At the high school level, administrators prompted students to provide an answer aloud to the first Think About It prompt. Half of the six students offered suggestions related to the prompt. After the second Think About It prompt, half of the students offered unprompted answers, two students who provided suggestions after the first question and one who did not. One student

who provided a suggestion after the first question felt "stuck" after the second question and did not offer an explanation for their answer.

Use of Video. A video was used as a way to offer multiple means of expression and support in decoding information in the Precursor- and Target-level testlets. The video was related to the content of the testlet, but viewing or comprehending the video was not required to answer any of the items. Eight middle school and six high school students experienced videos. An example of the video is shown in Figure 4.

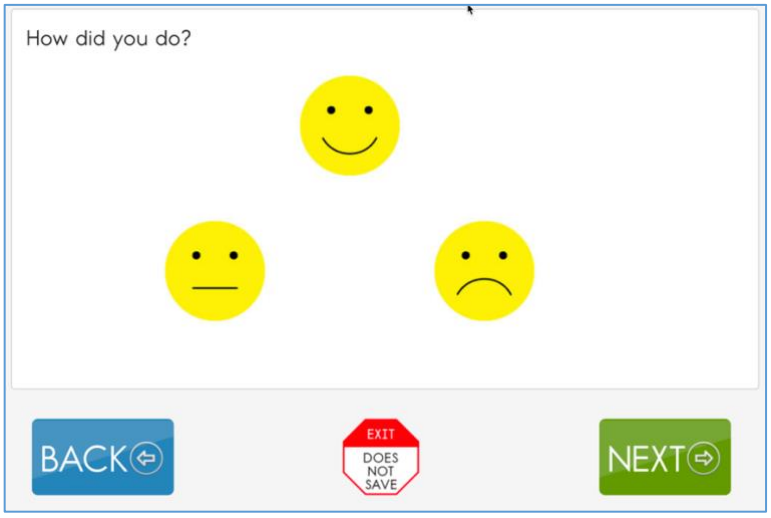
Figure 4
Use of Video Example



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. Two students were neutral, and others did not comment. One student with a positive opinion of the videos stated that it was her favorite part of the testlet. One of the students from Group 3 with a more neutral opinion stated, "I don't know. It's all in like a 15 second video. All I can see is this chicken is eating corn. I guess it was somewhat interesting to know how they eat."

Self-Assessment. An unscored self-assessment item was presented at the very end of the Precursor- and Target-level testlets as a way to support self-regulation. The item asked the student how they did and offered emoticon-style images as answer options in a familiar layout. The self-assessment item is displayed in Figure 5.

Figure 5
Self-Assessment Item



Twelve of the 14 students selected the smiley face answer option; two students chose neutral faces. One student answered with the smiley face but indicated through the verbal response that he had not done well. Another student performed well on the assessment but selected the neutral face.

Item Navigation. For the computer-based testlets (Precursor and Target level), selected-response and drag-and-drop items were used. 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. In terms of presentation order, this selected-response item was displayed four items later than the drag-and-drop item.

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

Because the design of the prototype testlet required more items than had been delivered previously though DLM science testlets, the time needed for each student to complete the cognitive lab was recorded during observations or after sessions based on media recordings. Each testlet measured four nodes, with three to four items per node. When coupled with unscored items, the number of total items for each testlet ranged from 14 to 17. The elapsed time information is exploratory only because the lab session was not an authentic assessment experience; unfamiliar observers, recording equipment, interview activities, and teacher input were present during each session. Some students at the Target level offered extended verbalizations that added to their elapsed time. Some students who took the choice-based testlet had longer response times due to sensory processing needs. Ranges of elapsed time are listed by testlet template and population in Table 5.

Table 5
Elapsed Time Range by Testlet Template and Population

Testlet template	Group	*N*	Number of items	Time range (minutes)
Choice-based	1	13	14–17	11:47–25:00
Extended narrative – Target	1	2	16	17:41–18:20
Extended narrative – Target	2/3	4	16	12:21–29:28

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

Precursor- and Target-Level Testlets. Prior to the cognitive labs, staff examined the item specifications for each Essential Element and noted an intended response process for each item based on the performance expectation it was designed to measure. Common misconceptions were also located on the specifications document. For example, one node at the Target linkage level for middle school Life Science Essential Element 2-2 requires students to use a model to describe the feeding relationship between two living things. We would expect a response process that shows students understand that an arrow in a food chain model points in the direction that matter moves between two living things, or in other words, from the food to the thing that eats the food. A common misconception is that the arrow represents which animal eats the other animal instead of how matter moves from one animal to another.

For students who were able to think aloud during, or provide retrospective comments after, answering an item, analysis of data for RQ 3 involved comparing student comments to the intended response process for the item and also to the list of misconceptions noted in the specifications. An I-SMART science test development staff member reviewed each answer and made a judgment based on the whether the comment was construct relevant (matched the intended response or matched a misconception), construct irrelevant (represented a guess or used a process of elimination), or was unknown. Judgments were reviewed by a senior staff member, and where judgments did not match, the team discussed the item until agreement was reached. Staff concluded that if students were answering based on the intended response process or a misconception, there was evidence that students were representing the construct in the same manner as the testlet designers, even if the students were not yet mature enough in their understanding to answer the item correctly.

At the middle school Target level, six students provided think aloud or retrospective comments. One of two students at the middle school Precursor level was able to verbalize thought processes. At the high school level with students from Group 1, only retrospective probes were used. Those students were less successful overall with verbalizing their thinking, and thus, more unknown ratings were used. Table 6 below lists the number of construct-relevant item responses by student population and testlet template.

Table 6
Construct-Relevant Responses by Population and Testlet Template

Testlet template	Group	*N*	Number of construct-relevant item responses by each student	Number of scored items
Choice-based – Precursor	1	1	8	14
Extended Narrative – Target	1	2	10, 11	14
Extended Narrative – Target	2/3	4	5, 10, 10, 14	14

Initial-Level Testlets. Eleven students completed Initial-level testlets, six at the elementary and five at high school. Determining construct relevance of responses was more difficult at the Initial level; students were not able to think out loud or respond to retrospective probes due to the extra cognitive load of these activities. I-SMART staff noted construct-irrelevant testing behaviors such as random picking of answer options, repeatedly picking answer options in the same position, or not looking at all the answer options. Additionally, session facilitators engaged students' teachers in helping with interpretation of student responses, especially relative to typical behavior during assessment or instruction.

To make a judgment about whether the students' responses represented the performance expectations, staff reviewed the session video tapes, observation forms, teacher interviews, and survey data to make a judgment for each student. Confirming and nonconfirming evidence for each case was also explored.

Of the six students in the elementary grade band, five did not have overall responses that represented the targeted constructs. Of these, one was unable to finish, three chose options primarily based on answer-choice position, and one had inconsistent responses using a communication device. For the student who showed the most evidence of construct-relevant answers (about half of the items), the teacher reported that she had previously instructed the student on the assessed science content.

Of the five high school students, two did not have overall responses that represented the targeted constructs. Both chose answers primarily based on position of the options, though the preferred answer position was not consistent between the two students. The other students made clear answer choices, selected answer options in all positions, and did not exhibit other significant construct-irrelevant behavior. These students also had more correct answers toward the beginning of the testlet, where teachers reported having delivered instruction, and fewer toward the middle and end of the testlet, where teachers said content had not been covered.

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

After each session, I-SMART staff interviewed students (where possible) and teachers about their experiences with the testlets using a semi-structured interview protocol. Interviews were transcribed and coded for themes.

Length and Difficulty. At the middle school level, three of nine students thought the test was too long. Of these three students, two were from Group 1 and one was from Group 2. Three of the six middle school students who competed Target-level testlets felt that the content was too easy. Of these three students, one was from Group 1 and two were Groups 2 and 3. Two of the middle school students, one from Group 1 and one from Group 3, stated that the questions seemed to repeat themselves. At the high school Target level, one student felt the test was too long. One student said that the test was easy, and two students felt that the test included easy and difficult questions.

At the Initial linkage level, three teachers felt that the content was too advanced for the students. Two teachers mentioned that they had covered part but not all of the content during instruction, "so we worked a lot on different animals. We've not drifted to who eats what yet this year. Last year, we talked about it a little bit and it was extreme [*sic*] above where the kids were at." One teacher also expressed concerns about the accessibility of the food-chain topic for her students who do not eat food orally, saying, "…I'm talking about food with these kiddos and who eats what but their energy does not come from orally eating. So, it's no motivation to them… it's a hard concept."

Media. During interviews at the Precursor and Initial linkage levels, three students mentioned liking the video and the pictures. One student's overall suggestion for testlet improvement was, "more pictures." One teacher mentioned that the pictures should be more realistic and larger on the screen.

Teachers at the Initial linkage level also had thoughts about the media. Two teachers mentioned that the pictures were not familiar to students, and thus, the students had difficulty comprehending the answer options. One teacher suggested that the pictures be more realistic.

General Usability Feedback. One teacher generally liked the flow of the storyline across the various media screens. The teacher also mentioned that the "wonder question" was an unfamiliar format and students did not know how to answer. Two students and one teacher felt that the format of the wonder question led to difficulty in knowing how to mark an answer. The teacher of the high school students at the Target linkage level felt that multiple-select multiple-choice questions were confusing to students.

Other screen-based suggestions for improvement included avoiding the word "not" in questions, bolding key words in text, avoiding placement of story narrative above key graphs on the screen, and making the food-chain arrows larger.

Discussion

The purpose of the I-SMART cognitive labs was to explore the extent to which students interacted with the I-SMART testlets as intended and the extent to which item format and response-process demands may have introduced unintended construct-irrelevant variance. While many sessions were completed, the number of completed sessions was below the number that were planned. With a small number of states, schools, and classrooms represented, results should be considered exploratory and formative in nature.

New Testlet Features

Several new UDL components were added to I-SMART testlets, including features to help students engage their natural interest and curiosity during the assessment experience. The study explored how students interacted with the item types and innovative testlet features.

Choice Options. From the early session results, neither variant of the choice-based testlet template, construct-relevant or character-based, was clearly preferred by students. From this information, I-SMART staff decided to move forward only with the construct-relevant choice option for the I-SMART pilot assessment and for subsequent cognitive lab sessions.

Students at the Precursor linkage level indicated their choice easily and with intention. At the Initial linkage level, where students have more communication challenges, eight of 11 students made intentional choices. Teacher input was key to interpreting whether students made an intentional choice. Future research on this feature could explore whether the offering choices led to increased student engagement during the testlet administration.

I Wonder. The I Wonder question was a novel feature that some middle school students found unfamiliar. We suspect that two middle school students did not mark an answer on this item because the answer choices were presented in a nontypical format; the answer choice text is longer than answer choices students would be familiar with on typical science assessments. Students at the high school level had less difficulty answering the I Wonder question; however, in most cases, the facilitators provided extra prompts that there was a question on the screen to answer.

While novel, about half of the middle school students were able to explain the underlying concept of the I Wonder question correctly. Additionally, some students stated that they changed their answer to the correct response based on the content of the testlet itself.

Difficulties with the feature may have been influenced by lack of student exposure to inquiry-based instruction. Potentially, the item could be enhanced by adding a sentence on the screen that directs the student to mark a response.

Think About It. The Think About It question consisted of text on the screen without a place to enter or mark an answer choice. Students were unsure what to do on this screen for the cognitive lab. Students did offer verbalizations about the question, especially when prompted by the lab administrator. It was unclear if students would have used their executive function to consider the question's content without prompting.

This feature was difficult to probe during lab sessions. New probes could potentially yield clearer evidence about the process students use when encountering the feature. Also, collecting and analyzing screen time data during a larger field test or pilot could help explore whether students read and considered the question as intended or clicked through.

Use of Video. Most students liked the short videos in the testlets. The videos were not required to answer the questions but were intended to engage the student in the overall narrative of the testlet.

The facilitators and students did discover that the videos loaded slowly and were overly large on the screen. Details about the video concerns were immediately communicated to the technology team who addressed the issues prior to the I-SMART pilot administration.

Self-Assessment. Offered at the very end of the testlet, the self-assessment item was intended to support student self-regulation. All but two students in the sample chose the smiley face. One student chose the smiley face even though they felt they had not done well on the testlet. With many observers present during the session, students may have chosen the happy face due to a halo effect. Further analysis of this item type from the pilot study will explore the distribution of the selected answer options and their relationships to testlet performance.

Item Navigation. Generally, students did not encounter issues with the drag-and-drop items. One related usability issue did surface; one item looked like a drag-and-drop item but was actually selected response. On further investigation, the item had been presented as selected response due to a technical difficulty with the item-authoring system. The test development team was informed and the technical issue was resolved prior to the I-SMART pilot administration.

Testlet Time

The specifications for the I-SMART testlet items coupled with the new features led to longer testlets than had been previously used in DLM assessment administration. While some students did feel that the testlets were too long, teachers did not express general concerns about testlet length, above what is typically encountered with DLM assessments. However, some students were unable to finish due to sensory processing difficulties. Testing times, although recorded during an inauthentic testlet experience, were within acceptable ranges. Thus, in preparation for the large-scale I-SMART pilot, item requirements for testlets were maintained. More data about testing time under authentic conditions will be reported based on the pilot administration.

Response Process and Science Performance Expectations

Related to testlet content and the science performance expectations, think aloud and retrospective comments from students supported that students were interpreting testlet content as intended. Some guessing is expected when students have knowledge gaps or have not received instruction, but many students were able to verbalize the response processes that designers intended based on the science learning map model.

While students at the Initial linkage level sometimes lacked the expressive communication skills needed to indicate their choices, teachers provided assistance with interpreting what student test-taking behaviors could be considered construct-relevant responses. Students at the Initial linkage level were able to answer some questions correctly and some without construct-irrelevant behaviors, especially with familiar content. Additionally, some teachers were concerned that some of the testlet content was too difficult because students had not received instruction on the concepts. Further information about the testlet difficulty will be gathered during the I-SMART pilot administration.

Testlet Perceptions from Students and Teachers

In general, students and teachers liked the testlet content and provided useful and actionable suggestions for improvement. Use of media was a clear theme during interviews, with pictures or video viewed both as an asset and an area for improvement. Several improvements to the testlet navigation and onscreen layout have already been implemented. Some teacher concerns related to accessibility of the content have been shared with test development staff for further evaluation.

Finally, teacher involvement with the cognitive lab sessions was invaluable. Teachers assisted by adapting the session to individual student needs, supporting student access to the testlet content, providing information about typical instruction and assessment practices, and providing overall context for each student's experience. For students in the I-SMART populations, including the teacher in the cognitive lab is a key method in collecting and interpreting response process evidence. Further strategies for encouraging teacher participation will be incorporated into future cognitive lab protocols.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association. https://www.aera.net/Publications/Books/Standards-for-Educational-Psychological-Testing-2014-Edition

Karvonen, M., Swinburne Romine, R., & Clark, A. K. (2020). *Response process evidence for academic assessments of students with significant cognitive disabilities* [Manuscript submitted for publication]. Accessible Teaching, Learning, and Assessment Systems (ATLAS), University of Kansas.