DYNAMIC®
LEARNING MAPS

USABILITY STUDY OF TECHNOLOGY-ENHANCED ITEMS IN SCIENCE

Technical Report #26-01

March 2026

Gholson, Melissa, L., Kobrin, Jennifer, L., Gane, Brian, D., Clark, Amy, K., & Thomas, J. (2026). Usability study of *technology-enhanced items in science*. (Technical Report No. 26-01). University of Kansas, Accessible Teaching, Learning, and Assessment Systems.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

Accessible Teaching, Learning and Assessment Systems (ATLAS) is developing new Dynamic Learning Maps® (DLM®) science assessments aligned to dimensions of the National Research Council's Framework and Next Generation Science Standards (NGSS). This study examined whether technology-enhanced items (TEIs) can validly measure targeted science content knowledge, skills, and understandings, and whether TEI features introduce unintended challenges for students with the most significant cognitive disabilities. The team analyzed how students interacted with each TEI and identified considerations of accessibility, presentation, and engagement for operational use.

Using modified cognitive-laboratory procedures, researchers observed students in two states as they worked through five TEI types. Observational evidence was paired with teacher ratings of student performance to examine item demands for students with a range of complexity-band assignments. Multiple data sources supported a holistic appraisal of each item type. Following each test administration, the research team interviewed test administrators.

The results found three of the item types—drag-and-drop, drop-down, and hot spot—showed the most promise. The interactions and means of responding to these items are similar to instructional activities and digital tools commonly used in classrooms, facilitating student access, and reducing extraneous demands. Students engaged successfully with drag-and-drop items with minimal practice or support. Drop-down and hot-spot items worked well for most students assigned to Complexity Bands 2 and 3 after the orientation item. Two TEI types—table match and simulation—were challenging. Limited familiarity with these item types led to inconsistent test administration, longer student response times, reduced student engagement, and a need for more intensive test-administrator support.

The findings of the study have implications for test design and item development. The study findings indicate that TEIs have the potential to validly measure multidimensional science constructs for students with significant cognitive disabilities. To ensure accessibility, additional considerations for TEI use must minimize navigation load, provide clear visual and interactive cues, render items consistently across devices, and include practice opportunities aligned with instruction. Conversely, formats that require unfamiliar interfaces or use heavy scaffolding risk conflating technology demands with science understanding and should be avoided.

BACKGROUND AND PURPOSE

ATLAS is developing new Dynamic Learning Maps® (DLM®) science assessments intended to support interpretations about what students know and can do and to support inferences about student achievement in science. The DLM Science Essential Elements (EEs), which serve as the extended content standards, were developed to include the three dimensions underlying the National Research Council's Framework and Next Generation Science Standards [1]. Science learning maps were used to identify critical knowledge, skills, and understandings at differing levels of complexity aligned to the EE. These levels, known as linkage levels, provide multiple access points to grade-band science content at reduced depth, breadth, and complexity appropriate for students with significant cognitive disabilities.

As part of the assessment design efforts for the new science assessments, we investigated the viability of using five TEI to measure knowledge, skills, and understandings. DLM science assessments are designed for students with the most significant cognitive disabilities. We conducted a study to determine whether students could respond to the item types as intended, or the technology enhancement in these item types might unintentionally create challenges for our student population. For the investigation, we evaluated the extent to which students were able to interact with five TEI types as intended for their use on an operational assessment. We considered aspects of accessibility, presentation, and engagement.

Technology-enhanced item types offer potentially engaging and efficient ways to assess content (e.g., Parshall et al., 2010) and measure targeted constructs better than multiple-choice items (e.g., Bryant, 2017). A previous DLM item-usability study evaluated the use of drag-and-drop, click-to-place, multiple-select multiple-choice, and select-text TEI types by students with significant cognitive disabilities who were at higher complexity bands (Dynamic Learning Maps Consortium, 2016). Following that usability study, multiple-select multiple-choice, select-text, matching lines, and drag-and-drop items were used on DLM assessments on a limited basis. However, the small number of drag-and-drop items developed were retired due to a lack of compelling evidence that this item type was needed to assess the targeted constructs. TEI are used in limited instances for English language arts and mathematics assessments, but the current science assessments do not include any.

[1] The new DLM science assessments are linked to the three dimensions of the National Research Council's 2012 Framework for K–12 Science Education and the Next Generation Science Standards: Disciplinary Core Ideas (DCI), Science and Engineering Practices (SEP), and Crosscutting Concepts (NGSS Lead States, 2013).

In the 10 years since the previous DLM item-usability study, there has been an increase in both informal (Isaksson & Björquist, 2021) and instructional uses of technology (Gunderson et al., 2017). Students with significant cognitive disabilities may be more familiar with functionalities present in TEI types (Holyfield et al., 2024). Little research on technologies and item types has been conducted specifically for this population (Braun et al., 2025; Karvonen et al., 2024). The Kite Suite used to deliver DLM assessments currently has item types available that have not yet been evaluated for use with students with significant cognitive disabilities. However, students taking DLM assessments vary widely in their disabilities, educational placement, expressive and receptive communication, sensory characteristics, access needs, and academic skills (Burnes & Clark, 2021). Therefore, it is important to evaluate how students taking DLM assessments engage with TEI types before using them for an operational assessment.

RESEARCH QUESTIONS

The purpose of the study was to evaluate the extent to which students can interact with the TEI types as intended, whether the TEIs were engaging and relevant, and identify features of specific TEI types that show promises or challenges for operational use, including any accessibility considerations. We were interested in determining which item types have potential for targeting intended cognition of the science linkage levels and the extent to which they introduce construct-irrelevant factors into the cognitive response process. We evaluated five TEI types: drag and drop, drop-down (cloze), hot spot, table match (matrix), and simulation.

The study addressed the following research questions:

1. What are accessibility considerations for each of the item types?
2. What are the features of presentation that affect student responses?
3. Were the item types engaging and relevant for students?
4. What are the considerations to support accessibility, presentation, and engagement when pursuing each item type?

PURPOSE OF THE REPORT

In this report we present findings from the student interactions with the items, to determine which TEI types have potential for targeting cognition represented by the new multidimensional science EEs. In the results section, we describe response barriers, accessibility considerations, and aspects of TEIs that affect cognitive load related to the item format that might limit students' ability to demonstrate their knowledge, skills, and understanding. In addition to student interactions with the TEI types, we also describe test administrators' impressions and any challenges they faced in supporting students as they completed the TEIs.

Findings from this study can provide evidence for evaluating the appropriateness of these TEI types at different grade bands and with science content, especially with respect to their ability to measure the intended constructs. This report can be used as evidence of technical quality to support test design and item development for assessment. The results of this study serve to inform decisions about which TEI types are useful for measuring the intended skills for DLM science and guide decisions for test design, item development, and considerations for operational test administration.

METHOD

Cognitive labs are often used during item tryouts and purposefully obtain evidence of test takers' response processes. Traditional cognitive labs may pose challenges for students with significant cognitive disabilities due to the cognitive load required to verbalize their thinking (Karvonen et al., 2024; Marion & Pellegrino, 2006). For this study, we used a modified cognitive lab approach established for this population (Karvonen et al., 2024; Tiemann & Karvonen, 2019). Students interacted with the TEI and were interviewed using a structured protocol about their experience.

RECRUITMENT

We visited two states for study. In State A, we visited two districts. District 1 is a public school district of 23 schools serving approximately 9,000 students. We observed test sessions in five different schools (three elementary, one middle, and one high school). District 1 schools were in both rural and suburban settings with a student–teacher ratio of 16:1 and a free-and-reduced-lunch rate of 37%. District 2 is a public school district of 26 schools serving 11,000 students. We observed test sessions in one high school that had a population of 1,700 students from a mixture of suburban and rural areas, a student–teacher ratio of 17:1, and a free-and-reduced-lunch rate of 42%. In State B, we visited District 3, an urban school district with 25,000 students in 69 schools serving only students with disabilities. We observed three schools in District 3, which have an average free and reduced-lunch rate of 89% and a student–teacher–paraprofessional ratio of 8:1.1.

For recruitment, the state leaders identified participating districts and the district's contact leads. We met district leaders and discussed the study and recruitment information. District leaders distributed a letter to recruit prospective test administrators from their district to participate. We requested volunteer test administrators with experience giving DLM assessments and, from the returned letters of interest, we recruited fourteen test administrators.

We used Zoom to meet separately with each participating district and share information with the recruited test administrators. We gave an overview of the study and discussed methods to recruit students eligible to try computer-based item types. Despite our

communication efforts and frequent requests to districts, fewer test administrators at the high school level volunteered to participate. Thus, we had less flexibility in selecting students to participate at that level. Test administrators signed informed-consent documents and sent out parent-consent letters. Test administrators identified their students and collected parent consent forms. We asked test administrators to overrecruit to account for attrition during the study. Only students with completed parent-consent forms were considered for participation in the study.

RECRUITED PARTICIPANTS

Test Administrators

Fourteen test administrators participated, eight from State A and six from State B. Test administrators were teachers with experience in administering DLM assessments who were familiar with the students selected to participate (i.e., rostered teachers in the system and/or typical test administrators). Most test administrators were female, White, non-Hispanic, and with various levels of experience; this pattern is consistent with the population of teachers that deliver DLM assessments. Table 8 (Appendix A) further describes the characteristics of the test administrators.

Most test administrators administered the usability-study items to multiple students. In both states, the median number of students per test administrator was 2 (range = 1–5). Most test administrators did not administer all five item types. The median number of item types across all their students was 2 (range = 1–5).

Participating Students

The eligibility criteria for student participation included students who received computer-delivered testlets. The study team reviewed the complexity-band assignment to select students at the local site. We used a maximum-variation sampling plan to deliberately include a range of students at each science-complexity band for each item type and included other key characteristics. Characteristics of participating students are provided in Table 1.

Table 1

Student Characteristics

Characteristic	State A*	State B*	Total*
Gender			
Female	3	6	9
Male	11	8	19
Race			
American Indian	0	2	2
Black	0	4	4
White	14	7	21
Two or more races	0	1	1
Grade Band			
3–5	3	11	14
6–8	4	3	7
9–12	7	0	7
Final Science-Complexity Band			
Foundational	3	3	6
Complexity Band 1	0	8	8
Complexity Band 2	7	2	9
Complexity Band 3	4	1	5

Note: The numbers represent the number of students by characteristic by State. The total numbers of students who participated in the study.

Tables 9, 10, and 11 (Appendix B) show the distribution of participating students across the expressive-communication band and the science band. (Expressive-communication and science bands are used to compute the final science-complexity band reported in Table 1.)

Table 12 (Appendix C) details of the accessibility supports, devices, response tools, and expressive-communication modes used by participating students. For this study, students used human read aloud, which is consistent with operational DLM administration. Of note, 83% of students responded via touch screen on an iPad in this study, compared to 26% who responded via iPad during operational DLM administration (Dynamic Learning Maps Consortium, 2023).

TEI TYPES

We evaluated five TEI item types: drag and drop, drop-down, hot spot, table match, and simulation. In the drag-and-drop item type, students drag textual, numerical, or graphical components and drop them onto an image. In the drop-down item type, students select a word, phrase, number, symbol, or expression from a drop-down menu to complete a statement or expression. This item type is similar to the cloze procedure used for reading instruction. Cloze procedure uses a fill-in-the-blank strategy, or and is often used with an augmentative and alternative communication (AAC) device (Holyfield et al., 2024). In the

hot-spot item type, students select one or more components of an image, table, or chart. In the table-match item type, students use radio buttons to select answer choices in a table format. In the simulation item type, students manipulate one or more variables in the form of selectable options. After a student selects the desired option for each variable and activates the simulation, outputs are returned in a table format. The outputs include animated components.

TESTLET DESIGN AND ASSIGNMENT

Before the study began, ATLAS staff reviewed the new science EEs and linkage levels to determine the range of cognition being measured, which TEI types may appropriately measure the constructs, and a rationale for how the TEI type may measure the construct in ways that a multiple-choice item type may not. For this study, each testlet contained four items of a single TEI type: two *content-neutral* items and two *science-light* items. The content neutral items did not include any science content. The science light items contained science content at a reduced complexity. The combination of content-neutral and science-light items for each TEI was a design decision to maximize the information we collected.

The content-neutral items were designed to identify any TEI types with interface features or response requirements that would be too challenging or inaccessible for students. The first item of each testlet was an orientation item that gave students an opportunity to practice the item type and see how to respond. Students were allowed multiple attempts to complete the item, if needed, and test administrators could model the response. The second item, a content-neutral item, used content that did not rely on prior academic science knowledge. For example, a content-neutral, drag-and-drop item may ask students to drag an image and drop it in a specific location, but it could not ask students to drag-and-drop only the images of solids (not of liquids or gasses) because that would require using science knowledge to classify something. The purpose of the item was to determine whether students could independently respond, without accessibility barriers.

The science-light items were designed to help us infer whether students could successfully interpret and respond to each TEI type as it may look in a science item. The third and fourth items included science content. The science-light items were designed so that the fourth item was more complex than the third item. The science content used for items were developed using the elementary EEs (i.e., grades 2–5). Science content was selected that was considered widely familiar to students (e.g., day and night, moving objects, plants, and animals). The concepts were chosen to be unlikely to pose opportunity-to-learn concerns for students who have not yet received instruction on the new science EEs. Because these concepts are foundational to the more-complex knowledge, skills, and understandings in

the middle school and high school EEs, they were relevant to students in all grades. Appendix D includes examples of the items used in this study.

Each student received two testlets (i.e., responded to two TEIs). The testlets were ordered according to the test-development team's perception of the functional complexity of the item types and the amount of fatigue each item type may cause students. Ordered from least to most perceived complexity and potential for fatigue, the item types were drag and drop, drop-down, hot spot, table match, and simulation. The testlet with the less functionally complex item type was always presented first (e.g., drag and drop before hot spot). The testlets were delivered through Kite Student Portal, the online platform used to deliver DLM assessments. Because of a delay in development the simulation item types were not available for one of the state site visits; all students at the second state site visit received the simulation item type and one other item type, which allowed us to evaluate the simulation item type.

All item types were delivered to students that spanned the full range of final science-complexity band assignments (Foundational and Complexity Bands 1–3; see Table 2). Table 3 shows the number of students who interacted with each item type by grade band and complexity band. Because of the small sample sizes and because we were interested in evaluating each TEI type's potential for use for students across complexity-band assignments, we used a maximum variation sample. We later collapsed the complexity bands assignments into two categories: students assigned to Complexity Band 2 or 3 and students assigned to the Foundational Band or Complexity Band 1.

Table 2

Number of Students Taking Each Item Type, by Final Science-Complexity Band

Item Type	Final Science-Complexity Band			
	Foundational	Complexity Band 1	Complexity Band 2	Complexity Band 3
Drag and drop	4	3	4	1
Drop-down	2	2	5	3
Hot spot	3	4	2	2
Table match	1	2	5	3
Simulation	2	5	2	1

Table 3

Number of Students Taking Each Item Type, by Grade Band and Final Science-Complexity Band

Item Type	Grade Band			Total
	Elementary (3–5)	Middle (6–8)	High School (9–12)	
Complexity Band 2 or 3				
Drag and drop	2	2	1	5
Drop-down	3	3	2	8
Hot spot	0	2	2	4
Table match	3	2	3	8
Simulation	2	1	0	3
Foundational band or Complexity Band 1				
Drag and drop	3	1	3	7
Drop-down	3	0	1	4
Hot spot	5	0	2	7
Table match	2	1	0	3
Simulation	5	2	0	7
All bands				
Drag and drop	5	3	4	12
Drop-down	6	3	3	12
Hot spot	5	2	4	11
Table match	5	3	3	11
Simulation [a]	7	3	0	10

Note. [a] Simulation items were not ready in time for the initial observations. Recruitment efforts in the second state did not yield any opportunity to include high school students.

Accessibility supports were provided for participating students. We reviewed and determined the supports in the Kite Support Portal most likely to present accessibility challenges (magnification [2x, 3x],] color contrast, and both single- and two-switch use). During student selection, we strived to include students whose Personal Needs and Preferences (PNP) profile had these features enabled. Table 12 (Appendix C) describes the display enhancements and audio and environmental supports used by students in our sample. Only three students used magnification. Because we used a think-aloud method, we did not make spoken audio available; test administrators provided all reading support. For 21 students, human read-aloud support was identified in their PNP. None of our participants used switches or needed language or braille support.

STUDY PROCEDURES

ATLAS staff members participated in on-site visits to observe item-usability sessions and conduct test-administrator interviews in late February and early March 2025. The sessions were conducted in the classroom or in a separate room in the school building where the student typically completed DLM assessments. Test administrators were instructed to

follow typical DLM administration procedures to allow flexibility and student support. Students could use alternate communication to respond. We recorded each test administration using screen captures via Zoom.

Student-Lab Sessions

Test administrators who were familiar to the students and administered testlets conducted the sessions in a quiet location in the school building. Each session lasted 30–40 minutes. Using the first item in the testlet, test administrators modeled the item type and gave students time to interact and respond to the item. After the practice item, students could complete the next item independently. If a student did not respond, the test administrator prompted the student to respond to the item and gave them another opportunity to respond. If the student still did not respond, the test administrator supported the student using only DLM allowable practices for two trials. Test administrators followed this procedure with the other two science-light items on the testlet.

Test-Administrator Notetaking Form

The test-administrator notetaking form was designed to obtain data on each administered testlet from teachers who were attuned to their students and could evaluate students' performance against how they normally performed (Leighton, 2017). For each testlet in the test administration, test administrators completed a structured protocol with questions and space to take notes. The test administration protocol asked the test administrator to document students' responses to each of the four test items. Questions asked if students performed as expected, whether they gave consistent yes–no responses and asked test administrators to identify difficulties with the item type, and if the item type was effective for this student. A separate note taking form was used for each student and one set of questions for each item type administered. The research team reviewed the form and used the information to triangulate the information collected on the observation form and to inform the open-ended questions during the teacher interview.

At the end of each testlet, the test administrator interviewed the student using the notetaking form. The questions on the protocol used a yes–no response format like that used during instructional practice; this design is consistent with questioning used in other modified labs (Karvonen et al., 2024; Tiemann & Karvonen, 2019). Students were asked whether the activity was fun, how well they thought they performed whether the activity was easy or hard, whether they knew what to do, and whether anything was confusing. Because many of the students did not give a consistent yes–no response, we did not use those findings for analysis. Students responded positively, but some were able to elaborate and say they liked it or that it was fun and easy.

ATLAS Staff Observation

As students interacted with the items, ATLAS staff completed a structured observation form during each test-administration session. The form included fields to note student behaviors and interactions. Observers considered how students completed each of the four items in their testlet. Students were provided with allowable supports and were often able to complete an item independently (i.e., without any support). However, a few students needed test-administrator modeling.

Test-Administrator Interviews

After each student completed two testlets (i.e., two different TEIs), ATLAS staff used a semi-structured protocol to interview the test administrator. They asked follow-up questions informed by data collected from the test-administrator notetaking form (e.g., "Tell me why you think the functioning for this item type was hard for this student?"). Staff used item descriptions and printed copies of the test items to solicit test-administrator feedback on each TEI type. Interview questions obtained feedback on each item type. Staff probed for more information on opportunities and challenges of interaction, interpretation of students' responses, and whether the item type might more broadly challenge all their students who take DLM assessments. ATLAS staff also elicited feedback for evidence to support the potential use of each TEI.

DATA ANALYSES

We collected evidence to evaluate whether TEIs could support the design of testlets that are engaging and instructionally relevant. We evaluated each data source for evidence of whether the TEIs were accessible and elicited response options that reflected the knowledge, skills, and understandings of the student. We specifically sought to determine whether students understood the instructions or had difficulty with any part of item functionality (e.g., selecting, dragging), and what level of assistance students needed when interacting with the items. Additionally, we considered whether student interactions with the items may vary depending on student needs for accessibility support (e.g., magnification) or assistive technology (e.g., switch use). We noted whether any TEI types required special considerations for administration (e.g., test-administration guidance for students who are blind or have visual impairments).

We reviewed the screen captures from test administration as an additional source for reviewing performance and navigation, as well as to obtain the time required (minutes per each testlet administered). We used multiple sources to obtain information including observer notes and ratings, test-administrator interviews, and test-administrator responses on their test-administration protocol. Data from the observation form and test-administrator notetaking form were analyzed using descriptive statistics. Working with a

deductive coding approach, three ATLAS research-team members coded the test-administrator interview transcripts using a predetermined codebook that aligned each source to accessibility, item presentation, and student engagement (see Table 4).

One researcher coded each transcript independently. A second researcher reviewed the codes, added codes they felt were missing, and/or left comments when they disagreed with a code. These two researchers then met to discuss and resolve any discrepancies. A third researcher was used to settle any disagreements. The final codes were aggregated by item type and summarized in tables, along with illustrative examples and test-administrator quotes when applicable (see Tables 14-18 in Appendix E).

Table 4

Interview Coding Structure

Interview Code	Subcodes and Definitions
Accessibility	• Accessibility supports: Use of supports and impact of use of supports on student responses (magnification; human read aloud, test-administrator actions, such as entered responses, physical support, reduced answer choices, repeating directions) • Technology: Impact of technology on student responses (content rendered as intended, item-type functioning, scrolling effects) • Navigation: How students navigated screens (independently, after verbal prompts, after test administrator [TA] pointed or gestured, TA navigated and selected answers)
Presentation	• Display of information: Impact on student responses due to layout, size of fonts, images, graphics, video, spacing, sizing • Task design: Impact on student responses due to task prompt, directions, stimuli, visual/TA cues or distractors, complexity • Language: Impact on student responses due to language/text complexity • Test-administration directions: Clarity of directions, one step vs. multiple step
Engagement	• Effort: attention to task, time spent on task • Tasks that optimize independence • Student satisfaction • Student attention • Task relevance: Meaningful, familiar, prior opportunities to practice instructionally.

RESULTS

We report the results for each item type, summarize the relevant findings from the qualitative analysis of the data collected from each session. For each item, observers categorized how the student completed each of the test items, including how the student responded (independently, *with allowable supports*, *only with test-administrator modeling*, or *did not respond*. These ratings are presented for each item type, with two horizontal

panels corresponding to either students in final science-Complexity Bands 2 and 3 (e.g., Figure 1, top panel) or students in the Foundational band and Complexity Band 1 (e.g., Figure 1, bottom panel). Data in the figures are slightly jittered to prevent overplotting. We also describe quantitative results for each TEI type, including ratings from observers (see Tables 5 and 6 and Figures 1–5) and test administrators (see Appendix F).

DRAG-AND-DROP ITEMS

The pattern of student completion[2] across the four drag-and-drop items is displayed in Figure 1. All students in Complexity Bands 2 and 3 were able to complete all four drag-and-drop items independently or with allowable supports. Although one student in the Complexity Band 1 completed all items independently, most did so with allowable supports. All but one student in the Foundational band was able to complete the drag-and-drop items with allowable support and did not require test-administrator modeling. One student in the Foundational band required test-administrator modeling for the science-light items but not the content-neutral items. Another student in the Foundational band, who typically received teacher-administered testlets and not the computer-based assessment, experienced challenges responding to the drag-and-drop item type and did not respond after the orientation item (Items 3 and 4 were not administered).

In 100% of the observations, observers rated students' engagement with this item type as either *somewhat engaged* (*n* = 2, 17%) or *very engaged* (*n* = 10, 83%; see Table 5). Observers indicated all but one student were able to complete this item type effortlessly (*n* = 11, 92%; see Table 6). Test administrators felt this item type was *easy* (*n* = 4) or *about right* (*n* = 5) for all but one student (see Appendix F, Table 19) and that the item type was *somewhat effective* (*n* = 2) or *very effective* (*n* = 7) for all but one student (see Appendix F, Table 20).

Overall, students liked this TEI type and knew how to respond. Many students were familiar with this type of item because it is used instructionally and informally. All test administrators stated that students were successful with drag-and-drop functionality in other technology-based applications. Test administrators further offered that the features of drag and drop were very useful for demonstrating knowledge of sorting, ordering, classifying, categorizing, labeling images for part-to-whole relationships (e.g., parts of a plant), identifying sequences (e.g., water cycle), and labeling groups (e.g., group of flowering vs. nonflowering plants).

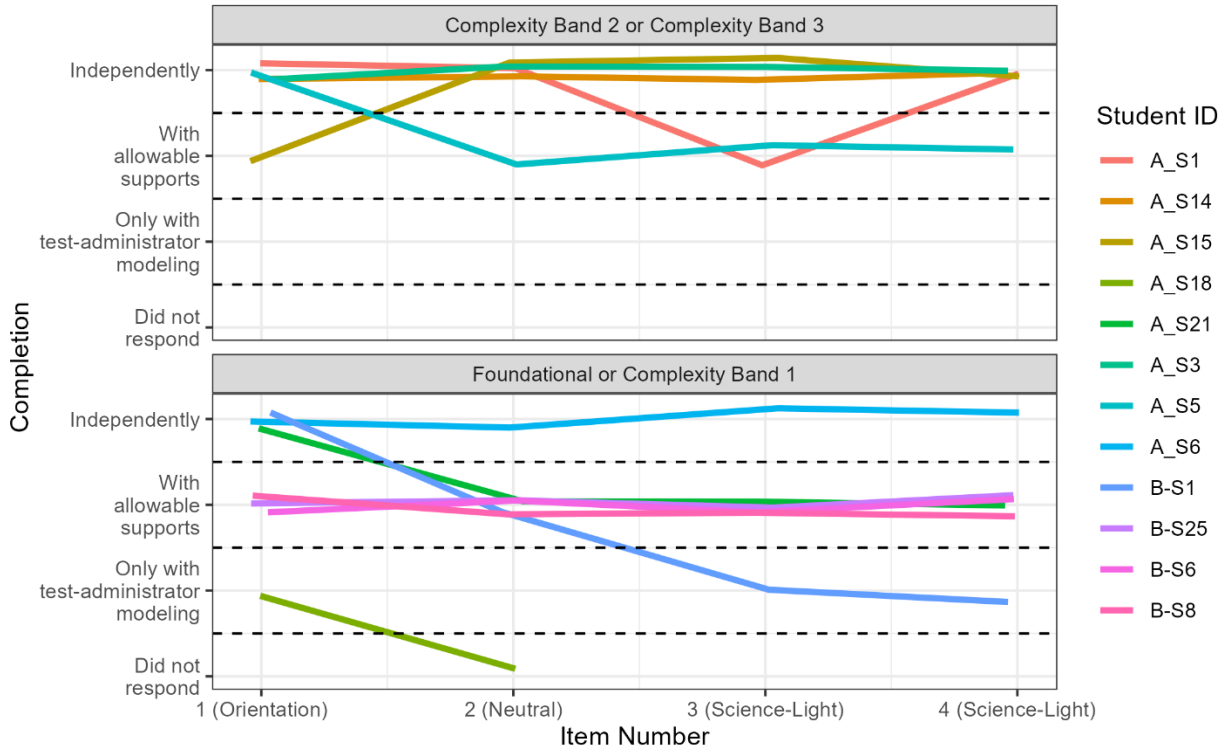
Most students could complete this item type without any test-administrator modeling. The dotted lines to mark drop targets (see example in Appendix D) helped students identify

[2] Based on the study team's observer ratings form during the session.

where to place the object. Students selected an image and focused immediately on the action. One student pointed instead of touching the screen, so the test administrator completed the action. The drag-and-drop item type was successful for students assigned to final science-Complexity Bands 1–3. At the Foundational band, the drag-and-drop item was very challenging for at least one student who took DLM teacher-administered testlets. Detailed results of how students interacted with this item type are organized by deductive codes and presented in Appendix E.

Figure 1

Student Completion of the Four Drag-and-Drop Items



DROP-DOWN ITEMS

All but one student in Complexity Bands 2 and 3 were able to complete the four drop-down items independently or with allowable supports (see Figure 2). One student in Complexity Band 2 required test-administrator modeling for the orientation item and again later for a science-light item. Three of the four students in the Foundational Band and Complexity Band 1 used allowable supports or required test-administrator modeling or both. In both sets of complexity bands, some students required test-administrator modeling only for the orientation item, which might have reflected difficulties with navigation or response selection.

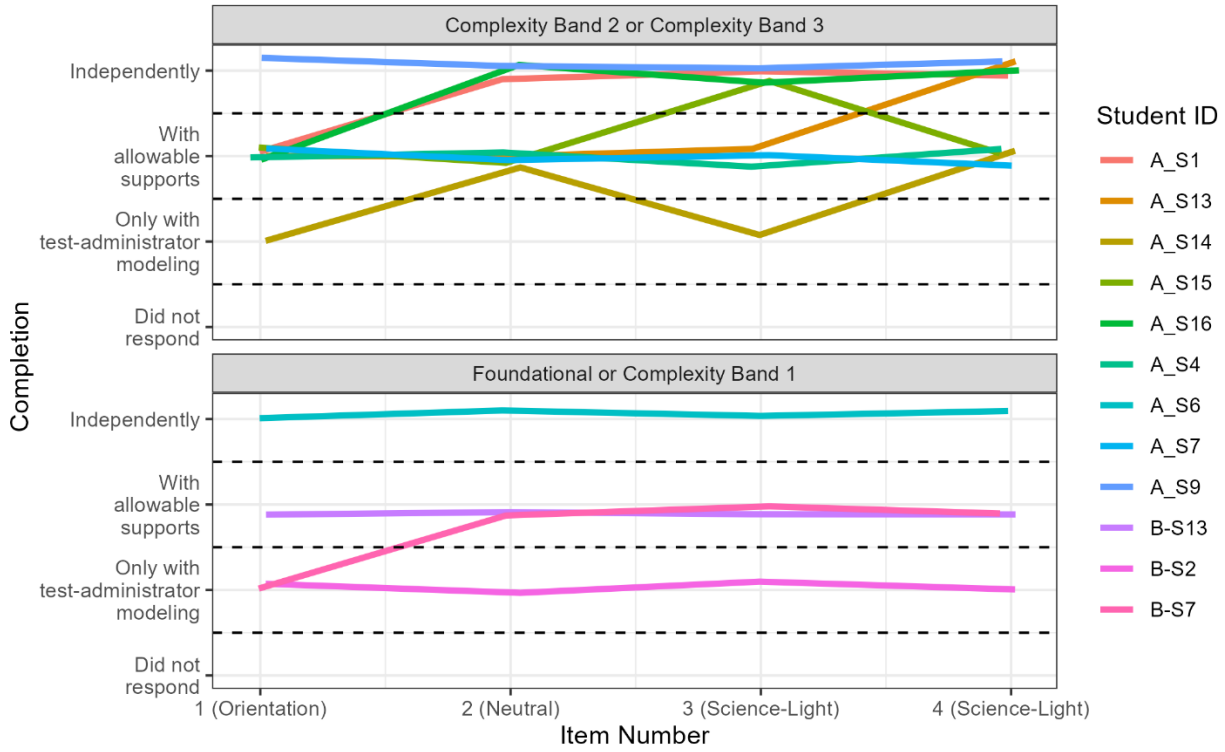
For items that presented one- or two-row response options, students selected the responses quickly (see example items in Appendix D). Students made more errors when

items had three rows of response items for the two categories. The item's functionality was relatively easy for students to understand, but somewhat hard to navigate independently for some students who had difficulty selecting the drop-down menu arrow button, which was extremely small.

Test administration varied. Some teachers navigated and presented the response options by reading each sentence separately with each response option; other test administrators read the sentence once with both response options. Beyond navigating the item, students either had to read all content or have all content read to them. One test administrator thought this item type may have measured listening.

Figure 2

Student Completion of the Four Drop-Down Items



Despite this variability, this item type received high ratings from observers and test administrators. Observers rated more than 80% of students' testlet completions as *effortless* (see Table 6). Test administrators rated the item type as *somewhat effective* or *very effective* for 90% of students who completed it and as *easy* or *about right* for 88% of students (see Appendix F). Some students in grade 4 and above had experience with this type of item, both instructionally and with assessment (e.g., i-Ready). Educators felt that picture-and-word combinations would support some of their students who did not read well. Based on observation data, students were more confident using numbers instead of words.

HOT-SPOT ITEMS

The hot-spot items showed different completion patterns for students in Complexity Bands 2 and 3 compared with students in the Foundational Band or Complexity Band 1. Three of four students in Complexity Bands 2 and 3 completed the hot-spot items independently (see Figure 3). Although none of the seven students in the Foundational Band or Complexity Band 1 completed the items independently, over half of them were able to complete the items with allowable supports. Two students in the Foundational Band or Complexity Band 1 required test-administrator modeling on the first or first and second items but were able to complete later items without the modeling. One student, who was assigned to the Foundational band using an AAC device and typically received teacher-administered testlets, refused to complete the third and fourth items.

Overall, students liked this item type and progressed quickly through these items. Students were able to select the hot spot with up to two response targets with few errors. The items were efficient and maintained students' attention. The hot spot items were *somewhat easy* for students to learn and perform. Test administrators liked this TEI type for all their students. This item type uses images which is similar to individual visual picture supports that students in this population use instructionally and for supporting communication (e.g., pointing, selection of picture supports).

One challenge with this item type was that students had to unselect their first answer before changing their answer. This process was frustrating when they wanted to change their response but did not know they had to unselect their first answer. Similarly, students who wanted to make a change after selecting the maximum number of responses did not know they had to unselect a response to make a change.

Observers indicated that students responding to this TEI type were engaged and that responding did not require a lot of effort (see Table 5 and Table 6). Test administrators indicated the item was *easy* or *about right* for all students except one (see Appendix F, Table 19). Test administrators noted two times that this item type was *not at all effective*; once for a student in the Foundational band and once for a student in Complexity Band 2 (see Appendix F, Table 20). Two different reasons were given for these ratings. The item was rated *not effective* for the student assigned to Complexity Band 2 because the test administrator felt that a "deeper level of knowledge" needed to be tested. For the student assigned to the Foundational band, the test administrator noted that the student needed physical manipulatives and did not understand how to interact with an iPad or computer.

Figure 3

Student Completion of the Four Hot-Spot Items

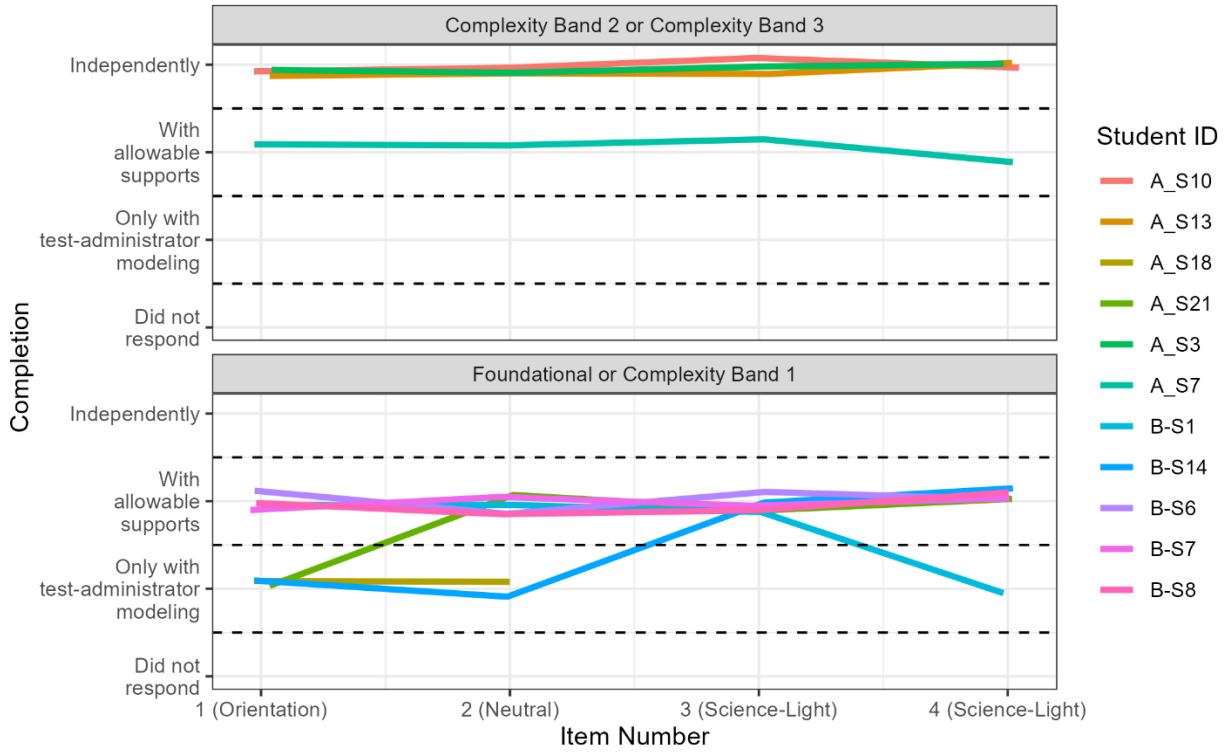
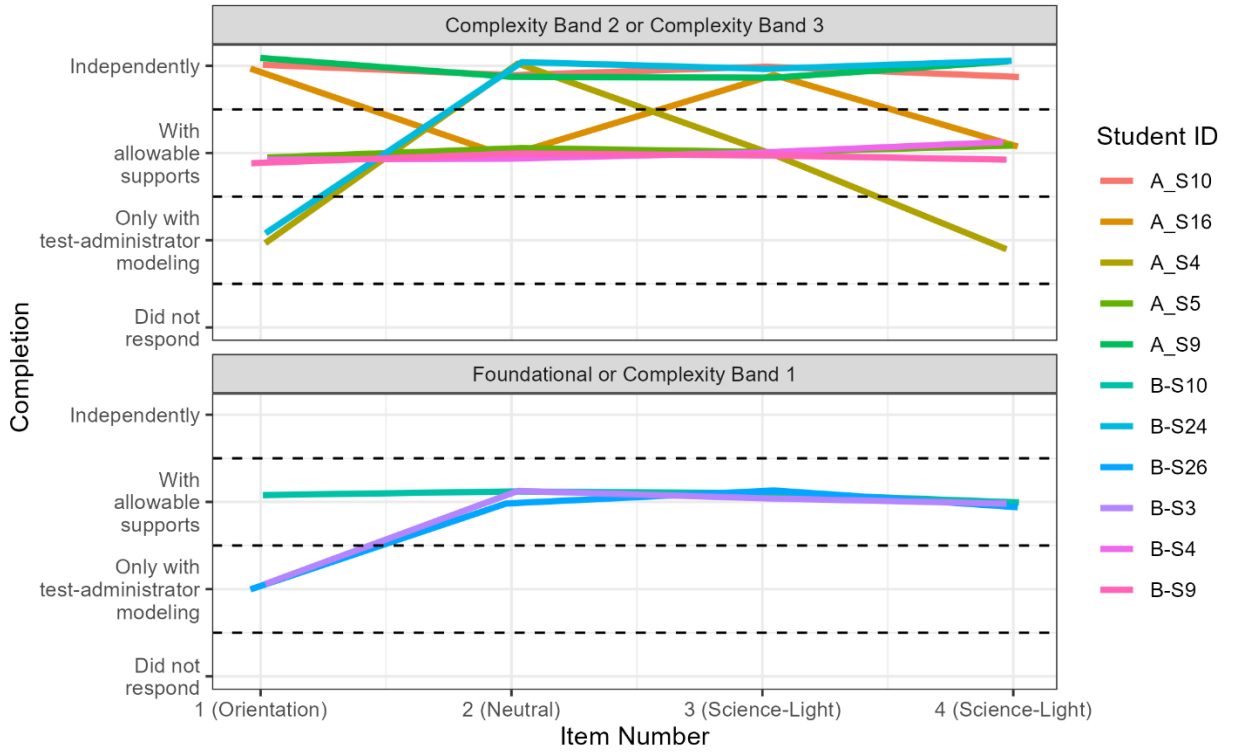


TABLE-MATCH ITEMS

Six of the eight students in Complexity Bands 2 and 3 completed all four table-match items independently or with allowable supports (see Figure 4). Three students in the Foundational band and Complexity Band 1 completed the second, third, and fourth items with allowable supports; none did so independently. In both groups of complexity bands, some students (*n* = 4; 36%) could complete only the orientation item with test-administrator modeling.

Figure 4

Student Completion of the Four Table-Match Items



The table-match item type used radio buttons, which presented a challenge for a few students. The small size of the buttons and their location required focused attention for response, causing some students to lose track of their intended target. Radio buttons were extremely small on the iPad; students had difficulty navigating the item and had to click on the buttons several times.

Variability of test administration occurred with this item type. Some administrators reread the table title for each item and read the possible responses sequentially for each individual option, which led to more time in test administration. Other test administrators did not reread the category of the table title. Students were more successful when test administrators consistently pointed to the top of the table and repeated the category for each item. Students could not remember the column headings or were confused when the table match had more than one set of matches (i.e., additional rows). Their responses often defaulted to selecting the first button in a row when more than two rows were presented. One teacher stated that students had "difficulty with tables, charts, and organizers," and some students may not have been familiar with their format.

Some test administrators did not believe this item type worked well for their students and thought it required more time and effort than other item types. In terms of test administrators' ratings of effectiveness and item functioning, this TEI type had substantial

variability in test administrators' ratings of effectiveness and item functioning, spanning all categories (see Appendix F). Observer ratings indicated students showed high engagement but needed to exert more effort than with drag-and-drop, drop-down, and hot-spot items (Table 5 and Table 6). Observational data indicated longer student response times for this item type relative to the drag-and-drop, drop-down, and hot-spot items.

SIMULATION ITEMS

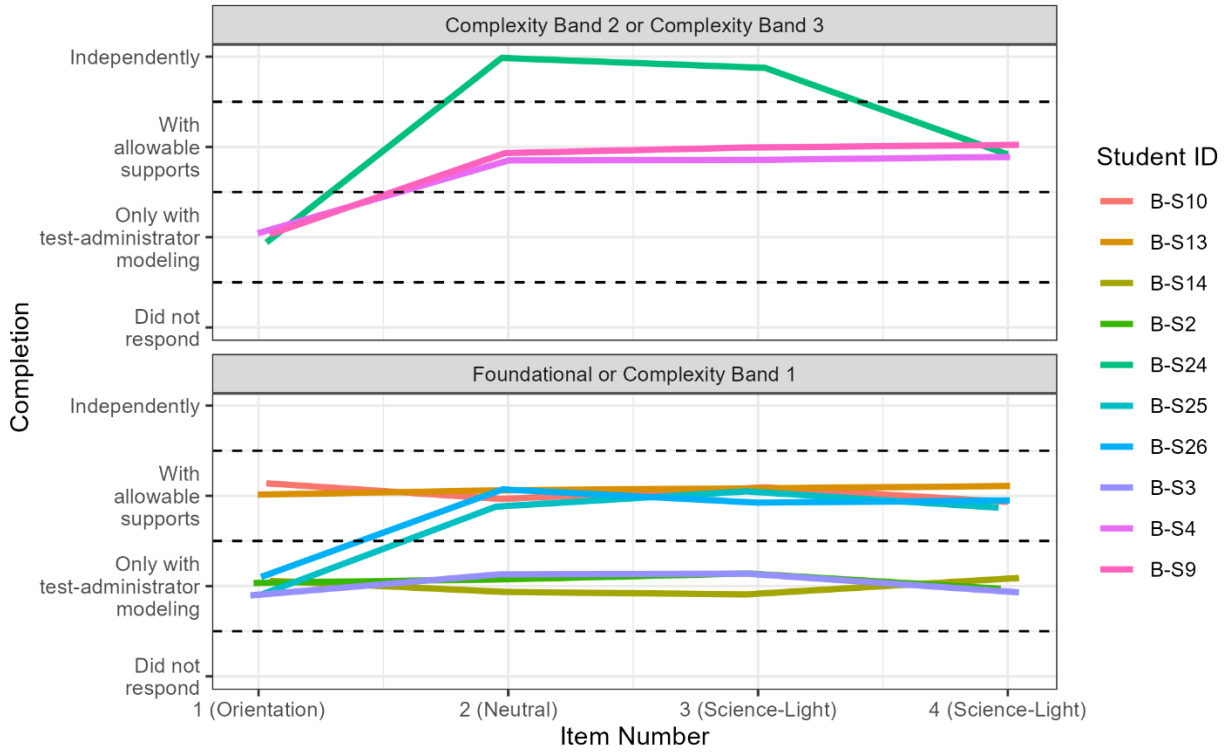
Because of extended time to develop items and recruitment constraints, the simulation items were not administered to high school students (see Table 3), who might have been a more appropriate audience. Seven of the 10 students who completed this item type were assigned to either the Foundational Band or Complexity Band 1. Of these students, none were able to complete any of the four items independently, and about half completed the items with test-administrator modeling throughout the session (see Figure 5). Almost all students, including the three students assigned to Complexity Band 2 and 3, required test-administrator modeling on the orientation item but successfully completed the subsequent items independently.

The item type required substantial interaction and direction from the test administrator. Administration time was lengthy, which led to decreased attention; for example, one student left the session, and the test administrator had to help the student return. Simulation items required students to perform multiple activities, which students found confusing. Students often sought guidance from the test administrator and relied on the test administrator to navigate the item. Test administrators redirected students throughout the test administration. Students often did not know what to do. Test administrators commented that the item type had promise; they had no opportunity to use these instructionally. Test administrators rated the simulation item type *hard* or *somewhat not effective* more frequently than other item types (see Appendix F).

Test administrators were unfamiliar with the administration of this item type. The items required multiple steps, which added complexity, and administration time for simulation item types was the longest of all item types. The item prompts were sometimes unclear to both test administrators and students. Nevertheless, students enjoyed many aspects of the items (e.g., videos, content of cars, ice cream), even when they did not understand what they were doing. Although students were very engaged (see Table 5), very few students were able to complete them effortlessly (see Table 6). This item type shows some promise and was motivating for several students but was most successful for students at higher complexity bands. Simulation items may be more relevant to high school students.

Figure 5

Student Completion of the Four Simulation Items



RATINGS AND TIMING DATA FOR ALL TEI TYPES

After each session, observers used a Likert-scale item which rated the student's engagement and effort. Overall, students were *highly engaged* across all item types (see Table 5). The drag-and-drop item required the *least effort*, and the simulation item required the *most effort* (see Table 6).

Table 5

Observer Ratings of Students' Engagement, by Item Type

Item Type	***Not Engaged***		***Somewhat Engaged***		***Very Engaged***	
	N	%	*n*	%	*n*	%
Drag and drop (*N* = 12)	0	0	2	17	10	83
Drop-down (*N* = 12)	0	0	2	17	10	83
Hot spot (*N* = 11)	0	0	3	27	8	73
Table match (*N* = 11)	0	0	1	9	10	91
Simulation (*N* = 10)	0	0	2	20	8	80

Table 6

Observer Ratings of Students' Effort, by Item Type

Item Type	***Took a Lot of Effort***		***Took Some Effort***		***Effortless***	
	n	%	*n*	%	*n*	%
Drag and drop (*N* = 12)	1	8	0	0	11	92
Drop-down (*N* = 12)	0	0	2	17	10	83
Hot spot (*N* = 11)	1	9	2	18	8	73
Table match (*N* = 11)	1	9	4	36	6	55
Simulation (*N* = 10)	0	0	8	80	2	20

After each testlet was administered, the test administrator rated how well the item type functioned for students and how effectively it measured what the students knew. Not all test administrators completed and returned this form for each testlet; as a result, the counts in Table 19 and Table 20 in Appendix F are lower than the number of testlets administered. The drag-and-drop, drop-down, and hot-spot item types were rated as *mostly easy* or *about right*, whereas the simulation item type had the highest frequency of *hard* ratings. Ratings were positively skewed and educators may have desired to please the research team. The drag-and-drop and drop-down items were rated as *somewhat effective* or *very effective* in all but one case.

Timing data were obtained from the recording of the test administration for each testlet. The average response time for completing all four items per testlet for each item type was calculated (see Table 7). Students needed the least amount of time to complete the drag-and-drop item types. The student responses demonstrated the highest efficiency in terms of time to respond and required the least amount of support from test administrators of all item types. In contrast, the response time for simulation items was longer and

necessitated greater assistance from administrators. Factors that may influence the response time of simulation items is that the study include fewer students in higher complexity bands were not available for the study. The simulation item type required students to be engaged for longer amounts of time and included abstract concepts and multi-step directions.

Table 7

Response Times, by Item Type

Item Type	Average Response Time per Testlet, in Minutes
Drag and drop	1.2
Drop-down	3.3
Hot spot	2.5
Table match	3.5
Simulation	10.4

DISCUSSION

This study evaluated the viability of using five item types in DLM science assessments. The findings indicate that while some TEI types are immediately viable, others present significant challenges that must be addressed to ensure valid measurement of student knowledge. Beyond the specific item's performance, broader themes regarding accessibility, technology, and test development emerged.

COMMON THEMES ACROSS TEI TYPES

Across all TEI types, an important consideration is whether students understand how to perform and/or respond to make use of the technology-enhanced aspect of the item. Test administrators recognized that students' independent responses for a given item type took less time when the student had prior experience or instructional familiarity with the item type. In this study, we consistently observed this outcome in the performance of items across all types. Students who had prior experience with an item type interacted with the items immediately, either independently or with allowable supports. In a review of all item types, we found item-type accessibility to be a good predictor for successful response to an item type. Particularly, we found familiarity, consistency, and format to impact student responses.

- Familiarity
    - Success was strongly linked to whether a student had prior experience with the specific technology-enhanced format. When students recognized how to interact with an item type, their independent performance improved.
- Consistency

    - After students successfully engaged with the first item in a testlet, they were generally able to manage the mechanics of subsequent items.
- TEI Format vs. Content
    - From the outset, performance challenges were consistently linked to the accessibility of an item type, and not to the introduction of science-light content. We did not observe any significant increase in support needs solely because the science-light content was added.

Accessibility challenges across all item types can be categorized into visual, cognitive, technical, and support-related issues.

- Visual Presentation and Layout
    - Images and text were often too small to be easily accessible.
- Cognitive Load and Memory
    - Complex formats (e.g., table match, simulation) taxed students' short-term memory, causing them to forget the task instructions while attempting to manipulate the item.
- Fatigue
    - Items that required sustained attention (e.g., table match, simulation) led to student fatigue and decreased focus.
- Complex Directions
    - Lengthy or multistep directions (e.g., table match, simulation) were difficult for students to process and remember.
- Screen Usage
    - Content was sometimes left-justified rather than using the entire screen, which unnecessarily limited the size of the interactive elements.
- Technical Functionality
    - Technical issues in simulation items, such as videos failing to play or slow loading times, created barriers to access by breaking students' engagement and focus. For example, students having to click multiple times for the arrow in drop-down items to render.
- Scrolling
    - Some simulation items required scrolling, which hindered accessibility and caused students' attention to wander. Magnification increased scrolling in all item types.
- Overreliance on Administrator Support
    - Items that relied on text-only options (e.g., drop-down, simulation) forced students to depend on test administrators to read the answer options aloud, reducing their ability to respond independently. Complex items (e.g., table

match, simulation) required administrators to help students navigate the interface rather than just demonstrate content knowledge.

IMPROVING ACCESSIBILITY AND TECHNOLOGY

Accessibility is paramount to ensure students can demonstrate their knowledge for any selected item type. While some TEI types worked better than others, we identified ways to improve usability and accessibility for all five TEI types, especially DLM computer-based assessments. We also identified specific accessibility challenges for each item type (see Table 21 in Appendix G).

The study revealed ways to improve user experience and create more intuitive and accessible TEI items for students who take DLM assessments. These recommendations include considerations for universal design and the POUR (Perceivable, Operable, Understandable, and Robust) principles of web accessibility[3]. We provide specific areas for improvements and recommendations for better accessibility of each TEI (see Appendix H).

Visual and Interface Design

A recurring challenge was the efficient use of screen space. Many items were left-justified, resulting in significant unused whitespace and forcing interactive elements to be smaller than ideal. To enhance accessibility, we recommend all item types use the full width of the screen. Information and user-interface components must be presented in ways that users can easily perceive. Engagement tends to be significantly higher when content is fully rendered on a page. Expanding the spatial layout to occupy the entire page would allow larger images and text, benefiting students with visual impairments.

Interactive elements that are essential for navigation—such as drop-down menu arrows, table-match radio buttons, and hot-spot target regions—should be enlarged to accommodate students with fine-motor challenges (Braun et al., 2025). Increasing the size of buttons, fonts, and images is critical for accessibility, and these interface enhancements do not detract from the science content. To improve accessibility, consider the following suggestions:

- Hot-Spot Interaction
    - The current requirement to unselect a response before making a new one is counterintuitive. The interface should allow for immediate changes in selection, which would enhance both operability and understanding.

[3] Universal design focuses on flexibility, simplicity, and efficiency, aiming to create from the start that are inherently accessible. The POUR principles are a blueprint for accessible design originating from the Web Content Accessibility Guidelines (World Wide Web Consortium, 2018).

- Magnification Considerations
    - We observed that magnification tools often interfered with item layout on smaller devices, frequently pushing content off-screen and necessitating scrolling. This added physical navigation may impede independent interaction. All items should be evaluated to understand the impact of 2x and 3x magnification. It is vital that user-interface components and navigation remain operable and perceptible.

Device Variability

The variety of devices used for assessment has significant implications for all DLM test takers. While testing devices vary across states and districts—74% of students used Chromebooks or personal computers (DLM Consortium, 2023)—this study was limited to observing two devices: iPads and laptops with touchpads. Most participants (79%) in this study used an iPad. According to teachers, most students use iPads for instruction, providing them with many opportunities to use touch screens. Consequently, we observed that device familiarity played a major role in how students responded.

Using unfamiliar devices can negatively affect student performance. For instance, we observed a student who exclusively used an iPad for instruction but was assigned a laptop for assessment. The test administrator acknowledged the student could have responded more easily on an iPad, but the district had mandated the laptop for testing. The student became overly focused on navigating the unfamiliar touchpad, shifting attention back and forth between the science content and the mechanics of inputting a response.

Furthermore, when item content does not render as intended across different devices, it can create unintended consequences and construct-irrelevant variance. Additional tools can also affect perceivability; specifically, we observed an interaction between magnification and increased scrolling. On some smaller devices, item content appeared incomplete. When magnification was selected, it increased the need for vertical and horizontal scrolling, making independent interaction difficult. This display issue may lead to student fatigue caused by the physical demand of scrolling or may cause students to view the task as a series of unrelated items.

This study did not include any single-switch users or other PNP options, such as color contrast adjustments. Future studies should investigate effects across a broader range of testing devices, examine the impacts of other PNP options, and evaluate the response process for full grade-level science content rather than simplified science-light versions.

TEST DEVELOPMENT OF TEIS

The selection of TEIs must be driven by their ability to elicit the specific knowledge, skills, and understandings represented in the science EEs. To achieve this result, it would be beneficial to conduct a needs analysis to identify which science knowledge, skills, and understandings are best assessed using TEIs. Crucially, universal design and accessibility must be integral to the task-design process.

To support the creation of high-quality items, test developers should implement the following strategies:

- Universal Design and Item Specifications
    - Future item development requires rigorous adherence to universal design and POUR principles. Item-writing guidelines for TEIs must be specific and detailed.
- Prototypes
    - Developing a prototype for each TEI type will ensure consistency and guide item writers. These prototypes should include specifications for graphical file types, response markers, and physical layout. For example, for a hot-spot item, writers need to specify supported file types, define the response marker displayed, and clearly denote both correct (i.e., key) and incorrect regions on the image. Additionally, considering the physical layout on the screen and the instructions required to respond will serve as a useful guide for future development.
- Cognitive-Load Management
    - Item writers must be cautioned against designs that unnecessarily tax students' short-term memory.
- Administrator Support
    - Test developers may need to provide additional administration directions to support consistent, construct-relevant delivery. Test administrators require detailed guidance on how to assist students with new item types. This guidance includes:
    - Student Directions
        - Directions provided to students during administration should be chunked to facilitate understanding. It is vital that students clearly understand how the item type functions and what is expected of them.
    - Teacher Guidance
        - Future development should reflect on the importance of detailed instructions. The Testlet Information Pages for teachers must explicitly address how to present each TEI type effectively.

TIME CONSIDERATIONS

Time spent on task is a critical factor for engagement, particularly for students with significant cognitive disabilities. In this study, we observed a direct, inverse relationship between time on task and student engagement across TEI types. Shorter items fostered higher engagement and independence, while longer items led to fatigue and increased reliance on test-administrator support. For example, the drag-and-drop item type, which averaged 1.2 minutes to complete, received the highest ratings for student independence and *effortless* completion. Engagement remained high (83% rated *very engaged*) because the task was quick, and feedback was immediate. Conversely, as the time required to complete an item increased, students' ability to attend the task decreased. During the simulation items (averaging 10.4 minutes to complete), students became visibly tired; in one instance, a student physically left the testing session.

The study confirms that extended time on task disproportionately affects students who cannot sustain focus for long periods. Additionally, longer administration times adversely affect students with motor challenges, as sustaining physical effort (e.g., clicking, dragging) leads to physical fatigue. Long items also often require more interactions with the test administrator, increasing the likelihood of interrupting the test administration, adding to the time burden to the student. Here are some strategic considerations for time management.

- Initial Learning Curve
    - New item types inevitably take longer because of unfamiliarity, which lowers student confidence and engagement. Providing practice items and clear, chunked instructions can mitigate this effect. During the study, test administrators noted the importance of more practice opportunities for all the item types; therefore, it is vital to make practice items available and consider how to integrate these item types into regular instruction.
- Pacing and Complexity
    - The population of learners with significant cognitive disabilities is uniquely heterogeneous. Longer items disproportionately affect students with specific attention, communication, or motor challenges. To preserve pacing and accommodate this range of learners, DLM science testlets should target an average duration of approximately three minutes (i.e., less than one minute per item).
- Design Strategy
    - If a TEI regularly exceeds target times (as seen with simulation items), developers should reevaluate the design. Alternatives include splitting tasks into smaller steps, adding scaffolds, or determining whether a different item type would better measure the construct while maintaining engagement.

Time costs are a key factor when introducing new items. DLM science assessments are delivered one testlet at a time (i.e., nine testlets per assessment), whereas this study presented only four items. States prioritize minimizing testing times, and DLM should prioritize low administration times across the collection of testlets. Future designs must balance the cognition being measured with the practical need to keep task time low and engagement high. Obtaining timing data for students across various complexity bands is a critical component of decision-making to ensure that new TEI formats do not inadvertently introduce construct-irrelevant barriers.

CONCLUSION AND SIGNIFICANCE

This study examined the potential for using TEIs to measure science constructs for students with significant cognitive disabilities. The findings suggest that item format strongly affects student performance. The findings reveal a clear divergence in the viability of the formats tested. TEIs were most effective when the item type aligned with existing instructional practices and used digital tools that were already familiar to students. TEIs that used unfamiliar interfaces or imposed heavy cognitive demands risked conflating technology skills with science understanding. The use of any TEIs must address the identified accessibility challenges.

VIABLE ITEM FORMATS

Three item types—drag and drop, hot spot, and drop-down—demonstrated strong potential for operational use.

- Drag and Drop
    - This type was the most promising item format, proving effective for students across Complexity Bands 1, 2, and 3. Students engaged with these items independently or with minimal support. Because of its intuitive nature, drag and drop is promising for all grade levels and was particularly effective for younger students (i.e., elementary) compared to other item types.
- Hot Spot
    - This type was highly successful for students in Complexity Bands 2 and 3, who completed items independently or with allowable supports. Although students in Complexity Band 1 required modeling for initial items. The touch-to-select mechanism proved intuitive across all bands, mirroring the visual instructional activities (e.g., pointing, touching) used at all grade levels for this population.
- Drop-Down
    - This type was successful for students in Complexity Bands 2 and 3 but presented specific design challenges. The navigation of the drop-down menu

arrow proved difficult for some students. Additionally, regardless of complexity band, all students relied heavily on test administrators to read the text-only options, suggesting that the text-heavy design creates a dependency on support. Consequently, this item type was recommended by test administrators for introduction at grade 4 or above.

CHALLENGING ITEM FORMATS

Two item types—table match and simulation—presented significant barriers and would be challenging to use on an operational test in their current form. Limited prior exposure to these formats led to uncertainty for both students and administrators, resulting in longer administration times, strained engagement, and a frequent need for intensive administrator support (e.g., navigation, reading, step-by-step modeling).

- Table Match (Matrix)
    - Although students in Complexity Bands 2 and 3 were generally successful, it required significant effort to complete these items. The difficulty stemmed largely from the format rather than the science content.
        - Cognitive Load
            - The format imposes a high, extraneous cognitive load caused by the split attention effect. To respond to an item, students must track a row header, look across column headers, and find the intersection point. This task taxed short-term memory and often caused students to get lost in the grid.
        - Administration Variability
            - Performance partially depended on the administrator's style rather than the student's knowledge. Students were notably more successful when administrators provided a high level of support (e.g., rereading headers for every row, physically pointing to track options). Without this visual anchoring, students often reverted to defaulting behaviors, such as selecting the first button in every row.
        - Physical Barriers
            - The extremely small radio buttons posed a barrier to independence, often forcing administrators to navigate the cursor or click on behalf of the student.
- Simulation
    - This item type presented the most significant challenges. Students at Complexity Band 3 had limited success, and no student in any band demonstrated true independence with this format. Although students

enjoyed the visuals, the cognitive and navigational demands were too high for them to demonstrate science knowledge without significant adult intervention.

- Validity Concerns
  - Unless the design is significantly simplified, these items risk measuring a student's ability to navigate a complex interface instead of their science understanding.
- Cognitive and Technical Barriers
  - Multistep directions limited student ability to access the item, causing students to forget the task while navigating. Technical issues, such as video-loading delays and rendering errors, introduced randomness into the testing process, breaking student focus. Furthermore, content rendering varied across devices (e.g., requiring scrolling on some screens but not others), introducing construct-irrelevant variance.
- Potential Use
  - Currently, this format appears too difficult for the elementary and middle school students tested. However, it may hold promise for high school students, although further research with that specific population is needed.

SIGNIFICANCE

This study contributes to the field of large-scale assessment by providing validity evidence for the use of science TEIs with students with significant cognitive disabilities. This study provides evidence to differentiate between item types that are candidates for operational use and types that are currently non-viable without further refinement. Specifically, drag-and-drop TEIs show immediate utility for measuring the intended science constructs across all complexity bands. Both hot-spot and drop-down show promise complexity bands two and three once the accessibility challenges are addressed. Drag and drop, hot spot and drop-down formats successfully align with students' instructional experiences and allow for the assessment of multidimensional science skills without introducing significant barriers. Conversely, table-match and simulation formats, in their current iterations, present cognitive and navigational barriers that introduce construct-irrelevant variance by conflating science knowledge with the ability to navigate complex digital interfaces.

The findings offer important contributions to research and practice. This study emphasizes the critical balance between technological innovation and accessibility. The research highlights specific design requirements necessary (e.g., screen real estate, touch-target

size, managing cognitive load) to make sure TEIs remain barrier-free. The results confirm that future science assessments can measure multidimensional science skills rather than students' ability to navigate complex technology. We found that students with significant cognitive disabilities can interact meaningfully with TEIs, but only when design choices minimize navigation load, provide clear visual cues, and ensure consistent rendering. Moving forward, the thoughtful selection of TEIs and adherence to accessibility guidelines will be essential to ensuring valid, barrier-free measurement for this population. Ultimately, the successful integration of TEIs into DLM science assessments will depend on balancing the need for innovative measurement with the necessity of an accessible, barrier-free user experience.

This study addresses a significant gap in the literature by finding TEIs can be used to measure multidimensional science constructs for students with significant cognitive disabilities. Study findings challenge decades-old assumptions that students with significant cognitive disabilities lack the digital literacy required for complex assessment items. The findings demonstrate that the increased use of touch-based devices in daily instruction has supported these students in developing the digital fluency necessary to navigate TEIs independently. Specifically, this study found that TEIs mimicking familiar touch-and-move and touch-to-select mechanics (e.g., drag and drop, hot spot) allow students to successfully demonstrate their science knowledge. This evidence suggests that when assessments align with students' existing digital experiences, construct-irrelevant barriers are minimized.

REFERENCES

Braun, M., Menschik, C., Wahl, V., Etges, T., Löwe, L. D., Wölfel, M., Kiuppis, F., Kunze, C., & Renner, G. (2025). Current digital consumer technology: Barriers, facilitators, and impact on participation for persons with intellectual disabilities – a scoping review. *Disability and Rehabilitation*, 1–22. https://doi-org.www2.lib.ku.edu/10.1080/09638288.2025.2471567

Bryant, W. (2017). Developing a strategy for using technology-enhanced items in large-scale standardized tests. *Practical Assessment, Research, and Evaluation, 22*(1). https://doi.org/10.7275/70yb-dj34

Burnes, J. J., & Clark, A. K. (2021). Characteristics of students who take Dynamic Learning Maps® alternate assessments: 2018–2019 (Technical Report No. 20-01). University of Kansas, Accessible Teaching, Learning, and Assessment Systems (ATLAS). https://doi.org/10.1111/bld.12501

Dynamic Learning Maps Consortium. (2016). *2014–2015 technical manual—year end*. University of Kansas, Center for Educational Testing and Evaluation. https://dynamiclearningmaps.org/sites/default/files/documents/publication/Technical_Manual_YE_2014-15.pdf

Dynamic Learning Maps Consortium. (2023). *2022–2023 technical manual update—science*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems. https://2023-sci-techmanual.dynamiclearningmaps.org/

Gunderson, J. L., Higgins, K., Morgan, J. J., Tandy, R., & Brown, M. R. (2017). Cognitively accessible academic lessons for students with intellectual disabilities using the iPad. *Journal of Special Education Technology, 32*(4), 187–198. https://doi-org.www2.lib.ku.edu/10.1177/0162643417715750

Holyfield, C., Zimmerman, T. O. N., MacNeil, S., Caldwell, N. S., Patel, P., Griffen, B., & Vucetic, S. (2024). Preliminary investigation of context-aware AAC with automated just-in-time cloze phrase response options for social participation from children on the autism spectrum. *Folia Phoniatrica et Logopaedica*, 1–21. https://doi.org/10.1159/000542304

Karvonen, M., Swinburne Romine, R., & Clark, A. (2024). Response process evidence for academic assessments of students with significant cognitive disabilities. *Practical Assessment, Research, and Evaluation*, *29*(13): 1–15. https://doi.org/10.7275/pare.2060

Kellems, R. O., Rickard, T. H., Okray, D. A., Sauer-Sagiv, L., & Washburn, B. (2017). iPad® Video Prompting to Teach Young Adults With Disabilities Independent Living Skills: A

Maintenance Study. *Career Development and Transition for Exceptional Individuals*, *41*(3), 175-184. https://doi-org.www2.lib.ku.edu/10.1177/2165143417719078 (Original work published 2018)

Leighton, J. P. (2017). Using think-aloud interviews and cognitive labs in educational research. Oxford University Press.

Marion, S. F., & Pellegrino, J. W. (2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice, 25*(4), 47–57. https://doi.org/10.1111/j.1745- 3992.2006.00078.x

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press. https://doi.org/10.17226/13165

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. The National Academies Press.

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. (2010*). Innovative items for computerized testing*. In W. J. van der Linden & C. A. W. Glas (Eds.), Elements of Adaptive Testing (2nd ed.). New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-85461-8_11

Tiemann, G., & Karvonen, M. (2019). *Evaluating innovative assessments: Evidence from I-SMART cognitive labs*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems. https://ismart.works/sites/default/files/documents/Publications/I-SMART_Goal_2_Cognitive_%20Lab_Report_FINAL.pdf

World Wide Web Consortium. (2018). *Web content accessibility guidelines (WCAG) 2.1*. https://www.isakssw3.org/TR/WCAG21/

APPENDIX A. TEST-ADMINISTRATOR CHARACTERISTICS

Table 8

Test-Administrator Characteristics

Characteristic	West Virginia	New York	Total
Gender			
Female	6	4	10
Male	2	2	4
Ethnicity			
Hispanic/Latino	0	2	2
Non-Hispanic	6	4	10
Choose not to disclose	2	0	2
Race			
Black or African American	0	1	1
White	7	3	10
Choose not to disclose	1	2	3
Years of experience administering DLM science assessments			
0–1	3	3	6
2–3	1	0	1
4–5	0	1	1
More than 5	4	2	6
Years of experience in pre-K–12 science			
0–1	0	0	0
2–5	3	2	5
6–10	1	2	3
11–20	3	1	4
> 20	1	1	2
Years of experience working with students with significant cognitive disabilities			
0–1	0	0	0
2–5	3	3	6
6–10	1	1	2
11–20	4	1	5
> 20	0	1	1

APPENDIX B
STUDENT COMPLEXITY BANDS: EXPRESSIVE-COMMUNICATION AND SCIENCE BANDS

Table 9

Expressive-Communication and Science Bands of Students, by State

Complexity Band	State A	State B	Total
Expressive-communication band			
Foundational	2	0	2
Complexity Band 1	1	5	6
Complexity Band 2	3	4	7
Complexity Band 3	8	5	13
Science band			
Foundational	1	3	4
Complexity Band 1	1	7	8
Complexity Band 2	8	3	11
Complexity Band 3	4	1	5

Table 10

Number of Students Taking Each Item Type, by Expressive-Communication Band

Item Type	Expressive-Communication Band			
	Foundational	Complexity Band 1	Complexity Band 2	Complexity Band 3
Drag and drop	2	3	4	3
Drop-down	1	1	4	6
Hot spot	1	5	2	3
Table match	0	1	1	9
Simulation	0	2	3	5

Table 11

Number of Students Taking Each Item Type, by Science Band

Item Type	Science Band			
	Foundational	Complexity Band 1	Complexity Band 2	Complexity Band 3
Drag and drop	2	4	5	1
Drop-down	1	2	6	3
Hot spot	2	4	3	2
Table match	1	2	5	3
Simulation	2	4	3	1

APPENDIX C: STUDENT ACCESSIBILITY SUPPORTS, DEVICES, AND EXPRESSIVE-COMMUNICATION MODE

Table 12

Accessibility Supports, Devices, and Expressive-Communication Mode, by Item Type

Category	Drag and Drop (*N* = 12)		Drop-Down (*N* = 12)		Hot Spot (*N* = 11)		Table Match (*N* = 11)		Simulation (*N* = 10)	
	n	%	*n*	%	*n*	%	*n*	%	n	%
Accessibility Supports										
Magnification	2	17	2	17	1	9	2	18	3	30
Human read aloud	11	92	10	83	10	91	9	82	10	100
Device Type										
Laptop	1	8	2	17	1	9	3	27	5	50
iPad	11	92	10	83	10	91	8	73	5	50
Observed Student Expressive-Communication Mode [a]										
Touch pad	0	0	1	8	0	0	1	9	0	0
Touch screen	11	92	10	83	9	82	8	73	5	50
Speech	9	75	12	100	6	55	10	91	8	80
AAC	1	8	0	0	3	27	2	18	3	30
Gestures	8	67	5	42	6	55	5	45	5	50
Other	1	8	0	0	0	0	1	9	1	10

Note. [a] These modes are not mutually exclusive. AAC = augmentative and alternative communication device.

Table 13

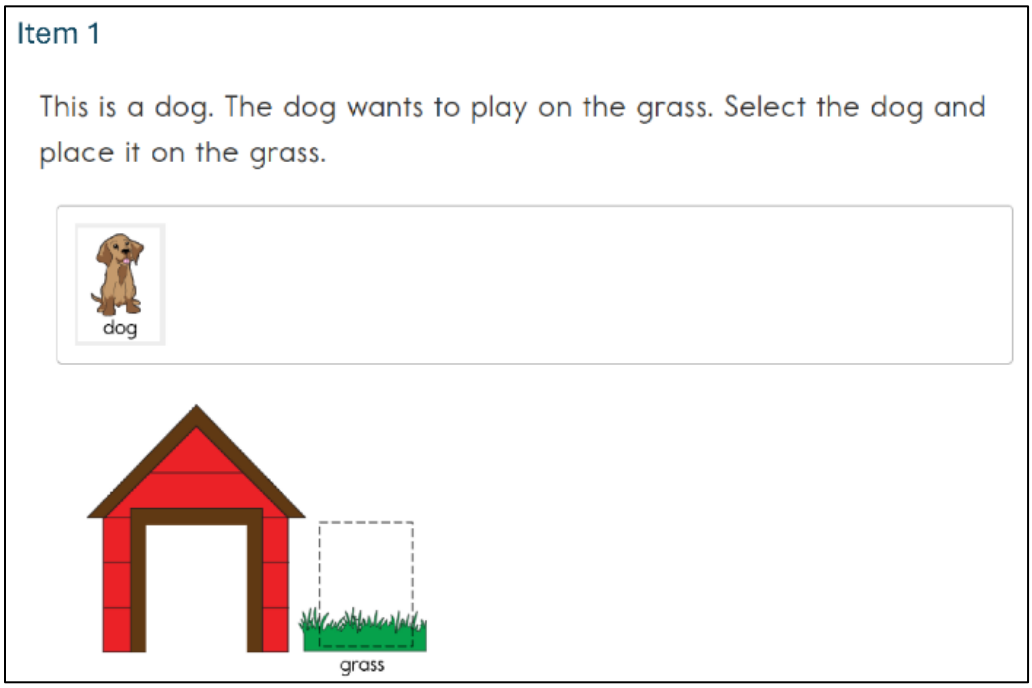
Number of Students With Specific Supports Indicated for Their Kite Student Portal and DLM Assessments

Support Category	Specific Support	Students (*N*)
Display	*Magnification*	3
Display	Overlay Color	2
Display	Contrast Color	2
Auditory	Spoken Audio	23
Auditory	*Switches*	1*
Environmental/other	Alternate Form—Visual Impairment	1*
Environmental/other	Individualized Manipulatives	12
Environmental/other	Calculator	3
Environmental/other	*Human read aloud*	21
Environmental/other	*Administrator enters response*	15
Environmental/other	Partner assisted scanning	3

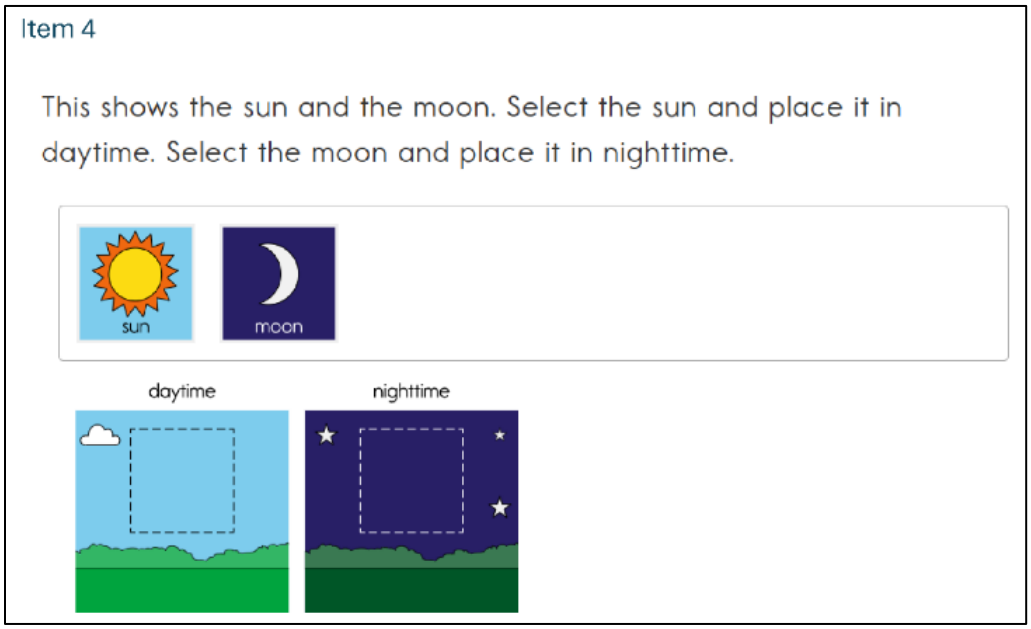
Note. The four italicized supports were provided during the usability study. Despite the alternate form identified on the Personal Needs and Preferences survey, one student was able to participate without a blind/visually impaired form. The test administrator said the student's survey had been filled out previously and that it was not accurate in the system. The student who had switches enabled used an augmentative and alternative communication device to support responses, and a switch was not needed.

APPENDIX D: EXAMPLE ITEMS

Example of Content-Neutral Drag-and-Drop Item



Example of Science-Light Drag-and-Drop Item



Example of Content-Neutral Drop-Down Item

Item 1

A student has school supplies. The student has a pencil, a marker, and a pen.

The sentence is about the student's school supplies. Choose a word to finish the sentence.

The student has a pencil, a marker, and a [ ].

Example of Science-Light Drop-Down Item

Item 4

This is a plant. The plant has 5 leaves. The leaves are small and green.

The table is about the plant leaves. Choose a number or word to finish the table.

Plant Leaves

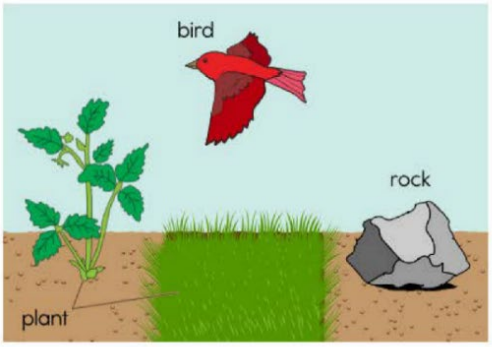
1. Number of leaves	[ ]
2. Size of leaves	[ ]
3. Color of leaves	[ ]

Example of Content-Neutral Hot-Spot Item

The picture shows a room. The room has a table, books, and two chairs. Select the **two** chairs.

Example of Science-Light Hot-Spot Item

The picture shows an outdoor area. The area has two plants, a bird, and a rock. Select the **two** plants.



Example of Content-Neutral Table-Match Item

Look at the pictures of different rooms. Are the rooms in a home or a school? Choose **Home** or **School** for each room.

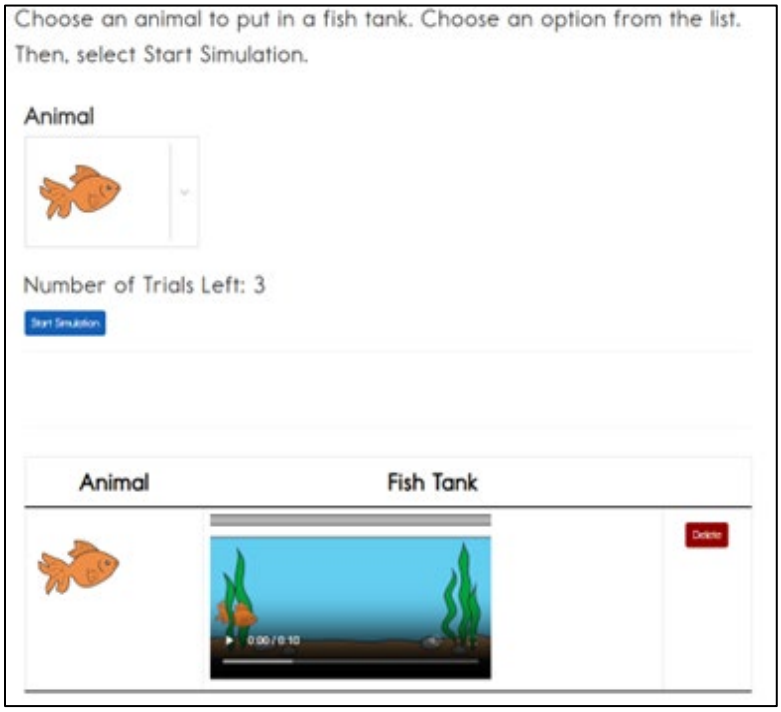
	Home	School
 classroom	○	○
 gym	○	○
 bedroom	○	○

Example of Science-Light Table-Match Item

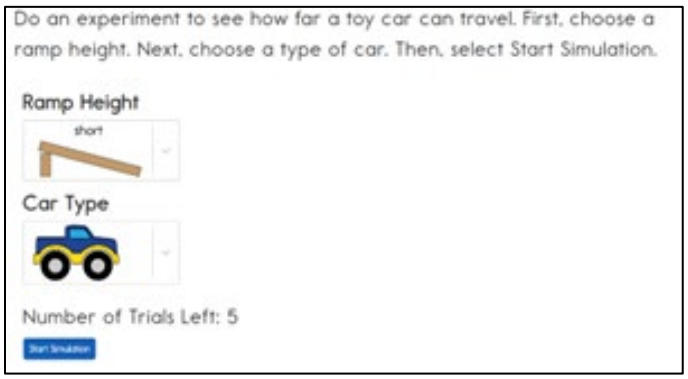
Look at the pictures. Do the pictures show an animal or not an animal? Choose **Animal** or **Not an Animal** for each picture.

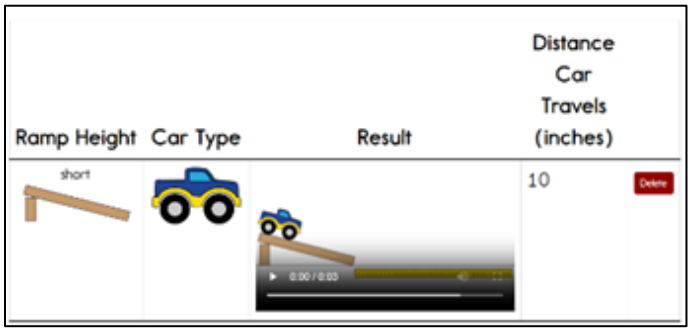
	Animal	Not an Animal
 dog	○	○
 duck	○	○
 rock	○	○

Example of Content-Neutral Simulation Item



Example of Science-Light Simulation Item





APPENDIX E: SUMMARY OF INTERVIEW SESSION CODING, BY TEI TYPE

Table 14

Drag-and-Drop Interview Summary

Category	Subdomain	Findings
Accessibility	Accessibility Supports	• Two teachers thought drag-and-drop items might not be accessible for students with fine-motor and visual impairments without TA support. This comment was unrelated to a study participant. • One teacher mentioned that drag-and-drop supports responses for students who use gestures (e.g., pointing).
	Technology	• One teacher mentioned the impact of scrolling on a student's response: *"They have to scroll down to find the next button. It stumbles them up every time, you know where it's kind of clunky, in fact, that it doesn't fit all on one screen."* • Left-justified content makes item content smaller than needed rather than using full screen real estate. • One teacher recognized that drag-and-drop might not be accessible for all students if the item clicks back to start or they cannot drag it all the way; this was not observed but was a concern.
	Navigation	• Four teachers felt that the drag-and-drop items were accessible for all students. Students could navigate these items independently or with allowable supports. • Quote: "I think it is really great for the lowest level of kids.... The drag and drop seem to be right at their level.... I think the functionality of it worked very well for them, but the items that they were asked were, you know, far too easy for some of my kids."

Presentation	Display of Information	• One teacher thought the images were too small.
	Task Design	N/A
	Language	N/A
	Test Administration	N/A
Engagement	Effort	• Most students needed little effort to interact with drag-and-drop items.
	Tasks Optimize Independence	• Five teachers felt the drag-and-drop items optimized student independence and did not think other students would have any difficulties with this item type. For example, • "He was ahead of me at one point. He was like, 'Oh yeah, I know where these pictures go.'"
	Student Satisfaction	• Students enjoyed the item type.
	Task Relevance	• Seven teachers indicated that students were familiar with the drag-and-drop functionality and used it in instructional programs, games, and other informal uses.
	Complexity	• Quote: "The drag and drop, for my lower-level kids, is fantastic."

Table 15

Drop-Down Interview Summary

Category	Subdomain	Findings
Accessibility	Accessibility Supports	• Students could interact independently, with allowable test-administrator (TA) supports and practice. • One student would need TA to manipulate item.
	Technology	• The use of magnification affected the item presentation and covered up a sentence. • Arrow for drop-down was too small. • When the drop-down menu expanded, it blocked some of the sentence in the prompt and students could not see all content on the screen (using magnification). • Response not detected; sometimes students needed to click multiple times for their response to be accepted. • Screen needed to be sensitive enough to detect response; student needed to click multiple times for response to be accepted.
	Navigation	• Quote: "Students can click anywhere in the box (not just the down arrow), it will be easier for Ss [students] (especially those with difficulties with fine-motor control)." • Other TAs did not know this, and most students had to click on the arrow before making a choice.
Presentation	Display of Information	• Content was left-justified, leaving unused white space. • Content needed to be expanded to take up the whole screen. • Pictures were too small. • Pictures needed contrast for students with visual impairment.

Category	Subdomain	Findings
		• TA suggested highlighting around the drop-down box to make it stand out, so students knew where they had to click.
	Task Design	• Drop-down text had two words, an appropriate number for some students. More words might have made it harder. • TA liked having two response options but thought three might be better for some of her students (two might be too easy). • Use of words rather than pictures was more complex, but students could engage with TA support. • Task design may lead to students always selecting the second option as it is the last text they hear. • Quote: "It just so happened the first student that did it can't read. So, she was able to do it." • Drop-down format will be "fine for students because it's very similar to multiple choice."
	Language	• This item type reduced the number of words to select, which helped students with difficulty reading longer response options.
	Test Administration	• Some TAs read the entire prompt with each choice separately (e.g., presenting option 1 separately and then presenting prompt and option 2). • Other TAs read the two answer choices consecutively (option 1 or 2) without repeating the prompt.
Engagement	Effort	• Students responded with little effort. • TA thought this was easy for the student.
	Tasks Optimize Independence	• Students could respond independently after practice. • Students could read some words.
	Student Satisfaction	• Students knew what to do and stayed attentive.
	Task Relevance	• Teacher used task type during instruction or assessment.

Category	Subdomain	Findings
		• Students had prior experience using drop-down type items (i-Ready). • Quote: "The drop-down…[is] not [in i-Ready] until like 4th and 5th grade, and I don't have those students as often, but even then, it's not used as much."
	Complexity	• Content represented at appropriate complexity. • Quote: "For the higher level [students], I feel like the drop-down is probably more suitable for them."

Table 16

Hot-Spot Interview Summary

Category	Subdomain	Findings
Accessibility	Accessibility supports	• Nearly all teachers who saw this item believed their students could interact independently with hot-spot items with allowable test-administrator (TA) supports and practice.
	Technology	• One teacher noted that when students wanted to change a response, they first had to unselect it before clicking another item. Students did not know they had to unselect the first choice before making another selection. This confused some students; they could not respond because they did not know how to select the desired response. This kind of issue can add to students' cognitive load. We saw evidence of students giving up when this occurred. • One teacher noted scrolling effects; another noted small font when students took testlets on an iPad.

Category	Subdomain	Findings
	Navigation	• Most teachers indicated their students knew what to do and could respond independently. • One teacher noted possible fatigue caused by physical effort when the student used a mouse to respond. One student needed prompting to select two items; teacher was not sure if the student could progress independently when he selected only one item.
Presentation	Display of information	• One teacher liked the fact that the object turned pink when selected so that students knew when they had selected an object. • One teacher noted the small font size, while another thought the images needed to be larger. • White space and left-justified placement on the screen
	Task Design	• Several students had difficulty with the item asking them to select two plants. Some did not notice the word "plant" and the lines that pointed to the plant; others think that students did not know that grass was a plant. • One teacher felt that the change in directions caused difficulty; in some hot-spot items, students were asked to select one thing and in others they were asked to select two things.
	Language	N/A
	Test Administration	N/A
Engagement	Effort	• One teacher indicated that the hot-spot items were easy for her student.
	Tasks Optimize Independence	• Three teachers believed that their students would be able to respond to the hot-spot items independently with allowable supports and practice.

Category	Subdomain	Findings
		• Quote: "All he had to do was…look at the picture and identify what they were asking him, [identify] the two sorts of plants, so he was able to tap the screen, which didn't really require any…fine motor skills…. So, he did well with that one…."…"
	Student Satisfaction	N/A
	Task Relevance	• Four teachers said their students used hot-spot functionality in other instructional activities.
	Complexity	N/A

Table 17

Table-Match Interview Summary

Category	Subdomain	Findings
Accessibility	Accessibility Supports	• Test administrator (TA) did not need to enter all responses even though it is in the PNP. • Magnification: TA had to help navigate as all content did not fit on one screen.
	Technology	• Item response would be easier for students if the area for the response were increased (i.e., anywhere in the box, not just the little circle); to support students with difficulty in selection due to fine-motor skills. • The technology was not sensitive enough to accept the student's response, so he had to try four times. This student persisted, but other students may not. Scrolling effects • Impact of screen size

Category	Subdomain	Findings
		• Quote: “If it was all on one screen, …I feel like when I was moving it around a lot…it…was visually overwhelming for them….” • Using a laptop with touch pad, a student needed to make a lot of effort to click on the radio buttons in the table cells.
	Navigation	• Using a laptop with touch pad took a lot of effort to click on the radio buttons in the table cells. • Students could not respond independently; it was “hard for him to follow what to do.” • TA prompts and reminders were needed for students to complete each step of the item. • Column labels in the table show categories only once in the top row; as more rows were added, students had difficulty remembering which column was associated with which category. • TA had to support students to respond correctly to the table-match format (i.e., marking their response in the appropriate row). • Student could not respond independently; it was “hard for him to follow what to do.”
Presentation	Display of Information	• Table format led to students following a pattern after the second example: “She just followed the pattern.” [Tables with more than two items, animal/not an animal] • Buttons in middle and fewer columns would make it easier to respond. • Quote: “If we have such a large space in this table, why is that circle so [far] in the corner?”
	Task Design	• Two categories and two objects maximum

Category	Subdomain	Findings
		• TA believed sorting skills would be better assessed with a different format (e.g., drag and drop, which is more familiar). • Students did not understand what was being asked of them. • TA did not think item type was appropriate for all students; it may have encouraged students to select all responses in the same column. • TAs wanted clear graphics combining pictures and words.
	Language	N/A
	Test administration	• Unclear directions • TAs read each category once and then presented the options. Two TAs read the two-column categories aloud each time an item was presented, which supported students' ability to respond. It would be helpful to include this option (i.e., repeat the column categories each time you present an item) in the instructions as an allowable practice. • A student clicked in the header a few times to respond rather than inside the cells.
Engagement	Effort	• This item type required effort and time to respond.
	Tasks optimize independence	• This item type is appropriate for some students who respond independently.
	Student satisfaction	• One student was confident in her response.
	Task relevance	• The table format was novel, and students were not familiar with responding to this item type. • Table was not the best format for categorizing; students had no experience with that format. • Difficulty reading tables and charts

Category	Subdomain	Findings
		• Drag-and-drop items (i.e., sorting) would have been better because students are used to them (used in i-Ready).
	Complexity	• Quote: "For the higher level [students], the charts [i.e., table match] [are] probably more suitable for them."

Table 18

Simulation Interview Summary

Category	Subdomain	Findings
Accessibility	Accessibility supports	• Test administrator (TA) had to navigate for students, as they did not know they had to scroll down to see the table with the simulation results.
	Technology	• Magnification and increased scrolling • The entire item was not visible on the screen. • Video had to be enlarged to be visible. • Quote: "*I feel like the scrolling is difficult. And then maybe, like the video being a little bit bigger or like the images being a little bit bigger in the video. Yeah, I just like the measuring [item], I feel you couldn't really see what the number was. I know it had it next to it. But like things like that.*"
	Navigation	• TA had to prompt and model each step. • Drop-down arrow was small and hard to select. • The video was difficult to play (not automatic, needed to be enlarged). • The videos were small and difficult to manipulate or expand, especially on iPads. • The simulation item type rendered differently across operating devices; researchers were unable to discern the nature of the video not working consistently. We were not sure whether it was because of bandwidth, device

Category		Subdomain	Findings
		effects, or that the size of the video file was too big to play effectively. The technology team may need to determine the minimum and maximum length or size of video files for simulation items.	
Presentation engagement	Display of information	• Text was too small to be legible in the item.	
	Task design	• TA did not think her students understood what they were supposed to do (the purpose of the item). • TA was uncertain if all students would be able to follow all the steps in the simulation item types.. • Having multistep directions for an item was challenging; it was better to break it up into separate steps for each page. • There is no way for test administrators to "*monitor progress during testing"*: for example, change color of arrows and buttons (or make them glow or blink) to make them stand out and prompt students so they do not forget to click it to progress through the item; or order the steps numerically with check boxes).	
	Language	• Students were not familiar with the word "simulation."	
	Test administration	• The items did not have a question to answer, so the test administrator and students were not sure what to do. • Quote: "I felt like if there was a question with those with the simulations, maybe it could have like more accurately measured the skill it was asking." • TAs need guidance on what is expected for the item-type response.	
	Effort	• Quote: "I love the fact that he enjoyed it, and you could see...,,it kept him engaged... ,because he kept wanting to do it."	
	Tasks optimize independence	• Students needed TA support to navigate the directions.	

Category		Subdomain	Findings
	Student satisfaction	• Quote: “He definitely enjoyed playing with it, so that to me would be the next step is understanding, getting him to understand, like, how the distance change based on the ramp….”	
	Task relevance	• Students did not use similar item types in other instructional activities. • TA felt the item type was meaningful and instructionally relevant.	
	Complexity	• Some students could not understand what they were being asked to do. • Quote: “*All my kids weren’t verbal, so I wouldn’t feel like explaining it to them that way would make it make sense for them, …even giving them fast and slow, and it doing it and say, did the car go fast and slow? They would most likely pick the 1st answer that they see, because there is no real understanding of what exactly transpired.*”	

APPENDIX F: TEST ADMINISTRATOR RATINGS OF ITEM EFFECTIVENESS ACROSS COMPLEXITY BANDS

Table 19

Test Administrators' Rating of Item Functioning, by Item Type and Students' Final Science-Complexity Band

Item Type	Item Functioning for the Student Was ***Easy*, *Hard*, or *About Right***		
	Easy	*About right*	*Hard*
Complexity Bands 2 and 3			
Drag and drop (*N* = 3)	2	1	0
Drop-down (*N* = 8)	3	4	1
Hot spot (*N* = 4)	3	1	0
Table match (*N* = 7)	2	4	2
Simulation (*N* = 2)	0	1	1
Foundational Band and Complexity Band 1			
Drag and drop (*N* = 7)	2	4	1
Drop-down (*N* = 3)	1	2	0
Hot spot (*N* = 5)*	2	1	1
Table match (*N* = 2)	1	0	1
Simulation (*N* = 7)	1	3	3
All Bands			
Drag and drop (*N* = 10)	4	5	1
Drop-down (*N* = 11)	4	6	1
Hot spot (*N* = 9) [a]	5	2	1
Table match (*N* = 9)	3	4	2
Simulation (*N* = 9)	1	4	4

Note. [a] Test administrator did not respond to the prompt for this item type for one student.

Table 20

Test Administrator's Rating of Item Effectiveness, by Item Type and Students' Final Science-Complexity Band

Item Type	Effectiveness of Item Type in Measuring What the Student Knows			
	Not at all effective	*Somewhat not effective*	*Somewhat effective*	*Very effective*
Complexity Bands 2 and 3				
Drag and drop (*N* = 3)	0	0	0	3
Drop-down (*N* = 8)	0	0	3	5
Hot spot (*N* = 4)	1	0	2	1
Table match (*N* = 7)	1	1	3	2
Simulation (*N* = 2)	0	1	0	1
Foundational Band and Complexity Band 1				
Drag and drop (*N* = 7)	1	0	2	4
Drop-down (*N* = 3)*	0	1	0	1
Hot spot (*N* = 5)*	1	0	0	3
Table match (*N* = 2)	0	0	1	1
Simulation (*N* = 7)	0	2	3	2
All Bands				
Drag and drop (*N* = 10)	1	0	2	7
Drop-down (*N* = 11)*	0	1	3	6
Hot spot (*N* = 9) [a]	2	0	2	4
Table match (*N* = 9)	1	1	4	3
Simulation (*N* = 9)	0	3	3	3

Note. [a] Test administrator did not respond to the prompt for this item type for one student.

APPENDIX G: STRENGTHS, PROMISES, AND CHALLENGES OF TEI TYPES

Table 21

Summary of Item Types' Strengths, Challenges, and Ratings of Promise

Strengths	Challenges	Rating: Promising or Challenging
Drag and drop		
• Effortless student response • Students were familiar with item type via informal and instructional use. • All students in Complexity Bands 2 & 3 completed all items independently or with allowable supports, which supports use at Proximal Precursor and Target linkage levels and may support use at Distal Precursor linkage level. • Shortest time to administer of all item types • Picture/text supported	• Left-justified items • Images and text should be larger.	Promising • Students knew how to intuitively use the item type. • Students use this type of item in a variety of instruction and informal activities and can demonstrate knowledge of skills and understandings for a wide range of applications (e.g., sorting, categorization, patterns, models). • Student responses were intuitive, effortless, and consistent with informal and instructional activities. • Most-efficient time to administer (1.2 minutes) • Minimal test-administrator burden Accessibility • Allowed responses for students who use gestures. • Left-justified content makes item content smaller than needed rather than using full screen real estate. Presentation • Increase size of image or text.

Strengths	Challenges	Rating: Promising or Challenging
		Engagement • Drag-and-drop items optimized student independence and enjoyment
Drop-down		
• Some students knew what to do. • Students are familiar with instructional use of item type (cloze tasks). • All but one student in Complexity Bands 2 & 3 completed all items independently or with allowable supports, which supports use at Proximal Precursor and Target linkage levels. • Some picture/text supported.	• Tryout items contained a lot of text. • Time to administer was longer and depended on the student reading or having the items read. • Interaction of drop-down hid some item content. • Small arrow hard to select. • Increased scrolling • Test administrators believe students use listening as a strategy to make responses.	Promising • The time to administer the items was efficient (3.3 minutes). • The test-administrator burden was moderate, mostly related to navigation of the item type. Accessibility • The use of magnification affected item presentation and increased scrolling made it difficult for students to see all content, increasing the cognitive load. Test administrators supported navigation when magnification was enabled. • Item content should be rendered on one screen for students to understand or respond independently. • The images need to be enlarged, and the response area increased. • The drop-down arrow should be enlarged. • The item response required clicking and often had to be multiple clicks to accept the response.

Strengths	Challenges	Rating: Promising or Challenging
		Presentation • Test administrators felt this was more appropriate for "higher level students."
Hot Spot		
• Students knew what to do. • Nearly effortless • Familiar and intuitive picture selection including AAC users • All students in complexity bands 2 & 3 completed all items independently or with allowable support and offers support for use at Proximal Precursor and Target linkage levels. • Minimal test-administration time • Picture/text supported • Highlighted responses • Nearly all teachers believed their students would be able to interact independently or with allowable supports.	• The process of changing responses is not intuitive. • Spatial layout due to left justification • Need to enlarge images. • Very image dependent on presentation and response	Promising • Students knew how to intuitively use the item type. • Students could respond independently. • Students could respond without reading. • Task relevant: This kind of response is used in instructional activities and on other devices. • Nearly effortless for student response • The time to administer was very efficient (2.5 minutes). • The test-administrator burden was minimal. Accessibility • The entire object should be clickable and have physical space for fine motor. • The source of challenge for changing the responses must be addressed if this item type is used. • Increase the size of images and text. • Left justification of items is an inefficient use of space. • Consider the effects of scrolling across devices.

Strengths	Challenges	Rating: Promising or Challenging
		• Consider differences in usability based on response differences between using a touch screen versus mouse, mouse pad, switch.
Table Match		
• Students were familiar with and liked the images.	• Students were unfamiliar with this type of item. • Students had difficulty understanding the item functioning. • Test administrators support was needed for directions, scrolling, and read aloud. • Students had difficulty clicking on the radio buttons in the table cells. • Student who did not know what to do sought to identify a pattern (e.g., selecting all the buttons in a row). • The length of the item caused increased scrolling. • All the item content was not on one page, which obscured the stem when responding.	Challenging • Students did not understand how to use this item type. • Students sought visual patterns to select responses. • The time to administer this item type was moderate. • The test-administrator burden was significant. Accessibility • The item response would be easier for students if the area for the response were increased (i.e., anywhere in the box, not just the little circle). • Item compatibility for laptop versus iPad users (scrolling and size of the radio button). Presentation • Students could not recall the category after the second row. If the purpose of the item is to sort or categorize, consider whether this is the most efficient item type, or another type, such as drag and drop, is more efficient.

Strengths	Challenges	Rating: Promising or Challenging
		Engagement • This item type required more student and test-administrator effort and time to respond (3.5 minutes).
Simulation		
• Students liked the videos. • Test administrators felt the item type was instructionally relevant for science.	• Students and test administrators did not understand what they were supposed to do. • Students were unfamiliar with this item type. • Students could not respond without significant interaction with test administrators (directions, navigation). • Language complexity (terminology e.g., start simulation) • Longest time to administer of all item types • Greatest test-administrator burden of all item types	Challenging • Students did not know how to use this type of item. • Test administrators had difficulty administering items. • Accessibility challenges: small images and text, video display. Presentation challenges • The item-stem expectations need clarification. • Multistep directions and too much content per page caused scrolling and added to the cognitive load required to complete the items. Engagement • Students enjoyed playing with the content (video and image simulation). • The time to administer this item type was the longest of all item types (10.4 minutes). • The test-administrator burden was significant to support students in completing this item type.

APPENDIX H: RECOMMENDATIONS FOR IMPROVING ACCESSIBILITY OF ITEM TYPES

DRAG-AND-DROP ITEMS

The drag-and-drop item type was one of the most promising formats evaluated. Drag-and-drop items offer a valid, efficient, and accessible way to measure science knowledge for most of the DLM population (Complexity Bands 1–3), provided that future designs maximize screen usage to enlarge images and text. Students intuitively understood how to interact with the items (selecting an object and moving it). This success was attributed to high familiarity with drag-and-drop mechanics from instructional apps and games used in the classroom.

- Performance
    - All students in Complexity Bands 2 and 3, and many in Complexity Band 1, completed these items independently or with allowable supports. Students intuitively understood how to respond, largely because they were familiar with similar features from both informal and instructional settings.
- Engagement
    - Students perceived the tasks as "easy" and "fun," leading to high engagement levels (83% rated as very engaged).
- Efficiency
    - This item type required the shortest time to administer, required very little support from test administrators, and was perceived by students as easy to use. Because students could work independently, test administrators did not need to provide heavy scaffolding, modeling, or navigation support.
- Design
    - The inclusion of pictures combined with words effectively supported student understanding. Test administrators viewed this item type as a successful transition from using physical manipulatives to responding with pictures and words.
- Support
    - The combination of pictures and text effectively supported student understanding. The combination of pictures and words helped bridge the gap between concrete manipulatives (used in instruction) and digital assessment.
- Application
    - This item type is suitable for testing a wide range of skills, such as sorting, categorization, patterns, and models.

- Challenges
    - The items were often left-justified, leaving a large amount of unused white space on the screen. This inefficient use of space limited the size of the content. Because of the layout issues, the images and text were often too small, creating a barrier for students with visual impairments or fine-motor difficulties.

Recommendations for improving accessibility for this item type include expanding the spatial layout of the item to use the entire page for item presentation. The item presentation in the piloted items was left-justified, which left substantial unused white space. Improving the spatial layout would allow images to be larger and more accessible for students with vision impairments. Developers should use full-screen width to maximize the size of images and text, ensuring they are easily perceivable for all students.

DROP-DOWN ITEMS

The drop-down item type showed promise, although it presented specific accessibility hurdles. While functional, this type is less effective than drag-and-drop or hot-spot items because of the text-heavy nature and small navigation targets. It is recommended for use only when it offers a clear construct advantage over multiple-choice items and the interactive elements (i.e., arrows) are significantly enlarged. The arrow button to open the menu was small, making independent navigation difficult for students with fine-motor challenges. The lack of picture support in the drop-down menu forces students who cannot read text to rely entirely on the administrator. Test administrators rated the item type *effective* and *easy* or *about right* for approximately 90% of students. Students in grade 4 and above often had prior experience with drop-down menus from other assessments (e.g., i-Ready) or instruction. Older students may need to navigate this kind of item type instructionally.

- Performance
    - All but one student in Complexity Bands 2 and 3 were able to complete the drop-down items. Even among students in lower complexity bands, many could complete the items with support. Some students in fourth grade had some prior experience with drop-down menus in computer-based instruction but relied heavily on test administrators to read the text-only options aloud, consistent with DLM procedures.
- Efficiency
    - These items took longer to administer than drag-and-drop or hot-spot items.
- Accessibility
    - Small navigation targets

- Text Accessibility
    - The lack of picture support in the drop-down menu forced students who could not read text to rely entirely on the administrator. If drop-down items w not read aloud during selection of the drop-down target, then consider whether the item offers a significant advantage over traditional multiple-choice items, particularly given that both formats often require the test administrator to read the content to the student.
- Administration Variability
    - Inconsistent test administration was observed. Some teachers read sentences with every option, while others read the sentence once.
- Validity Concern
    - Because the administrator often must read the options and sometimes navigate the small arrow, the item may inadvertently measure listening comprehension or the administrator's support rather than the student's science knowledge.
- Design Consideration
    - Cognitive and Load and Layout
        - Error rates increased when the menu contained three or more rows of options.
    - Response Options
        - Observations suggest students were more confident using the menus when selecting numbers rather than words.

Recommendations for accessibility of items to consider the motor control of users. The interactive arrow button required to open the menu was described as "exceedingly small." This created a significant barrier for students with fine-motor challenges or those using touch screens, often preventing independent navigation. The menu options were mostly text only. This design choice forced students—regardless of their science ability—to rely on the test administrator to read the options aloud. This reduces student independence and potentially measures listening comprehension rather than science knowledge. Error rates increased when the menu contained three drop-down options.

Hot-Spot Items

The hot-spot item type was highly effective and shared many of the positive attributes of the drag-and-drop items. Students at Complexity Bands 2 and 3 intuitively understood that they needed to touch the image to select a response. It was rated as highly effective for measuring student knowledge. It is engaging and familiar to students, but operational success depends on improving the "unselect" function and ensuring that images and hit zones are large and centrally located.

- Performance
    - Students intuitively understood how to select responses. All students in Complexity Bands 2 and 3 (and more than half in the Foundational band and Complexity Band 1) successfully completed these items independently or with allowable supports.
- Efficiency
    - Like drag-and-drop items, hot-spot items were quick to administer (averaging 2.5 minutes) and maintained student attention without requiring heavy intervention from test administrators.
- Design
    - The use of pictures combined with words was noted as effective for supporting student comprehension and response. Administrators noted that this format mirrored common classroom visual supports (e.g., pointing to a picture on a communication board or worksheet).
- Presentation
    - Visual Layout
        - Some items used left-justified images, which failed to use full screen real estate. Administrators requested that images be centered and enlarged to maximize visibility.
    - Stimulus Clarity
        - In specific items, students missed critical details (e.g., a small arrow pointing to grass) because the visual cues were not prominent enough.
- Challenges to Navigation
    - The hot-spot navigation was not intuitive for students at Complexity Band 1. Some students did not know how to unselect a response.
        - Selection and Deselection Mechanics (Usability Issue)
            - The biggest functional challenge was deselection for changing an answer. Students did not intuitively know they had to deselect (i.e., click again) a wrong answer or a choice after reaching the maximum selection limit. This prevented the student from selecting a new response, which caused frustration.
        - Touch-Target Size
            - Like other item types, the hot (i.e., clickable) area was sometimes too small.

To improve accessibility, we recommend increasing the hot-spot area to provide students with a larger response area. Like other item types, the hot (i.e., clickable) area was

sometimes too small or not well-defined. Visual clarity in specific items, critical visual cues (e.g., a small arrow pointing to grass) were not prominent enough, causing students to miss the target entirely. For example, in one item most students did not select the grass as one of the plants and did not notice the small word “plant” or the arrow that pointed to the grass. All picture–word combinations need to be large enough to be easily perceived. Increase the size of the active response area and ensure images are centered and high contrast.

To improve usability, we suggest changes aimed at making it easier and more intuitive for students to change their answer (i.e., addressing the current version where students have to click to unselect an answer before selecting a new one). The requirement to unselect an answer (i.e., click it again to remove it) before choosing a new one was not intuitive. This complication added an invisible cognitive step that caused frustration when students wanted to change their answer but could not.

TABLE-MATCH ITEMS

The table-match item type presented substantial challenges. Test administrators noted that many students do not use charts or organizers in daily instruction, making the design format itself, not only the science content, a barrier to performance.

- Cognitive Load
    - This item type taxed the short-term memory of some students, who frequently forgot what they were being asked to do while trying to complete the task.
        - Evidence of split attention:
            - To answer correctly, a student must track the intersection of a row and a column. This spatial tracking proved difficult, and students often got lost in the grid.
            - Defaulting to guessing when faced with multiple rows—confused students often reverted to selecting the first button in a row regardless of the correct answer, indicating a breakdown in understanding the format.
- Engagement
    - The items required sustained attention, leading to student fatigue and decreased focus.
- Administration
    - These items took longer to administer, with students requiring significant support from administrators to navigate the grid, understand the directions, and read the content.

    - Administrators varied significantly in how they read the table. Some repeated the column headers for every row (reducing memory load), while others did not (increasing difficulty). This introduced unfair variability in the testing experience.
- Accessibility Barriers
    - The visual layout and interactive elements were poorly optimized for the target population. The radio buttons used for selection were described as "extremely small." Students struggled to target them accurately, especially on touch screens (e.g., iPads). The small size required a level of fine-motor control that many students did not possess, forcing administrators to take over the task of clicking. Other concerns were visibility on smaller devices or when magnification was used, and the entire table often did not fit on the screen. This required scrolling or obscured column headers and row options simultaneously.

We recommend changes to improve accessibility. The table-match format had a split-attention effect. The grid format imposes a high cognitive load, requiring students to visually track the intersection of rows and columns. Without visual anchoring (e.g., the administrator pointing), students frequently lost their place. The radio buttons used for selection were described as "extremely small." This issue required a level of fine-motor precision that many students did not possess, forcing administrators to take over the physical task of navigation. On smaller devices or when magnification was used, the entire table often did not fit on the screen. This issue required scrolling, which prevented students from seeing column headers and row options simultaneously, effectively hiding the instructions needed to answer the question.

SIMULATION ITEMS

The simulation item type was the most challenging format evaluated. It proved difficult for students to complete and for test administrators to manage due to a combination of technical, cognitive, and design factors.

- Technical Issues
    - Similar to findings by Tiemann et al. (2019), technical delays—such as videos loading slowly or displaying incorrectly—caused students to lose attention. When content required scrolling or did not render immediately, engagement dropped. For this student population, even a short delay caused a break in engagement. When content did not render immediately, students lost interest or became distracted.
- Comprehension

    - Although students generally enjoyed watching the videos, they often could not remember or understand the subsequent task they were asked to perform.
- Administrator Reliance
    - Directions for simulations were often lengthy and multistep. Test administrators had to spend time deciphering these complex instructions to guide students through the items effectively. Students relied heavily on administrators not just for reading, but for navigating the interface and interpreting what the item was asking them to do.
- Cognitive Load
    - While students enjoyed the visuals (e.g., videos of cars or ice cream), they often forgot the specific task or question by the time the video finished playing. They could not retain the necessary information to respond.
- Task Design
    - The complexity of the task stripped students of their ability to work autonomously. None of the students in the Foundational band or Complexity Band 1 could complete these items independently. Even students in higher bands (2 and 3) required significant modeling. Administrators noted that this type of complex simulation is not typically used in instruction, meaning neither they nor the students had a frame of reference for how to interact with it. The design led to lengthy test administration, which in turn led to significant student fatigue. In one instance, a student physically left the testing session because of the length and difficulty of the task.

Improvements to design are needed before using this item type. This item type showed the most inconsistency across devices. Content rendered differently on iPads compared to Chromebooks (e.g., requiring scrolling on one but not the other). Both students and test administrators were confused about what was expected and how the items functioned. Specific improvements may include changing the pacing and chunking directions and tasks into smaller, more manageable steps. Doing so may enhance student comprehension and maintain student engagement and motivation. Instructional opportunities to practice using these item types is necessary for students to learn how this item type works.

Improving the mechanics of the human–computer interface (e.g., timing and navigation for videos to play and render as intended, regardless of device) would improve usability. Heavy media elements caused video-loading lags and rendering errors. For the intended population, even short delays broke attention and engagement. Items frequently required scrolling to see all components. Students often missed content located below the screen

when scrolling was necessary leading to incomplete responses. The multistep directions and complex interface (e.g., manipulating variables, running simulations, interpreting tables) often taxed cognitive load. Students often forgot the task while navigating the technology. To be accessible, simulations must be simplified (i.e., chunked into smaller steps) and be rigorously tested to ensure instant rendering across all supported devices.