Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 29 Number 13, August 2024 ISSN 1531-7714

Response Process Evidence for Academic Assessments of Students with Significant Cognitive Disabilities[1]

Meagan Karvonen, *University of Kansas: ATLAS*
Russell Swinburne Romine, *University of Kansas: ATLAS*
Amy K. Clark, *University of Kansas: ATLAS*

This paper describes methods and findings from student cognitive labs, teacher cognitive labs, and test administration observations as evidence evaluated in a validity argument for a computer-based alternate assessment for students with significant cognitive disabilities. Validity of score interpretations and uses for alternate assessments based on alternate academic achievement standards (AA-AAAS) for students with the most significant cognitive disabilities require nontraditional sources of evidence about student-item interactions and the influences teachers have on those interactions. Our findings provide evidence that the assessment has been designed so students can show what they know and can do on multiple choice, technology-enhanced, and teacher administered item types and that teachers administer the assessments in a way that allows students to respond as intended. We conclude with a discussion on how the findings inform future test development, limitations, and implications for the use of these research methods for gathering validity evidence for an AA-AAAS.

Keywords: validity, response process, alternate assessment, special education, intellectual disability

Introduction

Professional standards for educational assessment recommend that validity evidence for assessments include five critical sources of evidence, including evidence based on response process (American Educational Research Association et al., 2014). Response process data demonstrate the extent to which students engage in the cognitive processes the assessment items intend to measure, which may range from following ordered steps to complex cognitive thinking (Ercikan & Pellegrino, 2017), and the extent to which construct-irrelevant response processes are minimized (Thomas et al., 2023).

Direct evidence of response process typically comes from tasks that elicit external responses that indicate the internal cognitive processes needed to respond to items. Common methods for evaluating student response process include think-aloud protocols and cognitive labs (e.g., Haertel, 1999; Leighton, 2013; Padilla & Leighton, 2017), in which

[1] This work was funded in part by the U.S. Department of Education, Office of Special Education Programs, under Grant 84.373X100001. The views expressed herein are solely those of the authors, and no official endorsement by the U.S. Department of Education should be inferred.

students verbally describe their thinking as they respond to assessment items. Think-aloud protocols are used to gather evidence of student problem-solving, while cognitive labs, or cognitive interviews, yield evidence of student comprehension of the content being assessed (Leighton, 2017). Cognitive lab methods also may include interspersed interview questions to probe for additional evidence such as student reasoning or perception. For cognitive labs to yield useful data, participants must have knowledge of the assessment content to answer items, but also strong working memory and metacognitive knowledge to verbalize their response process.

The additional cognitive load required for cognitive lab participants introduces challenges for using these methods to gather response process evidence for alternate assessments based on alternate academic achievement standards (AA-AAAS), a type of large-scale academic assessment used since 2000 for students with the most significant cognitive disabilities. The term "significant cognitive disability" does not designate a specific disability, but rather describes the group of students (approximately 1% of the population) who are eligible to participate in AA-AAAS for school and state accountability purposes because they cannot meaningfully participate in general education assessments even with accommodations (Every Student Succeeds Act [ESSA], 2015). The most frequent disability labels for students who participate in AA-AAAS include intellectual disabilities, autism, or multiple disabilities (Burnes & Clark, 2021; Thurlow et al., 2016). AA-AAAS participants vary in their modes and complexity of communication. For example, in a 2019 survey of more than 90,000 students identified by their teachers as being eligible for AA-AAAS, an estimated 76% of students used speech to meet their expressive communication needs (Burnes & Clark, 2021). Among students who did not yet communicate with speech, sign language, or augmentative and alternative communication systems, approximately half (52%) used conventional gestures or vocalizations to communicate intentionally; and 47% exhibited behaviors that were not intentionally communicative but might be interpreted by others as such (Burnes & Clark, 2021). Challenges with working and short-term memory, as well as metacognition, are common for students with significant cognitive disabilities (Kleinert et al., 2009).

The combination of students' expressive communication and cognitive disabilities poses challenges with using cognitive lab or think-aloud protocols as a means of gathering response process evidence from students taking AA-AAAS (Marion & Pellegrino, 2006). To date, published, empirical response process evidence for AA-AAAS has been limited to evidence of how teachers interpret and rate students on a skills checklist type of AA-AAAS (e.g., Elliott et al., 2009; Goldstein & Behuniak, 2011) and analyses of item difficulty and level of support provided during administration (Carrizales & Tindal, 2009). Johnstone et al. (2006) included students with cognitive disabilities (not necessarily significant cognitive disabilities) in cognitive labs to evaluate item features and discovered that, although students with other types of disabilities were able to participate, those with cognitive disabilities had difficulty understanding what was expected of them and verbalizing their thoughts succinctly. Response process evidence used to inform the design of alternate English language proficiency assessments has been based on observation of student-administrator interactions and administrator interviews (Gholson et al., in press).

The design of AA-AAAS introduces questions about what evidence of response process is needed. Most published validity evidence is based on AA-AAAS designed before 2010, which were primarily portfolios or sets of performance tasks (Altman et al., 2010). With these formats, teachers mediate the assessment experience more than they would with multiple-choice tests due to their role in administration and scoring or selection of content for inclusion. To account for the teacher's role in supporting and mediating student-item interactions, evidence of implementation fidelity may be another source of response process evidence. For example, Hager and Slocum (2008) reviewed six sources of validity evidence for a performance-based AA-AAAS and included evidence related to the test administration process. Trained raters evaluated video-recorded administrations for evidence of fidelity. While relatively high rates of fidelity were reported for some criteria (e.g., whether the teacher presented the prescribed directions), complete fidelity to all expectations was relatively low. Since many AA-AAAS are designed to have some degree of flexibility in their administration, evaluating implementation fidelity for AA-AAAS requires consideration of "intended

variability" (Marion & Pellegrino, 2006, p. 53): the ways in which a standardized assessment is designed to be administered differently to provide access for a heterogeneous group of students.

By meeting the criteria to participate in AA-AAAS, students have significant cognitive disabilities that affect their ability to participate in validity studies on response process. However, to support assertions that knowledge and skills demonstrated on an assessment reflect students' true knowledge and skills, assessment items must "elicit cognitive processes associated with the underlying cognitive model so that observed item responses can lead to valid inferences about the construct under investigation" (Ketterlin-Geller, 2008, p. 10). The purpose of this paper is to describe methods and findings for three sources of response process evidence for AA-AAAS: modified student cognitive labs, teacher cognitive labs, and test administration observations. The illustrations are grounded in the Dynamic Learning Maps (DLM) Alternate Assessment System, which uses an argument-based approach to validation (Clark & Karvonen, 2020).

Dynamic Learning Maps Alternate Assessment System

The DLM Alternate Assessment System features assessments in English language arts (ELA), mathematics, and science in grades 3 through 8 and high school. Assessments are administered as a series of "testlets." Each testlet contains a nonscored engagement activity and three to nine items. There are two modes of administration, and the mode depends on the testlet's content. In about 80% of ELA and mathematics testlets and about 67% of science testlets, students interact directly with the computer, using human- and technology-delivered accessibility supports as needed. These testlets are called "computer administered." Item types used in computer-administered assessments include single-select multiple choice; multiple-select multiple choice; and, on a limited basis, technology-enhanced item types that require students to sort, match, or select text within a passage.

The remaining testlets are also delivered via computer but are referred to as "teacher administered" because teachers use the online content to guide administration of performance tasks and record the responses that the student expressed offline. These testlets present on-screen instructions to the teacher that guide a structured interaction with the student that occurs outside the system. The teacher observes responses to specific prompts that are part of the activity and records the student's response in the system. Item types include single-select multiple choice and multiple-select multiple choice. These testlets are typically used at lower complexity levels for students who are still working toward consistent symbolic communication. In all grades and at all levels of complexity, writing testlets are also teacher administered. The writing testlet guides the test administrator to deliver a structured writing task to the student. The test administrator evaluates student writing processes and products as the student responds to each item and enters responses into the online system.

Design of the DLM assessment system was guided by principles of universal design for assessment (Ketterlin-Geller, 2008) to allow flexibility during administration. For example, varying the timing and length of a test session, the choice of test setting and device, and the use of adaptive equipment are all options the test administrator may use as a matter of routine. Accessibility supports are available to any student, but the teacher must select supports for each student in a Personal Needs and Preferences profile maintained in the online assessment management system. Example supports include magnification and synthetic spoken audio, switch system, human read aloud, and the use of individualized manipulatives.

When making decisions about using accessibility supports for computer-administered testlets, teachers are encouraged to follow two general principles: (a) a student should respond to the content independently, and (b) a student should be familiar with the chosen supports because they have been used consistently during routine instruction. When making decisions about additional supports for teacher-administered testlets, test administrators are encouraged to use options that support each student's needs for access to the content and means of making a response. At the same time, test administrators must maintain consistency in the student's interaction with the concept being measured. Students do not have to interact with identical materials or respond using the same response mode, but they should complete the

same cognitive task. Questions cannot be rephrased, and items cannot be rearranged or modified. Teachers are trained on how to administer both types of testlets with fidelity, how to make decisions about accessibility supports, and how to use options for flexibility in test administration. Teachers must annually complete required training and pass quizzes about the contents of the modules to be eligible to administer DLM assessments.

Approach to Validation for DLM Assessments

The DLM Alternate Assessment System follows an argument-based approach to validity (Clark & Karvonen, 2020; Kane, 2006), which consists of claims, comprising a theory of action, that evaluate the intended uses of assessment results. Two such claims focus on student demonstration of knowledge as items are administered.

1. Students interact with the system to show their knowledge, skills, and understandings.
2. Teachers administer the assessments with fidelity.

The second claim is important because teachers' involvement in the administration process may introduce additional variation in item responses due to variables unrelated to either the student or the construct being measured.

Each of these claims in the theory of action has multiple underlying propositions that must be evaluated. For example, students should interact with the system as intended, and responses to items should reflect their knowledge, skills, and understandings. Students should be able to respond regardless of disability, health, or other constraints. Regarding test administration practices, teachers are expected to allow students to respond as independently as possible. When teachers must enter student responses on a student's behalf, the entries should accurately reflect the student's demonstration of the skill.

To evaluate the propositions underlying each claim in the theory of action, multiple sources of evidence are collected during the test administration process. Methods used to collect this evidence draw from well-established practices, but in some cases they are modified to fit the distinctive characteristics of the student population and the assessment. This study focuses on modified student cognitive labs, teacher cognitive labs, and test administration observations. Examples demonstrate how the evidence is used to evaluate propositions underlying the claims in the validity argument. All data were collected with Institutional Review Board approval, and instruments are available from the first author upon request.

Response Process Evidence

Evidence of response process was collected from three sources: modified student cognitive labs, teacher cognitive labs, and test administration observations.

Modified Student Cognitive Labs

Cognitive labs typically are used to elicit statements during the assessment process that allow the observer to know whether the item taps the intended process or construct-relevant knowledge (Leighton, 2017). However, because of the challenges previously described in using these methods with the population of students who are eligible for AA-AAAS, we present a modified cognitive lab study appropriate for students with the most significant cognitive disabilities. Rather than gathering evidence to confirm construct-relevant response processes, we evaluated whether the response demands of various item types introduced construct-irrelevant factors into the response process. Our concern with technology-enhanced item types was that they might be challenging for students with significant cognitive disabilities due to the items' cognitive demands, lack of familiarity, and physical access barriers related to students' fine-motor skills.

We recruited schools and districts administering DLM assessments to at least 3–5 eligible students to maximize onsite time. Sites identified eligible students from tested grades (3 through 8 and high school) who had sufficient symbolic communication systems to be able to interact with the content of on-screen items, without physical assistance, using a keyboard/mouse, tablet, or other assistive technology. Inclusion criteria also required students have some verbal expressive communication. We conducted modified cognitive labs with all students who provided parental consent. This included 27 students from five sites in three states. Of the 27, 18 were male, 6 were female, and 3 were unspecified; 17 used a computer and mouse, 5 used a computer and trackpad, 3 used a SMART board, and 2

used an iPad to respond. The modified cognitive labs were conducted by an examiner and observed by a second researcher who used a protocol to document student responses and actions during the session; we also video recorded the cognitive labs for subsequent review. The researchers consisted of six DLM staff members who had content and/or special education background. Four had doctorate degrees and one was a doctoral student. All researchers received training on the protocol ahead of site visits.

The modified cognitive labs focused on student interaction with four types of technology-enhanced items: drag-and-drop, click-to-place, select-text, and multiple-select multiple choice. The first three item types were designed specifically for DLM assessments and are administered through a user interface designed for this population. Drag-and-drop and click-to-place items are used for sorting. The difference between them is that drag-and-drop items require continuous selection (i.e., clicking and dragging) while click-to-place items require clicking on the origin and clicking on the intended destination. The latter item type is accessible for students who use switches to interact with computers, but one theory was that students who do not use switches would also find clicking without dragging easier than drag-and-drop because the process placed fewer demands on fine-motor skills. Both the drag-and-drop and click-to-place items were built to require a similar cognitive response process: sorting objects into categories. To facilitate comparisons with drag-and-drop and click-to-place items, multiple-select multiple-choice items were also constructed to access a response process requiring the student to select the answer options that matched a category. Select-text items are used only in some ELA assessments. In a select-text item, answer choices are marked with a box around the word, phrase, or sentence. When a student makes a selection, the word, phrase, or sentence is highlighted in yellow. To clear a selection, the student clicks it again.

To avoid relying on items that might be too difficult and therefore inappropriate for use in cognitive labs (Johnstone et al., 2011), four-item testlets were constructed with content that did not rely on prior academic knowledge. For example, a sorting item that required sorting of shapes required students only to interpret the instructions and move the shapes to boxes. They did not have to have knowledge of shape classification to complete the item. Each testlet contained one type of item. For drag-and-drop and click-to-place item types, both the number of objects to sort and the number of categories varied, with more-complex versions of the item type appearing later in the testlet. Each student completed two testlets (one per item type) except one student who completed only one. Testlet ordering assignments were counterbalanced across students. Fifteen students completed drag-and-drop, 11 completed click-to-place, eight completed select-text, and 11 completed multiple-select multiple-choice testlets. The eight students who completed select-text testlets also completed a testlet that used the same content as the select-text items, but presented in a single-select multiple-choice format.

For each item type, the examiner looked for evidence of challenges students encountered with each step of the item-completion process (e.g., for drag-and-drop items, the steps are initial item selection, manipulation, and item placement) and whether the student experienced challenges based on the number of objects to be manipulated per item. For all item types, the examiner also looked for evidence of the student's understanding of the task. If the student was not able to complete the task without additional assistance, the examiner provided further instructions on how to complete the task.

Students were not asked to talk while they completed the items. Instead, they were asked questions at the end of each testlet and after the session. These questions were more simplified than those described by Johnstone et al. (2011; e.g., "What makes you believe that answer is the right one?") and required only yes/no answers (e.g., "Did you know what to do?"). Students were asked the same four questions, in the same sequence each time. The yes/no response requirement and identical sequence mirror instructional practice for many students who are eligible for AA-AAAS. A researcher reviewed videos to confirm that the ratings of students' response challenges were correctly recorded by the on-site observer.

Students could encounter challenges when responding to drag-and-drop and click-to-place item types due to response demands when selecting the desired object, maintaining continuous selection, or selecting the group. Students could also have difficulty indicating their response when the item contained a

large number of objects that required sorting. Despite the many similarities between these two item types, students tended to have more difficulty with click-to-place items than with drag-and-drop items (see Table 1).

Students' sources of challenge in responding to multiple-select multiple-choice items (n_s = 11 students, n_i = 44 items) included difficulty with the multiple-select concept (n_i = 18, 40.9%), difficulty with selection of the first object (n_i = 4, 9.0%), and difficulty with selecting subsequent objects (n_i = 6, 13.6%). Nine students (20.5%) needed assistance to complete the item. The select-text items required less manipulation of on-screen content and only one selection to respond to the item. Across eight students and 32 items, there were only two items (6.3%) where the student had difficulty selecting the box and two (6.3%) where the student needed assistance to complete the item.

Table 2 summarizes student responses to post-hoc interview questions. Drag-and-drop and select-text items were most often liked, perceived as easy, and required a response process that students understood. A smaller percentage of students reported that they liked multiple-select multiple-choice items or knew how to respond to the item. Students reported liking click-to-place items the least and fewer students knew how to respond to click-to-place items. One student reported drag-and-drop and select-text items were easy and hard. Observers of the modified cognitive labs noted challenges with these item types that were generally consistent with student interview responses.

Teacher Cognitive Labs

Teacher cognitive labs are an additional source of response process evidence that have been recommended for AA-AAAS where item responses consist of teacher ratings (e.g., Goldstein & Behuniak, 2011) rather than student responses. This approach was used for DLM teacher-administered testlets because teachers interpret student response behaviors and respond to items about the student's response.

We asked participating DLM state education agencies to identify teachers of students with significant cognitive disabilities to participate in the teacher cognitive labs, separate from the student cognitive labs. State education agency staff identified 15 teachers in five schools across two states. Nine participants were female, two were male, and four did not specify gender; three taught 6th grade, four taught 7th grade, three taught 8th grade, four taught high school, and one did not specify grade(s) taught.

Teacher cognitive labs included think aloud and interview components. Teachers completed think-aloud procedures while preparing for and administering teacher-administered testlets in reading, writing, and mathematics. During preparation they were presented with a Testlet Information Page, which is a short document that provides the background information needed to prepare to administer the testlet. For example, a Testlet Information Page may contain instructions about materials needed, guidelines

Table 1. Number of Items Presenting Challenges During Item Response

Source of challenge	Drag-and-drop (n = 60 items)		Click-to-place (n = 44 items)	
	n	%	*n*	%
Difficulty with object selection	6	10.0	16	37.2
Difficulty with continuous selection	7	11.5	—	—
Difficulty with group selection	6	10.0	26	60.5
Difficulty with number of objects	2	3.0	10	23.3
Needed assistance to complete	7	11.5	26	60.5

Note. Drag-and-drop items administered to 15 students. Click-to-place items administered to 11 students.

Table 2. Affirmative Student Responses to Post-Hoc Interview Questions

Question	DD (*N*= 15)		CP (*N* = 11)		MSMC (*N* = 11)		ST (*N* = 8)	
	n	%	*n*	%	*n*	%	*n*	%
Did you like it?	15	100.0	7	63.6	9	81.8	8	100.0
Was it easy?	15	100.0	8	72.7	10	90.9	8	100.0
Was it hard?	1	6.0	1	9.0	1	9.0	1	12.5
Did you know what to do?	14	93.3	6	54.5	8	72.7	8	100.0

Note. DD = drag-and-drop; CP = click-to-place; MSMC = multiple-select multiple choice; ST = select text.

for substitution of materials, instructions about alternate text to be read aloud when describing pictures to students with visual impairments, and an indication that calculator use is appropriate on a mathematics testlet.

Teachers were asked to think out loud as they read through the Testlet Information Page. Next, the teacher gathered materials needed for the assessment and administered the testlet. In-vivo probes were sometimes used to ask about teacher interpretation of the on-screen instructions and the rationale behind decisions they made during administration. When the testlet was finished, teachers also completed post-hoc interviews about the contents of test administration instructions, use of materials, clarity of procedures, and interpretation of student behaviors.

All teacher cognitive labs were video recorded, and an observer took notes during the administration. Observers included two DLM test development staff members; one was a special education teacher and doctoral student, the other was a general education English language arts teacher. Analysis involved recording evidence of intended administration and sources of challenge to intended administration at each of the following stages: (a) preparation for administration, (b) interpretation of teacher directions within the testlet, (c) testlet administration, (d) interpretation of student behaviors, and (e) recording student responses. Through this lens, evidence related to fidelity (a, b, c, e) as well as response process (d) was identified.

Evidence of teachers' interpretation of student behaviors indicated that the ease of determining student intent depended in part on the student's response mode. Teachers were easily able to understand student intent when the student indicated a response by picking up objects and handing them to the teacher. In a case when a student touched an object rather than handing it to the teacher, the teacher accepted that response and entered it but wondered whether the student was just choosing the closest object. When a student briefly touched one object and then another, the teacher entered the response associated with the second object but commented that she was not certain if the student intended that choice. When a student's gesture did not exactly match one of the response options, the teacher was able to verbalize the process of deciding how to select the option that most closely matched the student's behavior. Her process was consistent with the expectations set forth in the Test Administration Manual (DLM Consortium, 2023). Understanding student intent from eye gaze required more interpretation by the teacher. For example, one teacher held objects within the student's field of vision and put the object that represented the correct answer away from the current gaze point so that a correct response required intentional eye movement to the correct object.

Test Administration Observations

A test administration observation protocol is used to gather information about how teachers administer testlets. This protocol was developed by project staff to give observers a standardized way to describe how a DLM testlet was administered, regardless of the observer's role or experience with DLM assessments. Test administration observations are collected by state and local education agency staff and DLM project staff. The observation protocol is used only for descriptive purposes. It is not used to evaluate or coach the teacher or to monitor student performance and is therefore anonymous. The protocol captures data

about student actions (e.g., navigation, answering), teacher assistance, variations from standard administration, and engagement and barriers to engagement during administration of a single testlet. Most items are a direct report of what is observed, for instance, how the test administrator sets up for the assessment and what the test administrator and student say and do. One section asks observers to make judgments about the student's engagement during the session. No information is collected about the student or teacher. Accuracy of the observations may depend on situational factors, such as the observer's location in the room relative to the computer screen (for instance, an observer may note if they were unable to see the computer screen).

During computer-administered testlets, students can interact independently with a computer, using special devices such as alternate keyboards, touch screens, or switches as necessary. In teacher-administered testlets, the test administrator is responsible for setting up the assessment, administering it to the student, and recording responses in the user interface. The observation protocol contains different questions specific to each type of testlet.

During observations, the student's typical test administration process was observed with the student's actual test administrator. Data described in this paper are based on observations made in five states. Of the 147 test administration observations collected, 117 (79.6%) were of computer-administered assessments and 30 (20.4%) were of teacher-administered testlets. Of the 147 observations, 70 (47.6%) were of ELA reading testlets, 32 (21.8%) were of ELA writing testlets, 40 (27.2%) were of mathematics testlets, and 1 (0.7%) was of a science testlet. Most testlets (81.6%) were administered in students' typical classrooms.

Several parts of the observation protocol correspond to specific propositions in the validity argument. For example, one proposition addressed is: "Test administrators allow students to engage with the system as independently as they are able." The observation protocol includes a multiple-choice item where observers select the best description of the student's interaction with the system, independently or with supports. For computer-administered testlets, as Table 3 shows, clarifying directions (26% of observations) removes student confusion about the task demands (a potential source of construct-irrelevant variance) and supports the student's meaningful, construct-related engagement with the item. In contrast, using physical prompts such as hand-over-hand guidance (also 26% of observations) indicates that the teacher directly influenced the student's answer choice.

Interaction with the system includes interaction with the assessment content as well as physical access to the testing device and platform. The fact that teachers navigated one or more screens in 73% of the observations does not necessarily mean the student was prevented from engaging with the system as independently as possible. Depending on the student, teacher navigation may either support or minimize students' independent, physical interaction with the assessment system. Navigating for students who are able to do so independently would be counter to the proposition that students are able to interact with the system as independently as possible. The observation protocol did not capture the reason the teacher chose to navigate, and the reason was not easily inferred by an observer.

Related to test administrators supporting students' independent engagement with the system is another proposition: "Students are able to interact with the system as intended." These results are also summarized in Table 3. Independent answer selection was observed in 39% of the cases, and the use of eye gaze (one unique form of independent selection) was seen in 21% of the observations. Verbal prompts for navigation and response selection are strategies that are within the realm of allowable flexibility during test administration. Although these strategies can be used to maximize student engagement with the system and promote the type of student-item interaction needed for a construct-relevant response, those practices also indicate that students were unable to sustain independent interaction with the system for the entire testlet without support.

Another proposition, "Students are able to respond to tasks irrespective of a sensory, mobility, health, communication, or behavioral constraint," was evaluated by having observers note whether there was difficulty with accessibility supports (including lack of appropriate available supports) during observations of teacher-administered testlets. Of the 30 observations of teacher-administered testlets, observers

Table 3. Actions Observed During Computer-Administered Testlet Administrations (N = 117)

Action	*n*	%
Test administrator		
Navigated one or more screens for the student	**85**	**72.6**
Repeated question(s) before student responded	**76**	**65.0**
Used verbal prompts to direct the student's attention	65	55.6
Defined vocabulary used in the testlet	34	29.1
Used physical prompts	*30*	*25.6*
Clarified directions	**30**	**25.6**
Repeated question(s) after student responded	11	9.4
Asked the student to clarify one or more responses	10	8.5
Reduced number of choices available to student	*6*	*5.1*
Student		
Selected answers with verbal prompts	53	45.3
Selected answers independently	**45**	**38.5**
Used manipulatives	30	25.6
Indicated answers using eye gaze	**24**	**20.5**
Navigated the screens independently	**19**	**16.2**
Navigated the screens with verbal prompts	8	6.8
Indicated answers using materials outside of system	4	3.4

Note. Bold text = supporting evidence; italic text = nonsupporting evidence; roman text = neutral evidence.

noted difficulty in two (6.7%) cases. Additional evidence for this proposition was gathered by observing whether students were able to complete testlets. Of the 147 test administration observations collected, in 132 cases (89.8%) students completed the testlet.

Observers provided evidence for another proposition—"Teachers enter student responses with fidelity"—by rating whether test administrators accurately captured student responses. To record student responses with fidelity, test administrators needed to observe multiple modes of response. In the 30 observed teacher-administered testlets, students responded via gesture (n = 12, 40%), verbal response (n = 7, 23.3%), eye gaze (n = 2, 6.7%), or other (n = 6, 20%). Five students (16.7%) made no response. Across all observations and student response modes, test administrators recorded responses with fidelity in 93.3% of observations. In the remaining instances, the observer did not record a response (e.g., due to position in room and being unable to see response).

Computer-administered testlets provided another opportunity to evaluate fidelity of response entry when test administrators entered responses on behalf of students. Test administrator response entry is a support recorded on the student's Personal Needs and Preferences profile and is recommended for a variety of situations (e.g., students who may have limited motor skills necessary to interact directly with the testing device even if they can cognitively interact with the on-screen content). Observers recorded whether the response entered by the test administrator matched the student's response. In 75 of 98 observations of computer-administered testlets (76.5%), the test

administrator entered responses on the student's behalf. In 98.6% of those cases (n = 74), observers indicated that the entered response matched the student's response. No further information was recorded about the one case that did not match. Collectively, this evidence generally supports the proposition that teachers entered student responses with fidelity.

Discussion

For AA-AAAS such as the DLM assessment system, characteristics of the student population and the central role of teachers as test administrators require the use of multiple sources of evidence to evaluate claims about response process. Our approach was to include as much direct evidence as possible from students, supplemented by less-direct evidence including observations. The validity argument includes testable propositions about intended (i.e., construct-relevant) and unintended (i.e., construct-irrelevant) mechanisms involved in the entire test administration process, not just student responses to items. Validity evaluation then considers the extent to which there is confirmatory evidence of intended processes and counterevidence of unintended processes to make judgments about the extent to which the claims are supported. This study was delimited to evidence about the test administration process that was gathered during initial assessment development and early operational administration; it did not consider other factors that may influence interpretations about student response (e.g., opportunity to learn).

Table 4 summarizes overall findings from the three studies described in this paper. These studies provide examples of the types of evidence test developers may collect to evaluate claims that an AA-AAAS was designed so students can show what they know and can do and that teachers administer the assessments in a way that allows students to respond as intended. Response process data are part of a larger body of evidence, including procedural and empirical data, spanning the complete set of validity claims for the DLM assessment system. While there is not yet consensus on how to evaluate validity evidence (e.g., Carrillo-Avalos et al., 2023), organizing the findings according to the associated validity propositions was a useful tool for integrating evidence when evaluating claims in the DLM validity argument (Clark & Karvonen, 2020). The DLM program concluded that the complete set of evidence, including the response process data described in this paper, provided sufficient support for the validity claims (DLM Consortium, 2022). Other programs should consider their intended interpretations and uses when evaluating the role of response process data collection and the extent that collected evidence supports conclusions drawn from results.

Observations and cognitive labs yield data on a relatively small sample, but these were the most direct types of evidence available during the development and early operational years. Larger-scale data collection provides additional but less-direct evidence. For example, evidence that teachers choose and implement appropriate supports is currently limited to teacher self-report on an annual survey. Until the assessment platform supports tracking of accessibility support use, there is no other mechanism for gathering large-scale data on actual testlet-by-testlet use of teacher-administered accessibility supports (e.g., human read aloud). State education agency staff could compare accessibility supports selected on the Personal Needs and Preferences profile against testing accommodations listed in students' IEPs, but that type of study would allow us to evaluate only correspondence of the choice of supports on the IEP and Personal Needs and Preferences profile, not necessarily which supports were used during administration.

Results described in this paper have already informed improvements in test development and resources to support test administration. For example, test development guidelines have been updated to specify the conditions under which technology-enhanced items may be used and how different item types may be combined in a single testlet. Similarly, the test administration manual and required test administrator training have been revised according to teacher misconceptions that emerged during test administration observations and teacher cognitive labs. These studies also led to improvements in the data-collection protocols. For example, test administration observation protocols were reorganized and streamlined to better match the flow of test administrations.

Table 4. Findings Associated With Propositions in the Validity Argument

Proposition from validity argument	Evidence
Students are able to interact with the system as intended.	• Students successfully completed select-text and drag-and-drop technology-enhanced items. Students experienced some difficulty with click-to-place and multiple-select multiple-choice items. • Some students navigated computer-administered assessments independently and selected answers without support. • Students used a variety of response modes to indicate selection of answers on computer-administered testlets.
Student responses to items reflect their knowledge, skills, and understandings.	• For select-text and drag-and-drop items, item selection and manipulation demands did not introduce construct-irrelevant variance into the response process. • Response demands in multiple-select multiple-choice and click-to-place items indicates responses in those formats may not fully reflect student knowledge, skills, and understandings.
Students are able to respond to tasks irrespective of a sensory, mobility, health, communication, or behavioral constraint.	• In the majority of observations of teacher-administered testlets (93.3%), students did not experience difficulty using supports. • In the majority of observations (89.8%), students were able to complete the testlet.
Test administrators allow students to engage with the system as independently as they are able.	• Teachers entered student responses in a majority of the observations (76.5%). This was sometimes due to students' physical access barriers and other times due to student behavior. However, in some cases teachers treated their role in navigation and response entry as part of the regular assessment routine. • Teachers engaged in supporting behaviors, such as navigating screens for the student or repeating a question before the student responded. In limited instances they engaged in nonsupporting behaviors, such as hand-over-hand (26%) or reducing answer choices (5%).
Test administrators enter student responses with fidelity.	• In almost all observations (93%), teachers who entered responses on a student's behalf chose responses that matched the student's behavior. • In observations of teacher-administered testlets, teachers interpreted responses across different response modes such as verbal, gesture, and eye gaze. In some instances, teachers indicated uncertainty in their interpretation of student behaviors in these modalities. • Teacher cognitive labs did not reveal evidence of teacher misinterpretation of student responses or of selecting a response that did not reflect the student's behavior.

Limitations

We identified limitations for each of the studies that can inform future study designs. The student cognitive labs were not designed to collect confirmatory evidence that students used the intended cognitive process. Instead, the labs evaluated the possibility that construct-irrelevant item features

would negatively affect the student-item interaction. In some respects, the cognitive lab protocol allowed for straightforward data collection on the construct-irrelevant parts of item response. However, even with simplified interview questions, the fact that one student rated a testlet as both easy and hard highlights the unreliability of self-report for some students in the population. Another limitation is that, in our attempt to reduce cognitive load and simplify the questions we asked, we did not design the study to gather evidence of how the student was interpreting the on-screen content.

The overall experience with the teacher cognitive labs was mixed. Labs were conducted in an authentic environment while administering testlets to students on their case load. The goal was to capture teachers' thought processes during administration, to understand how they made choices in the moment while interacting with students who are familiar but challenging to assess. We relied on teachers to select students for these labs, and teachers tended to choose students who typically worked on more-advanced academic content and who would not typically have been eligible to take teacher-administered testlets. Their choice of students removed some of the challenges that would occur with the teacher lab methods (e.g., needing to manage student behavior and health issues while also administering the assessment and thinking aloud about the process) but made for a less authentic experience and potentially incomplete data about response process for students for whom teacher-administered testlets were designed.

There were also limitations to the test administration observation protocol, especially for computer-administered testlets. While it is valuable to be able to observe teacher intervention (e.g., navigation, response entry), we didn't record the reason for their intervention. To support accuracy and consistency of observational data, we did not want observers to make inferences about teachers' behaviors. But a teacher's motivation for navigating or entering responses on a student's behalf determines whether test administrators' actions support or inhibit students' independent interaction with the system. The same ambiguity was true for behaviors such as clarifying student answers and repeating the question after an answer was given. Depending on the situation, these could be ways of confirming that the response they enter reflects the student's intended response. Alternatively, these actions could indicate an attempt to get a student to change an answer, a practice that is not allowed.

Implications for Future Study Designs

Each study also yielded ideas for future studies. Eye-movement tracking and screen-capture software could be used to evaluate which aspects of items might be prompting certain student actions, thereby expanding our understanding of students' response processes without requiring verbalization. Future cognitive lab studies could also look more deeply within patterns of student responses. For instance, within-case analysis of response challenges, correctness of item responses, and post-hoc interview responses could be used in combination to make judgments about student understanding and construct-relevant responses. Protocols with items that require academic content knowledge will likely require typical retrospective interview questions (e.g., Johnstone et al., 2011) with scaffolding to support students' ability to respond. Those studies will also benefit from clear inclusion criteria to ensure participants can articulate ideas to an unfamiliar researcher. Even with improvements to study design, we believe it is important to have the familiar teacher present to support the student's comfort with the novel conditions and use a post-hoc interview to check interpretations of student actions with an educator who knows the student.

A different approach to teacher cognitive labs would be to remove student participation and instead have the teacher think aloud during the testlet, with a hypothetical student in mind. This approach may produce richer verbalizations as evidence of the teacher's processes during administration, but the responses may be less authentic than they would be grounded in student actions during a live administration. Another approach would be to have teachers respond to video-recorded sessions with unfamiliar students and pause periodically to ask the teacher what their next steps would be, and why, if they were the ones assessing the student. If teachers were also instructed to answer the assessment item based on a student's behavior in the video, these data could also be used to evaluate interrater agreement.

Test administration observation protocols may serve a variety of purposes. Designers should consider all of the purposes of the protocol (which may extend

beyond response process evidence) and the balance between making the questions concise and easy to answer but still informative enough to address the evaluation questions. For example, a limited number of neutrally worded post-hoc interview questions may be useful for eliciting the rationale behind test administrators' choices.

Besides lessons learned about the three specific data-collection methods, this work has also led us to think more generally about AA-AAAS evidence in an argument-based framework. As Kane (2006) noted, there is a tendency toward confirmation bias with validity evidence collected during the test development phase. The data in this study were based on the development phase and, for observational data, the first years of operational assessment. As the assessment system matures and we shift into a more neutral stance for validity evaluation, and also consider the challenges of response process data collection, the growing body of evidence is likely to rely in part on inverse logic or counterevidence. More work is also needed where the least-plausible propositions in the validity argument intersect with the most-complex data collection. For AA-AAAS, the greatest challenge may be evaluating response process for students who do not yet have expressive communication systems that allow them to make consistent, intentional responses (Erickson et al., 2024). For other assessment developers and researchers contemplating response process evidence studies for AA-AAAS, we recommend studies be designed with careful consideration of the desired claims about what students know and can do, the assessment design and expected administration, and the cognitive and noncognitive characteristics of the student population.

References

Altman, J. R., Lazarus, S. S., Quenemoen, R. F., Kearns, J., Quenemoen, M., & Thurlow, M. L. (2010). *2009 survey of states: Accomplishments and new issues at the end of a decade of change*. University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/OnlinePubs/2009StateSurvey.pdf

American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association. https://www.testingstandards.net/open-access-files.html

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. https://dynamiclearningmaps.org/sites/default/files/documents/publication/Characteristics_of_Students_Who_Take_DLM_AAs.pdf

Carrillo-Avalos, B. A., Leenen, I., Trejo-Mejía, J. A., & Sánchez-Mendiola, M. (2023). Bridging validity frameworks in assessment: Beyond traditional approaches in health professions education. *Teaching and Learning in Medicine,* online preprint. https://doi.org/10.1080/10401334.2023.2293871

Carrizales, D., & Tindal, G. (2009). Test design and validation of inferences for the Oregon alternate assessment. In W. D. Schafer & R. W. Lissitz (Eds.), *Alternate assessments based on alternate achievement standards: Policy, practice, and potential* (pp. 275–299). Brookes Publishing Company.

Clark, A. K., & Karvonen, M. (2020). Constructing and evaluating a validation argument for a next-generation alternate assessment. *Educational Assessment, 25*(1), 47–64. https://doi.org/10.1080/10627197.2019.1702463

Dynamic Learning Maps Consortium. (2022, December). *2021–2022 Technical manual—Instructionally Embedded Model*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems. https://2022-ie-techmanual.dynamiclearningmaps.org/

Dynamic Learning Maps Consortium. (2023). *Test administration manual 2023-2024*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems. https://dynamiclearningmaps.org/sites/default/files/documents/Manuals_Blueprints/Test_Administration_Manual_IE_2023-2024.pdf

Elliott, S. N., Roach, A. T., Kaase, K. J., & Kettler, R. J. (2009). The Mississippi Alternate Assessment of extended curriculum frameworks: Purpose, procedures, and validity evidence summary. In W. D. Schafer & R. W. Lissitz (Eds.), *Alternate assessments based on alternate achievement standards: Policy, practice, and potential* (pp. 239–274). Brookes Publishing Company.

Ercikan, K., & Pellegrino, J. W. (Eds.). (2017). *Validation of scoring meaning for the next generation of assessments: The use of response process*. Routledge.

Erickson, K. A., Karvonen, M., & Shin, N. (2024). *Access to the general education curriculum for students who communicate using only non-symbolic modes*. [Manuscript in preparation].

Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177 (2015–2016). https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

Gholson, M., Sova, L., Hauck, M. C., Howley, L., & Albee, T. (in press). Extending principles of evidence-centered design for diverse populations: K-12 English learners with the most significant cognitive disabilities. *Assessment in Education.*

Goldstein, J., & Behuniak, P. (2011). Assumptions in alternate assessment: An argument-based approach to validation. *Assessment for Effective Intervention, 36*(3), 179–191. https://doi.org/10.1177/1534508410392208

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5–9. https://doi.org/10.1111/j.1745-3992.1999.tb00276.x

Hager, K. D., & Slocum, T. A. (2008). Utah's alternate assessment: Evidence regarding six aspects of validity. *Education and Training in Developmental Disabilities, 43*(2), 144–161. https://www.jstor.org/stable/23879926

Johnstone, C., Altman, J. R., & Moore, M. (2011). Universal design and the use of cognitive labs. In M. Russell & M. Kavanaugh (Eds.), *Assessing students in the margin: Challenges, strategies, and techniques* (pp. 425–442). Information Age Publishing.

Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners* (Technical Report 44). University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/OnlinePubs/Tech44/TechnicalReport44.pdf

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Praeger.

Ketterlin-Geller, L. R. (2008), Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practice, 27*(3), 3–16. https://doi.org/10.1111/j.1745-3992.2008.00124.x

Kleinert, H. L., Browder, D. M., & Towles-Reeves, E. A. (2009). Models of cognition for students with significant cognitive disabilities: Implications for assessment. *Review of Educational Research*, *79*(1), 301–326. https://doi.org/10.3102/0034654308326160

Leighton, J. P. (2013). Item difficulty and interviewer knowledge effects on the accuracy and consistency of examinee response processes in verbal reports. *Applied Measurement in Education, 26*(2), 136–157. https://doi.org/10.1080/08957347.2013.765435

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

Padilla, J.-L., & Leighton, J. P. (2017). Cognitive interviewing and think aloud methods. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 211–228). Springer International Publishing. https://doi.org/10.1007/978-3-319-56129-5_12

Thomas, E., Pinilla, R. K., Ketterlin-Geller, L. & Hatfield, C. (2023). Iterative cognitive interview design to uncover children's spatial reasoning." *Practical Assessment, Research, and Evaluation, 28*(1), 12. doi: https://doi.org/10.7275/pare.1918

Thurlow, M. L., Wu, Y., Quenemoen, R. F., & Towles, E. (2016). *Characteristics of students with significant cognitive disabilities: Data from NCSC's 2015 assessment* (NCSC Brief No 8). University of Minnesota, National Center and State Collaborative. http://www.ncscpartners.org/Media/Default/PDFs/Resources/NCSCBrief8.pdf

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

Corresponding Author:
Meagan Karvonen
University of Kansas: ATLAS
1122 West Campus Rd., Lawrence, KS 66045
Email: karvonen@ku.edu