Validity Evidence to Support Alternate Assessment Score Uses: Fidelity and Response Processes

Meagan Karvonen, Russell Swinburne Romine, and Amy K. Clark

University of Kansas

Paper presented at the 2016 annual meeting of the National Council on Measurement in Education, Washington, DC. This paper was developed under grant 84.373X100001 from the U.S. Department of Education, Office of Special Education Programs. The views expressed herein are solely those of the author(s), and no official endorsement by the U.S. Department should be inferred. Correspondence concerning this paper should be addressed to Meagan Karvonen, Center for Educational Testing and Evaluation, University of Kansas, karvonen@ku.edu. Do not redistribute this paper without permission of the authors.

Abstract

Validity of score interpretations and uses for new computer-based alternate assessments (AA-AAS) for students with significant cognitive disabilities require new sources of evidence about student-item interactions and the influences teachers have on those interactions. In this paper we present methods and findings from student cognitive labs, teacher cognitive labs, and test administration observations for a computer-based AA-AAS first administered in 2014-15. The paper concludes with a discussion on how the findings inform future test development and reflections on the use of these research methods for gathering validity evidence for an AA-AAS.

Validity Evidence to Support Alternate Assessment Score Uses: Fidelity and Response Processes

Alternate assessments based on alternate achievement standards (AA-AAS) are large-scale assessments designed for students with the most significant cognitive disabilities. These are students who cannot meaningfully participate in general education assessments even with accommodations. Alternate assessments were first implemented in most states in 2001 after IDEA 1997 required that they be made available. Shortly thereafter, NCLB (2002) required that alternate assessments be based on grade-level academic content standards but with alternate performance standards to the general education assessments.

In the first decade and a half of their existence, alternate assessments usually took the form of portfolios, rater checklists, or structured performance tasks. Altman et al. (2010) surveyed states about their AA-AAS and found that 25 states reported using a portfolio or body of evidence, 23 used performance tasks, 8 used multiple-choice responses, and 7 states were in the process of revising their AA-AAS. There are trade-offs in the choice of alternate assessment design (Hess, Burdge, & Clayton, 2011). For example, performance tasks support standardized administration and allow for more evidence (multiple items) across a broader range of academic content, but the design of the tasks may present barriers in students' ability to demonstrate their knowledge and skills. Portfolios, with content identified and evidence selected by the teacher, are flexible and tend to allow more accessibility supports, but they tend to have fewer entries and a more limited sampling of the content domain. Compared with general education assessments, teachers have a significant influence on AA-AAS results – whether that be through the evidence included in portfolios, the administration process for performance assessments, or the responses for checklists.

Alternate assessments are designed for a small but very heterogeneous population. The most frequent disability labels for students who participate in AA-AAS include intellectual disabilities, autism, or multiple disabilities (Nash, Clark, & Karvonen, 2015; Thurlow et al., 2016). AA-AAS participants also have varied complexity and modes of communication. For example, in a 2013 census of more than 40,000 students identified by their teachers as being eligible for AA-AAS, an estimated 76% of students use speech to meet their expressive communication needs (Nash et al., 2015). Of the 24% of students who do not, 71% combine three or more words according to grammatical rules. The remaining 29% only use one or two words at a time. Students who use symbols or signs instead of speech tend to use only one or two at a time. Among students who do not yet have speech, sign language, or augmentative and alternative communication (AAC) systems, nearly half (48%) use conventional gestures or vocalizations

to communicate intentionally, 14% use only unconventional vocalizations, gestures, or body movement to communicate intentionally, and 38% exhibit behaviors that are not intentionally communicative but may be interpreted by others as such.

Students with significant cognitive disabilities also have sensory and physical challenges that must be addressed for effective assessment. In the same census study (Nash et al., 2015), teachers reported that 19% of students use an AAC device, 7% are also blind or have low vision, and 5% are also deaf or hard of hearing. One-third (33%) have a health or care issue that interferes with instruction or assessment.

The combination of assessment design and student characteristics has required a deliberate approach to developing standards and methods for evaluating evidence of technical quality of AA-AAS. Many techniques used for decades in general education assessments do not transfer well to AA-AAS. For example, a state with fewer than 1,000 students across all tested grades participating in AA-AAS may require small-sample statistical techniques to evaluate comparability of results across similar subgroups. AA-AAS development, administration, reporting, and evaluation should still be guided by the professional *Standards* (AERA, APA & NCME, 2014). But evidence of assessment system quality does not always conform to standard practices in large-scale assessment.

Validity evidence to support intended score uses for AA-AAS is one area that has received some attention in the literature. Several authors have described ways of framing validity arguments and provided examples of validity evidence for alternate assessments (Goldstein & Behuniak, 2011; Marion & Pellegrino, 2006; Marion & Perie, 2009; Perie & Forte, 2011). Early validity studies provided evidence of interrater reliability as a prerequisite for validity (Crawford, Tindal, & Carpenter, 2006 Tindal et al., 2003) and content-related evidence such as the impact of opportunity to learn on AA-AAS outcomes (Karvonen & Huynh, 2007). Construct-related evidence came in the form of internal consistency and factor analysis (Crawford et al.; Tindal et al.) and analysis of AA-AAS results in relation to external variables (Elliott, Compton, & Roach, 2007). Although there are some examples of fully developed validity arguments for AA-AAS (e.g., Goldstein & Behuniak, 2011), not all sources of evidence are easy to collect, and states must weigh several factors when deciding which studies to pursue (Marion & Perie, 2009).

Student-Item Interactions in Next Generation AA-AAS

Two multi-state consortia, Dynamic Learning Maps and the National Center and State Collaborative, have recently developed new, computer-based AA-AAS based on rigorous college and

career ready standards. Both assessments still have a higher degree of teacher involvement than general education assessments, and both have made advances in their design that open up new opportunities for thinking about validity evidence. For instance, the transition to computer-based AA-AAS with multiple-choice items raises questions about how students interact with items in order to demonstrate their knowledge and skills.

Evidence of response process. Evidence of student response process can help test developers understand student-item interactions. Although response process evidence has been included in previous AA-AAS validity studies, it has focused on evidence of how teachers interpret and rate students on a skills checklist type of AA-AAS (Goldstein & Behuniak, 2011). We are aware of no published studies that gathered response process evidence directly from students with significant cognitive disabilities. In fact, Johnstone, Bottsford-Miller, & Thompson (2006) included students with cognitive disabilities (not necessarily *significant* cognitive disabilities) in cognitive labs to evaluate item features and discovered that while students with other types of disabilities were able to participate, those with cognitive disabilities had difficulty verbalizing succinctly and understanding what was expected of them. Challenges with working and short-term memory, as well as meta-cognition, are common for students with significant cognitive disabilities (Kleinert, Browder, & Towles-Reeves, 2009). By meeting the criterion to be included in AA-AAS, students have significant cognitive disabilities that impact their ability to participate in this type of research. Yet to support assertions that knowledge and skills demonstrated on an assessment reflect students' true abilities, 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).

Implementation fidelity. To account for the teacher's role in supporting and mediating student-item interactions, evidence of implementation fidelity may be useful. For example, Hager & Slocum (2008) reviewed six sources of validity evidence for a performance-based AA-AAS and included evidence related to the test administration process. Trained raters evaluated video recorded administrations for evidence of fidelity during the administration of a task. While there were relatively high rates of fidelity reported for some criteria (e.g., prescribed directions were presented in 97% of math and 98% of ELA tasks), complete fidelity to all expectations was relatively low (39% of math tasks and 60% of ELA tasks). Since many AA-AAS have some degree of flexibility in their administration by design, evaluating implementation fidelity for AA-AAS requires consideration of "intended variability" – i.e., the question of standardization versus flexibility (Marion & Pellegrino, 2006, p. 53).

The shift to computer-administered AA-AAS and selected-response item types requires re-conceptualization of the evidence needed to support score uses for AA-AAS – particularly evidence related to the administration process and student interaction with items. The purpose of this paper is to illustrate methods for collecting and evaluating validity evidence for a new, computer-based alternate assessment system. We tie the data sources to specific assumptions in the validity argument, and describe the data collection methods and findings to date. The paper concludes with a reflection on the methods and areas for future research.

DLM Alternate Assessment System

The Dynamic Learning Maps Alternate Assessment System (DLM), used by a multi-state consortium, features assessments in English language arts, mathematics, and science in grades 3-8 and high school. Assessments are delivered as a series of testlets. Each testlet contains a non-scored engagement activity and 3-8 items. There are two modes of delivery to the student. The delivery mode for each testlet depends on the content, and the system delivers testlets at various levels of complexity using an adaptive process.

In about 80% of the testlets, students interact directly with the computer, using human and technology-delivered supports as needed. Item types used in computer-administered assessments include single-select multiple choice and multiple-select multiple choice as well as several other technology-enhanced item types. These include select text items, which direct students to select an answer from a passage taken from a text; sorting items in which students sort items or objects into categories; and matching items in which students match items from two lists.

The remaining 20% of testlets are also delivered via computer but teachers use the online content to guide delivery of performance tasks and to record the student's responses as expressed offline. Item types include single-select multiple choice and multiple-select multiple choice. These teacher-administered testlets are typically used at lower levels of complexity for students who are still working toward 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. Similar to other DLM teacher-administered testlets, in the writing testlets the test administrator evaluates student writing processes and products offline and enters responses into the online system.

Design of the DLM assessment system was guided by principles of universal design for assessment. Consistent with these principles (cf. Ketterlin-Geller, 2008), much of the assessment system was designed for flexible administration. For example, the timing and length of a test session, the choice of test setting and device, and the use of adaptive equipment are all part of routine options that the test administrator has available. When making decisions about using additional supports for computer-delivered testlets, educators are encouraged to follow these two general principles: first, the student should respond to the content independently and second, the student should be familiar with the chosen supports because they have been used consistently during routine instruction. This means providing the same support, or a very similar one, during the student's computer-based classroom instruction. When making decisions about additional supports for teacher-administered testlets, educators are encouraged to provide flexibility in student access and response mode. This means that students should be able to indicate responses using their regular communication strategies and that the test administrator has flexibility to change typical administration setup and physical arrangement based on a student's physical needs and use of special equipment. 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 all complete the same task. Questions cannot be rephrased and items cannot be rearranged.

Furthering the guiding philosophy of accessibility by design, universal supports are available to any student but must be selected by the teacher in the Personal Needs and Preferences (PNP) profile. Examples include online supports such as magnification and synthetic spoken audio, communication switches, human read aloud, and the use of individualized manipulatives. Teachers are trained on how to administer both types of testlets with fidelity and how to make decisions about PNP supports to provide and options for flexibility in test administration. Teachers must complete required training and pass quizzes about the contents of the modules in order to be eligible to administer DLM assessments.

DLM Validity Argument

The DLM consortium uses an argument-based approach to validity and includes claims that support the intended uses of these scores. Two such claims focus on student demonstration of knowledge as items are administered.

1. The assessments have been designed to allow students to demonstrate their knowledge and skills in relation to academic expectations.
2. Teachers administer the assessments with fidelity so that students can respond to the items as intended.

Each of these claims has multiple assumptions to evaluate. For example, the engagement activities, images, and manipulatives should function as intended and should not distract or create barriers during the response process. Students should be able to respond regardless of disability, health, or other constraints. Regarding test administration practices, teachers are expected to select appropriate supports students need to respond to tasks and allow students to respond as independently as possible. When teachers must enter student responses on their behalf, the entries should accurately reflect the student's demonstration of the skill. And when teachers administer assessments offline, they must correctly interpret the instructions and administer the assessments with fidelity.

To investigate several of these assumptions, multiple sources of evidence are collected during the test administration process. Methods used to collect this evidence draw from some 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 test administration observations, cognitive labs conducted with students, and cognitive labs conducted with teachers. The relationships between validity argument assumptions and types of evidence collected is summarized in Figure 1.

Figure 1. Evidence associated with assumptions in the DLM assessment system validity argument

Assumption from Validity Argument	Type of Evidence Collected	
	Cognitive Lab	Test Administration Observation
Educators allow students to engage with the system as independently as they are able		✓
Students are able to interact with the system as intended	Student	✓
Students are able to respond to tasks irrespective of a sensory, mobility, health, communication, or behavioral constraint	Student	✓
Optional supports are used effectively by the student and don't distract	Student	✓
Teachers enter student scores/responses with fidelity	Teacher	✓

Validity Evidence

Test Administration Observations

Test administration observations were conducted in multiple states during field testing in spring 2014 and operational assessments in spring 2015 and 2015-16. We observed the student's typical test administration process, with the student's actual test administrator. We observed administrations for the full range of students eligible for DLM assessments (i.e., those with the most significant cognitive disabilities).

DLM uses a test administration observation protocol to gather information about how educators in the consortium states deliver testlets to students with significant cognitive disabilities. This protocol gives observers a standardized way to describe the way a DLM testlet was administered – no matter their role or experience with DLM assessments. The test administration observation protocol captured data about student actions (navigation, answering), teacher assistance, variations from standard administration and engagement & barriers to engagement. Test administration observations are collected by DLM project staff, as well as SEA and LEA staff. The observations protocol is only used for descriptive purposes. It is not used to evaluate or coach the teacher, or to monitor student performance. 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.

During computer-administered testlets, the intent is that 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, delivering it to the student, and recording responses in the KITE system. The test administration protocol contains different questions specific to each type of testlet.

Test administration observations were collected in 5 states beginning in 2015 through February, 2016. The numbers of observations collected by state are shown in Table 1.

Table 1. Observations by State (N=147)

State	n	%
Alaska	5	3.4
Iowa	45	30.6
Kansas	1	0.7
Mississippi	1	0.7
Missouri	95	64.6

Of the 147 test administration observations collected, 117 (79.6%) were of computer delivered assessments and 30 (20.4%) were of teacher-administered testlets. Of the 147 observations, 70 (47.6%) were of English language arts (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 were administered in students' usual classrooms (81.6%).

To investigate the assumptions that underlie the claims of the validity argument, several parts of the test administration observation protocol corresponded to assumptions. One assumption addressed is: "educators allow students to engage with the system as independently as they are able." For computer-administered testlets, related evidence is found in five items in Table 2. Bold items represent supporting evidence and italicized items represent non-supporting evidence. For example, clarifying directions (26% of observations) removes student confusion over the task demands as a 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 is a clear indicator that the teacher directly influenced the student's answer choice.

Table 2. Test Administrator Actions during Computer-Administered Testlet (N=117)

Action	n	%
Navigated one or more screens for the student	85	72.6
Repeated question(s) before student responded	76	65.0
Repeated question(s) after student responded	11	9.4
Reduced number of choices available to student	6	15.1
Used verbal prompts to direct the student's attention	**65**	**55.6**
Used physical prompts	*30*	*25.6*
Clarified directions	**30**	**25.6**
Defined vocabulary used in the testlet	34	29.1
Asked the student to clarify one or more responses	10	8.5

*Respondent could select multiple responses to this question

For DLM assessments, 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 is not necessarily an indication that 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. While not the same as interfering with students' interaction with the content of assessment, navigating for students who are able to do so independently would be counter to the

assumption that students are able to interact with the system as intended. The observation protocol did not capture the reason the teacher chose to navigate, and the reason was not always obviously inferred just from watching.

Related to the assumption that educators allow students' independent engagement with the system is another assumption: "students are able to interact with the system as intended." Evidence for this assumption was gathered by observing students taking computer delivered testlets. Again, bold items represent supporting evidence and italicized items represent non-supporting evidence 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 that was recorded separately) was seen in 21% of the observations. Verbal prompts for navigation and response selection are strategies that are within the realm of what is allowable flexibility during test administration. Although these strategies would 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 indicate that students were not able to sustain independent interaction with the system.

Table 3. Student Actions during Computer-Administered Testlets (N=117)

Action	n	%
Navigated the screens independently	**19**	**16.2**
Navigated the screens with verbal prompts	*8*	*6.8*
Selected answers independently	**45**	**38.5**
Selected answers with verbal prompts	*53*	*45.3*
Indicated answers using eye gaze	**24**	**20.5**
Indicated answers using materials outside of KITE	4	3.4
Used manipulatives	30	25.6

*Respondent could select multiple responses to this question

Another assumption, "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 noted difficulty in 2 (6.7%) cases. For computer delivered testlets, evidence to evaluate this assumption was observed by noting students' abilities to indicate answer to items using multiple response modes such as sign language, eye gaze, and using manipulatives or materials outside of KITE. A summary of the frequencies of these behaviors is shown in Table 4. Additional evidence for this

assumption 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.

Another assumption underlying the claims is that "teachers enter student responses with fidelity." Observers rated whether test administrators accurately captured student responses. In order to record student responses with fidelity, test administrators needed to observe multiple modes of such as verbal, gesture, and eye-gaze. Table 4 summarizes students' response modes for teacher-administered testlets.

Table 4. Primary Response Mode for Teacher-Administered Testlet (N=30)

Response mode	n	%
Verbal	7	23.3
Gesture	12	40.0
Eye gaze	2	6.7
Other	6	20.0
No response	5	16.7

*Respondent could select multiple responses to this question

Across all observations and student response modes, test administrators recorded responses with fidelity in 93.3% of observations.

Computer-administered testlets provided another opportunity to confirm fidelity of response entry when test administrators entered responses on behalf of students. This is a support recorded on the PNP 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, the test administrator entered responses on the student's behalf. In 98.6% of those cases, observers indicated that the entered response matched the student's response. This evidence supports the assumption that teachers entered student responses with fidelity.

Student Cognitive Labs

Cognitive labs are typically used to elicit statements that allow the observer to know whether the item is tapping the intended cognitive process (Ericsson & Simon, 1993). Due to the challenges in getting students with cognitive disabilities to verbalize in this manner (Altman et al., 2006), we instead framed this study as indirect evidence of response process because it allowed us to evaluate whether the item response demands were introducing construct-irrelevant variance.

With a move to computer-based testing, many assessment programs have introduced technology-enhanced items. When designing the DLM assessments, we considered the potential trade-offs of these new item types. On one hand, these items offer a means of assessing student knowledge using fewer items, which minimizes testing burden on a population that has difficulty with long tests. For example, a student's ability to classify objects could be assessed through a series of multiple choice items or through one item that involves sorting objects into multiple categories. Our potential concern with technology-enhanced item types was that they would be challenging for students with significant cognitive disabilities – in terms of their cognitive demands, lack of familiarity, and physical access barriers related to students' fine motor skills.

This phase of cognitive labs evaluated whether the technology-enhanced item types themselves introduced construct-irrelevant variance by creating challenges or confusion during the process of answering the item. Labs were conducted with 27 students from multiple states in spring 2014 and spring 2015. Eligible students were from tested grades (3-8 and HS) and had sufficient symbolic communication systems to be able to interact with the content of on-screen items without physical assistance, through keyboard/mouse, tablet, or other assistive technology. Inclusion criteria also required they have some verbal expressive communication and were able to interact with the testing device without physical assistance.

Labs focused on student interaction with four types of technology enhanced-items, including drag and drop (DD), click to place (CP), select text (ST), and multi-select multiple choice (MSMC) item types. The first three item types were designed specifically for DLM assessments and are delivered through a user interface designed for this population. DD and CP items are used for sorting. The difference between them is that DD requires continuous selection (clicking and dragging) while CP items require clicking on the origin and clicking on the intended destination. The latter item type is accessible for switch users, but one theory was that non-switch users would also find clicking without dragging to be easier since the process was less demanding on fine motor skills. Both the DD and CP items were built to require a similar response process, sorting objects into categories. To facilitate comparisons with DD and CP items, MSMC items were also constructed to access a response process requiring the student to select the answer options that matched a category. ST items are only used in some English language arts assessments. In an ST 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, Altman, & Moore, 2011), we constructed 4-item testlets with content that did not rely on prior academic knowledge. For example, while students who might be candidates for cognitive labs are highly likely to know their shapes, completing an item with shapes did not require an understanding of specific shapes (see Figure 2).

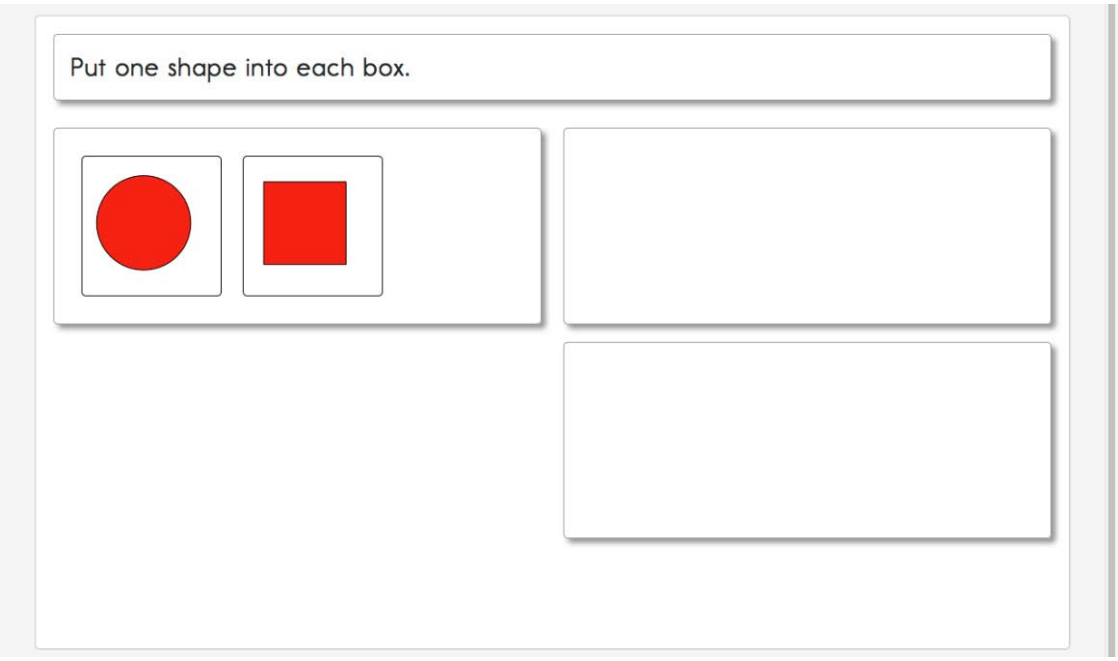

Figure 2. Sample DD item from cognitive lab.

Figure 3 shows a ST item that was similarly constructed to minimize the need for prior knowledge.

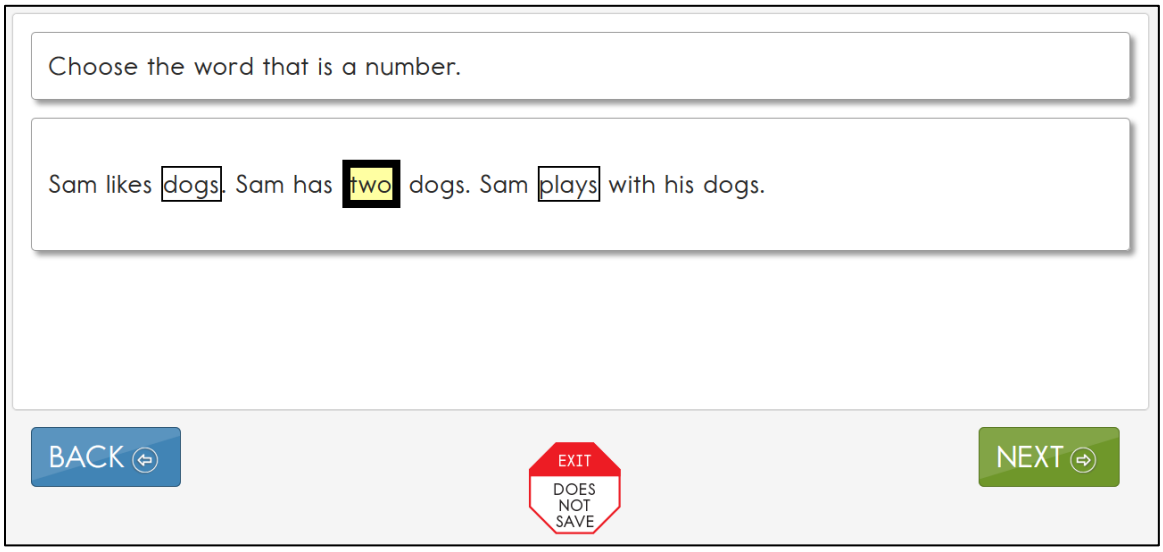

Figure 3. Sample ST item from cognitive lab.

Each testlet contained one type of item. For SD and DD item types, 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) and testlet assignments were counter-balanced. Fifteen students completed DD, 11 completed CP, 8 completed ST, and 11 completed MSMC testlets. The 8 students who completed ST testlets also completed a testlet that used the same content as the ST items, but presented in a traditional, single-select multiple choice format.

For each item type, the examiner looked for evidence of challenge with each step of the item completion process (e.g., for DD items, 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 additional 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 Altman et al. (2011 – e.g., "what makes you believe that answer is the right one?") and only required 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 parallel instructional practice for many students who are eligible for AA-AAS.

Videos were reviewed to confirm that the ratings of potential sources of challenge were correctly recorded. Data analysis reported in this paper consists of descriptive statistics for items in the observation protocol and frequency distributions for students' responses to interview questions.

Sources of challenge in responding to DD and CP item types were demonstrated when students had difficulty selecting the desired object, difficulty maintaining continuous selection, difficulty with group selection, or with number of objects. In general, students tended to have more difficulty with CP items than DD items (see Table 4).

Table 4. Sources of challenge in responding to drag-and-drop (DD) and click-to-place (CP) item types

	DD (n=15 students, 60 items)		CP (n=11 students, 44 items)	
Source of Challenge	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

Sources of challenge in responding to MSMC items was examined by observing difficulty with the selection of the first object, the subsequent object(s), the concept of needing to make more than one selection and needing assistance to complete the item. A summary of the sources of challenge in responding to MS items is shown in table 5. On 41% of the items, students had difficulty with the concept of making multiple selections.

Table 5. Sources of challenge in responding to multi-select multiple choice (MSMC) items (N = 11 students, 44 items)*

Source of Challenge	n	%
Difficulty with selection of first object	4	9.0
Difficulty with selection of subsequent objects	6	13.6
Difficulty with multi-select concept	18	40.9
Needed assistance to complete	9	20.5

*1 testlet not completed

ST items required less manipulation of on-screen content and only one selection to respond to the item. Across 8 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.

Finally, Table 6 summarizes student responses to post-hoc interview questions. DD and ST items were more often liked, perceived as easy, and required a response process that students understood.

MS items were viewed less positively and students reported the most difficulty with CP items. Rough rankings of item effectiveness based on sources of challenge noted by the observers were consistent with student interview responses.

Table 6. Affirmative student responses to post-hoc interview questions

	DD (n = 15)		CP (n = 11)		MS (n = 11)		ST (n = 8)	
Question	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

Teacher Cognitive Labs

Teacher cognitive labs are a potential source of response process evidence that have been recommended for AA-AAS where teacher ratings are the items (e.g., Goldstein & Behuniak, 2010). We used this approach for DLM teacher-administered testlets since teachers interpret student behavior and respond to items about the student's response. Most of these testlets involve teacher interpretation of responses for students who are working on consistent, intentional communication and who are working on foundational skills that promote their access to grade-level content. Writing testlets are also teacher administered at all levels of complexity.

Teacher cognitive labs were conducted in spring 2015 with 15 teachers in 5 schools across 2 states. Teachers completed think-aloud procedures while preparing for and administering teacher-administered testlets in reading, writing, and math. They were first presented with the Testlet Information Page (TIP), which is a short document that provides background information needed in order to prepare to administer the testlet. For example, a TIP may contain instructions about materials needed, guidelines 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 specific math testlet.

Teachers were asked to think out loud as they read through the TIP. 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 labs were video recorded and an observer took notes during the administration. The initial phase of analysis involved recording evidence of intended administration and sources of challenge to intended administration at each of the following stages: (1) preparation for administration, (2) interpretation of educator directions within the testlet, (3) testlet administration, (4) interpretation of student behaviors, and (5) recording student responses. Through this lens, we were able to look for evidence related to fidelity (1, 2, 3, and 5) as well as response process (4).

These 15 labs were the first phase of data collection using this protocol. Preliminary evidence on interpretation of student behaviors indicates 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 where the student touched the object rather than handing it to the teacher, the teacher accepted that response and entered it, but speculated as to 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 used eye gaze, the teacher held objects within the student's field of vision and put the correct answer away from the current gaze point so that a correct response required intentional eye movement to the correct object.
- 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 in the Test Administration Manual.
- In one case, the teacher's movement of objects to prepare for the next item led her attention away from the student and caused her to miss his eye gaze that indicated a response. She recorded "no response." However, this was observed for a student whose communication and academic skills were far beyond what was being assessed. The testlet was not appropriate for this student and his typical response mode for DLM testlets was verbal.

Analysis of teacher cognitive lab data is ongoing. Strengths and drawbacks of the method are described in the discussion section below.

Summary

Table 7 summarizes overall findings from the three studies described in this paper. Organizing these findings according to the associated assumptions helps us see how related evidence across data sources may be synthesized as we document the evidence in the technical manual.

Table 7. Findings associated with assumptions in the validity argument

Assumption from Validity Argument	Evidence
Educators allow students to engage with the system as independently as they are able	Teachers entered student responses in a majority of the observations. 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.
Students are able to interact with the system as intended	Overall, students were able to successfully complete two of the four types of technology-enhanced items.
	Some students were able to navigate computer-delivered assessments independently and select answers without support.
	Students used a variety of response modes to indicate selection of answers on computer-delivered testlets.
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, students did not experience difficulty using supports.
	In observations of computer delivered testlets, students were able to use different response modes such as verbal, gesture and eye-gaze.
	In the majority of observations, students were able to complete the testlet.
Optional supports are used effectively by the student and don't distract	In the majority of observations of teacher-administered testlets, students did not experience difficulty using supports.
Teachers enter student scores/responses with fidelity	In almost all observations, teachers who entered responses on a student's behalf chose responses that matched the student's behavior.
	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.

Discussion

Taken together, the evidence presented here provides preliminary evidence in support of claims that the DLM assessments have been designed so students can show what they know and can do, and that teachers administer assessments in such a way that allows students to respond as intended. These studies part of a larger body of evidence, including procedural and empirical data, and research is ongoing. We have not yet been able to collect evidence that would allow us to investigate some assumptions. For example, evidence that teachers choose and implement appropriate supports is currently limited to teacher self-report on an annual survey. With the exception of human read aloud, we have not seen the use of PNP features during routine test administration. Observational research is time intensive and samples are small, but we will need to do targeted recruitment to find teachers whose students use a broader array of supports during assessment.

Although preliminary in nature, results described in this paper have already informed improvements in test development and resources to support test administration. For example, we use technology-enhanced items sparingly and have added guidelines about how different item types may and may not be combined in one testlet. Click-to-place items will only be delivered to switch users since this item type did not have the anticipated benefits to all students. Similarly, when we observed teachers follow administration procedures that represented poor or mixed fidelity, their misconceptions that led to those actions have informed revisions to our test administration manual and required test administrator training.

These studies also point to areas for improvement in our data collection protocols. For example, while it is valuable to be able to observe teacher intervention (e.g., navigation, response entry) during computer-administered testlets, we did not record the reason for their intervention. To support accuracy and reliability of observational data, we originally did not want to require observers to make inferences about behaviors they saw. But the reason for navigating or entering responses on the student's behalf determines whether their action supports or inhibits students' independent interaction with the system. The same was true for checklist items about clarifying student answers and repeating the question after an answer was given. Depending on the situation, these could be ways of ensuring that the response they enter does indeed reflect the student's intended response. Or, this could be evidence of trying to get the student to change an answer – a practice that is not allowed. Future

revisions of the observation protocol will need to consider all of the purposes of the tool (which go beyond the purposes in this paper) and the balance between making the questions concise and easy to collect versus being informative enough to support our information needs. A limited number of neutrally worded post-hoc interview questions may be useful.

Our experience with the teacher cognitive labs has been mixed. We conducted these in an authentic environment with real students, hoping to capture teachers' thought processes while problem-solving during administration to students who are challenging to assess. But we relied on teachers to select students for these labs, and they tended to choose students who were not necessarily appropriate for the complexity of the testlets under investigation. Their choices of students removed some of the problems with interruptions during testing for in vivo probes (e.g., student behavior and health issues) but made for a less authentic experience. We may try additional labs that remove students from the room and instead have the teacher think aloud during the testlet while thinking about a hypothetical student. This approach may provide richer verbalizations but less authentic and immediate responses. Another possibility would be to have teachers respond to video recorded sessions and pause periodically to ask what their next steps would be, and why, if they were the ones assessing the student. If they also indicated how they would answer an item based on a student's behavior, this approach could also be a way of checking interrater agreement.

Student cognitive labs also generated some lessons learned about methodology. In some respects, the protocol allowed for straightforward data collection on the construct-irrelevant parts of item response. However, even with simplified interview questions, the fact that at least one student in each condition rated a testlet as both easy and hard is evidence of unreliability of student self-report. Also, we were not able to gather information about how the student was interpreting the on-screen contents. Future use of eye movement tracking software could expand our understanding of students' response processes without requiring verbalization.

Besides lessons learned about the three specific data collection methods, this work has also led us to think more generally about AA-AAS 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 1.5 years of operational assessment. As we shift into a more neutral stance and also consider the challenges of data collection, the growing body of evidence is likely to rely in part on inverse logic or counterevidence. For example, with cognitive labs, we were not able to collect

confirmatory evidence that students were using the intended cognitive process. Instead, we evaluated the possibility that construct-irrelevant item features would negatively impact the student-item interaction. Especially where the least plausible assumptions in the validity argument intersect with the most complex data collection, more work is needed.

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*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://www.cehd.umn.edu/NCEO/OnlinePubs/StateReports/2009_survey_of_states.htm

American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Crawford, L., Tindal, G., & Carpenter, D. M., II. (2006). Exploring the validity of the Oregon extended writing assessment. *Journal of Special Education, 40*, 16-27.

Elliott, S. N., Compton, E., & Roach, A. T. (2007). Building validity evidence for scores on a state-wide alternate assessment: A contrasting groups, multimethod approach. *Educational Measurement: Issues and Practice, 26*, 30-43.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data (Revised edition).* Cambridge, MA: MIT Press.

Goldstein, J., & Behuniak, P. (2010). Assumptions in alternate assessment: An argument-based approach to validation. *Assessment for Effective Intervention, 36*, 179-191.

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

Hess, K., Burdge, M., & Clayton, J. (2011). Challenges to developing and implementing alternate assessments based on alternate achievement standards (AA-AAS). In M. Russell & M. Kavanaugh (Eds.), *Assessing students in the margins: Challenges, strategies, and techniques* (pp. 171-213). Charlotte, NC: Information Age Publishing.

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). Charlotte, NC: Information Age Publishing.

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

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

Karvonen, M., & Huynh, H. (2007). The relationship between IEP characteristics and test scores on alternate assessments for students with significant cognitive disabilities. *Applied Measurement in Education, 20*(3), 273-300. doi: 10.1080/08957340701431328

Karvonen, M., Wakeman, S. L., & Kingston, N. (in press). Alternate assessment. In *Research Based Practices for Educating Students with Intellectual Disability.*

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*, 3–16.

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*, 301–326.

Marion, S. F., & Pellegrino, J. W. (2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice, 25*, 47–57.

Marion, S. F., & Perie, M. (2009). An introduction to validity arguments for alternate assessments. In W. D. Schafer & R. W. Lissitz (Eds.), *Alternate assessments based on alternate achievement standards* (pp. 113-125). Baltimore, MD: Paul H. Brookes Publishing.

Nash, B., Clark, A. K., & Karvonen, M. (2015). *First contact: A census report on the characteristics of students eligible to take alternate assessments*. Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation.

Perie, M., & Forte, E. (2011). Developing a validity argument for assessments of students in the margins. In M. Russell & M. Kavanaugh (Eds.), *Assessing students in the margin: Challenges, strategies, and techniques* (pp. 335-378). Charlotte, NC: Information Age Publishing.

Tindal, G., McDonald, M., Tedesco, M., Glasgow, A., Almond, P., Crawford, L., & Hollenbeck, K. (2003). Alternate assessments of students with significant disabilities: alternative approaches, common technical challenges. *Exceptional Children, 69*, 481-494.

Thurlow, M. L., Wu, Y., Quenemoen, R. F., & Towles, E. (2016, January). *Characteristics of students with significant cognitive disabilities: Data from NCSC's 2015 assessment* (NCSC Brief No 8). Minneapolis, MN: University of Minnesota, National Center and State Collaborative.