Gathering Evidence of Response Processes for Alternate Assessments (AA-AAS)

Russell Swinburne Romine, Meagan Karvonen and Amy K. Clark

University of Kansas

Author Note

Paper presented at the 2015 annual meeting of the National Council on Measurement in Education, Chicago, Illinois. 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 Russell Swinburne Romine, Center for Educational Testing and Evaluation, University of Kansas, swin0030@ku.edu. This paper should not be redistributed without permission of the authors.

Abstract

In order to make validity arguments, researchers commonly use cognitive labs as one source of evidence about student response processes. However, there are challenges in collecting such evidence for alternate assessments designed for students with significant cognitive disabilities (AA-AAS). We present findings from cognitive labs and test administration observation sessions for an AA-AAS.

Gathering Evidence of Response Processes for Alternate Assessments (AA-AAS)

Students with significant cognitive disabilities (SWSCD) have been assessed using large-scale alternate assessments based on alternate achievement standards (AA-AAS) since 2001. One significant challenge in the development of AA-AAS is to provide a standardized assessment that has adequate technical quality and minimizes construct-irrelevant variance, while maintaining flexibility regarding the individual access needs of students (Gong & Marion, 2006).

The Standards (AERA, APA & NCME, 2014) require that validity evidence should relate to the extent to which students' actual response processes align with the construct being assessed. Evidence based on response processes is often gathered by analyzing individual responses to test items or tasks. Messick (1995) describes the substantive aspect of construct-related validity as requiring evidence that the cognitive processes associated with the domain of the assessment are actually used by test-takers when they respond to assessment items. Padilla and Benitez (2014) reviewed methods for obtaining validity evidence of response processes which include interviewing, focus groups, and cognitive interviewing. Of particular interest to test developers who wish to support a validity argument are cognitive interviews. Typical methods for cognitive interviews include asking a student to explain his or her thinking while or after doing a task. In these interviews a researcher is then able to ask probing questions to understand the cognitive processes that underlie the student's response to the item or task (Willis, 2005).

The challenges for students with disabilities involved in methods such as cognitive interviews include areas of working memory, difficulty with expressive speech, and metacognitive development (Almond, et al., 2008). For students with significant cognitive disabilities (SWSCD), there are additional challenges associated with obtaining evidence of response process. Findings from a 2013 census of 44,782 students in the alternate assessment population in 14 states revealed 24% of the population does not use speech as a mode of expressive communication. Among those who do use speech to communicate, only 71% regularly combine three or more words to communicate for a variety of purposes. Furthermore, 40% of the students require support, whether human or from assistive technology, to interact with a computer (Dynamic Learning Maps, 2013). Many students who take an AA-AAS have not acquired symbolic communication (Towles-Reeves, et al., 2008). Given the unique and diverse challenges experienced by SWSCDs who take AA-AAS, it is not clear whether cognitive interviews will yield the information needed to evaluate student response processes. Alternate sources of evidence may be required.

All students who take an AA-AAS as a part of their education deserve an assessment that validly measures what they know and can do. As computer-based testing becomes implemented on a wider scale, including technology-enhanced item types, it is important to conduct research on the response processes used by SWSCDs when responding to assessment items. Additionally, research is needed to explore methods of collecting evidence based on response processes specifically with students with significant cognitive disabilities.

Research Questions

The purpose of this paper is to present findings from two related studies on an AA-AAS. Research questions include:

1. What evidence is there that students with significant cognitive disabilities use intended cognitive processes to respond to AA-AAS items?

2. How do students interact with various types of computer-delivered items, including innovative item types?

The Dynamic Learning Maps (DLM) alternate assessment system is a consortium of 17 states that have joined together to develop a computer-based, adaptive assessment for students with significant cognitive disabilities. DLM assessments are based on fine-grained, research-based learning maps. Unlike learning progressions, DLM learning maps are web-like networks of connected learning targets which represent multiple paths to learning objectives. In DLM learning maps, learning targets take the form of individual nodes, and these nodes are all interconnected to reflect how individual skills contribute to and provide the foundation for the development of subsequent skills.

From the beginning, the DLM system was designed to support learning and assessment with accessibility in mind. Students must understand what they are being asked in an assessment item and have the tools to respond in order to demonstrate what they know and can do. One aspect of DLM assessments that supports this goal is differentiated content. Using the structure of the learning map, DLM created testlets, which are short assessments each containing an engagement activity and 3-8 items. Testlets are available at five linkage levels for each content standard (Essential Element). Each linkage level provides access to content on key knowledge, skills, and understandings on the path to the grade level expectation for students with significant cognitive disabilities.

Most DLM testlets are designed for direct student interaction with the content via computer. In some instances, the student may need support from the test administrator to interact with the computer. Other testlets are designed to be delivered, outside the system, with the test administrator using instructions presented online to administer the assessment and record responses in the system. The recorded responses in these teacher-administered testlets require the test administrator to observe student responses to tasks completed outside the system and in some cases, evaluate the student's response according to a set of descriptions provided as answer options in the system, not unlike using a rubric.

In order to support the valid interpretation and use of scores from the assessment, the DLM validity argument includes several propositions related to student engagement with the assessment system. This includes assumptions that educators allow students to engage with the system as independently as they are able, that students are able to interact with the system as intended, and that teachers enter student scores/responses with fidelity. Because the DLM system is among the first computer-based platforms to be used to assess students with significant cognitive disabilities, research on evidence of response processes in this population is necessary.

This paper presents preliminary data from test administration observations of computer-delivered English language arts (ELA) and math testlets, and test administration observations of teacher-administered ELA writing testlets. Additionally, data from cognitive labs is presented which examines student use of technology-enhanced items delivered through the online testing platform.

Methods

This paper describes the use of two methods of data collection to answer the research questions. The first method, a test administration observation, included a researcher using standardized protocols to collect data about a teacher and student's experiences with test administration. One version of this method included open-ended questions to the test administrator about his or her experience. The second method used a cognitive lab in which a researcher sat with a student as he or she navigated a computer-delivered testlet created specifically to provide an experience with different types of items. Both methods can potentially yield information about student response processes.

Test Administration Observations

Test administration observations were conducted in 6 schools from 2 states during a spring 2014 field test. Eligible students were from tested grades (3-8 and HS) including the full range of students eligible for AA-AAS. There were no special inclusion criteria. Test administrations were conducted by the student's typical test administrator (often a classroom teacher) during the student's typical administration of a field test. A researcher watched the test administration and recorded data using a standardized protocol. One protocol was completed for each testlet administered. An additional set of test administration observations of ELA testlets that assess writing were conducted in 3 schools in 1 state during a spring 2015 field test.

The test administration observation protocol captured data about test administration time, student actions (navigation, answering), teacher assistance, variations from standard administration, and engagement and barriers to engagement. In spring 2014, 22 ELA and 17 math computer-delivered testlets were observed.

In a typical testing session that was observed, a student logged into the online testing system which uses a user interface specifically designed for SWSCDs. In both math and ELA testlets, the student starts the testing session with an engagement activity. The engagement activity provides a context for the assessment and may be useful to help activate prior knowledge related to the construct. In math testlets, this is often a sentence or sentences that describe a situation with a mathematical application and associated graphics. In ELA, the engagement activity involves reading a passage, either a story or an informational text. After the engagement activity, the student is presented with a series of items on screen, using the mouse to select his or her response.

In spring, 2015, 22 ELA writing testlets were observed. A modified protocol used specifically for observation of ELA writing testlets was used to capture data about how students responded to teacher-administered assessments of writing. In the writing testlets, the test administrator navigated the testlet using the online testing system. The writing testlets are constructed to provide instructions to the test administrator about the tasks and prompts to give to the student to guide him or her through a structured writing activity in which the student writes about an informational topic. As the student writes, the test administrator makes judgments about the student's responses to the task and enters responses into the system choosing from a set of descriptions the one that best describes the student's behavior. The protocol used for test administrations of the ELA writing testlets also included data collection about the student's use of a writing tool. The use of writing tools was included to specifically investigate students' use of tools to respond to the writing testlet tasks. Additionally, the modified protocol used to observe writing testlets included additional open-ended questions for the test administrator. The protocol was designed to gather evidence to refine the content and instructions of the testlet, examine response processes used by students and to examine how test administrators were making decisions about how to record student responses using the ratings included in the testlet. Test administrators were asked a question about typical writing instruction for the student. They were also asked what could be done to improve the design of the testlet, and asked to offer general feedback about the writing testlet.

Cognitive Labs

Cognitive labs were conducted with 14 students from 2 states in spring 2014. 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. Inclusion criteria also required they have some verbal expressive communication and were able to interact with the computer without physical assistance, through keyboard/mouse, tablet, or other assistive technology.

The cognitive labs focused on student interaction with item types other than single-select multiple choice. Using content that minimized reliance on the student's prior academic knowledge, 4-item testlets were constructed. Items were administered through the DLM test delivery engine using test design conventions of typical DLM items (e.g., constrained text complexity, use of graphics).

Each testlet contained one type of item: drag and drop (DD), click to place (CP), or multi-select multiple choice (MSMC). 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. Both the DD and CP items were built to require a similar response process, sorting objects into categories. The directions given to the student for the cognitive labs did not require any prior category knowledge. An example of one of the DD items is provided in Figure 1. MSMC items were also constructed to access a response process requiring the student to select the answer options that matched a category.

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. At the end of each testlet, students were asked several yes/no questions about what they had just done. At the end of the session, they indicated whether they preferred the first testlet or the second.

Each student completed two testlets (one per item type) and testlet assignments were counterbalanced. Nine students completed DD, 11 completed CP, and 6 completed MSMC testlets. Each lab was video recorded and an observer recorded student responses and evidence of challenge while the examiner administered the standardized protocol. 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 each source of challenge, per item type, and frequency distributions for students' responses to interview questions.

Results

Computer-Delivered Test Administration Observations

In the test administration observations that were conducted on computer-delivered assessments, initial findings from administration observations show that 28% of the time the teacher navigated all screens for students and 23% of the time the teacher entered responses on the student's behalf. Table 1 shows the categories of assistance observed during administrations of computer-delivered testlets. When test administrators navigated system components on the student's behalf, the observer recorded, when possible the reason for the assistance. This occurred due to student distraction (*n* = 3) or the student's self-injurious behavior (*n* = 1). In five observations the test administrator navigated or entered responses for no apparent reason related to the student. In these instances, the test administrator assumed s/he should be doing things on the student's behalf. The administration observations indicate there was high fidelity of student response entry. In 70% of observations students were rated as highly engaged and in 59% of observations students were rated as having a high level of independence.

Teacher-Administered Test Administration Observations

Students who participated in the test administration observations used a variety of writing tools as a part of their response to the items in the testlet. Table 2 shows the frequency of use of different types of writing tools. Most students used pen/pencil and paper (36%), a traditional keyboard (27%) or an alphabet chart/book (22%).

In the analysis of the open-ended responses from teachers about the design of the writing testlets, there were two major ideas repeated. The first idea was that the directions needed to be simplified on the items in the testlets. In response to "What could be done to improve the design of the testlet to help test administrators, eight respondents included the idea that the directions should be simplified. One test administrator described the directions as "wordy" and another test administrator described "missing a few directions" because too much information was presented on the screen. A second idea focused on the desire for expanded sets of answer options to cover a broader range of observed student performance. One test administrator described wishing there was "another answer option" for some of the testlet items. Another reported that she would have liked "another rating point" on one item.

Cognitive Labs

Of the students who participated in cognitive labs, 11 were tested on the computer with a mouse and 3 were tested via a touch screen or on a SMARTboard. Results are based on 10 students and 40 items for DD testlets, 11 students and 44 items on CP testlets, and 6 students and 23 items[1] on MSMC testlets. In general, students had more difficulty completing CP items than DD items. More students had difficulty with object selection, group selection, and number of objects in CP than DD items.

For MSMC items, object selection was not a source of challenge for most students but the concept of making multiple selections was difficult in the majority of cases (74%). Students most often needed assistance to complete the testlet for CP (61% of items), followed by MSMC (33%) and DD (10%). At the end of each testlet, students were asked several questions. Students tended to like DD items (89%) and find them easy (78%), more so than CP items (58% and 67%, respectively). Students perceived MSMC testlets as easy (83%), despite the fact that students needed assistance to complete one-third of those items.

Among students who received DD and CP (*n* = 7), all preferred DD. Among students who received MSMC and one of the other types (*n* = 6), three preferred MSMC, two preferred the non-MSMC item type, and one did not respond.

[1] One student did not complete the testlet.

Discussion

A validity argument for any large-scale assessment system requires evidence from response processes, yet the communication and metacognition difficulties for students who are eligible for AA-AAS introduce challenges in collecting this type of evidence. In this paper we reported on two types of evidence. The more typical evidence (cognitive labs) was collected for the subset of AA-AAS-eligible students who had sufficient verbal skills to provide evidence. Test administration observations were used to collect evidence that was visible through student behaviors for the entire population of SWSCDs.

In test administration observations of computer-delivered testlets, observers watched and recorded information related to test administrator assistance with navigation and entry. It was clear in this study that test administrator assistance is not automatically interfering with the response process. The frequency and types of assistance are relevant to considering evidence of student response process. In some cases, assistance with navigation may be helpful in allowing a student to demonstrate what he or she knows or can do. This would be true when the student cannot independently indicate a response due to a physical accessibility barrier. In other cases, however, assistance may prevent the student from using a potentially appropriate response processes to respond to the item. This occurred in a limited number of cases in the current study, when the test administrator assumed she should mediate the student-computer interaction even when the student's disability did not prevent independent access.

In the teacher-administered writing testlets, behavioral observations and open-ended questions allowed us to examine the student's response process as well as the teacher's. Where the teacher had difficulty interpreting the instructions and administering the test with fidelity, or could not find the answer option that corresponded with the student's behavior, the test introduced construct-irrelevant variance by limiting the student's opportunity to use the expected response process. The current study used post-hoc interviews to capture the sources of challenge during test administrations. An upcoming phase of research will expand on this work to include think-aloud protocols with in vivo probes to gather further evidence of the response processes teachers used when administering assessments.

In this study, cognitive labs were used to identify potential sources of construct-irrelevant variance due to physical access challenges and student misunderstanding of how to answer the item. The cognitive labs were only conducted with a subset of students who tend to work on higher-order skills within the content of AA-AAS. Part of the protocol relied on behavioral observations rather than think-aloud. Students answered four yes/no questions at the end of each testlet. While some students' answers to those questions corresponded with what we observed while they took the testlet, in other cases their responses were inconsistent. For example, a student might say a testlet was "easy" even though they were not able to successfully answer any item without instructions on how to do so. One student answered affirmatively when asked whether the testlet was hard, and then again when asked if it was easy. Students were not asked to think aloud during the task, in order to give them maximum opportunity to focus on the task itself. Our study raises questions about the quality of response process evidence that comes from post-hoc interviews and whether think-aloud protocols are viable for this population of students. Since this research is ongoing and the sample size is still small, we will continue exploring this approach with modified data collection methods.

Across both methods, the evidence collected has been more useful for evaluating the presence of sources of challenge that are unrelated to the construct of interest, rather than confirm that students use the intended response process. As we are in the early phases of a transition to online delivery of AA-

AAS, identifying sources of construct-irrelevant variance is useful for refining the delivery system, the assessment content, and the resources to support test administrators in delivering assessments with fidelity. As we are able to minimize more of these challenges, future evidence to confirm actual response processes match what was intended may require additional techniques that do not rely on verbal expression, such as eye gaze tracking software.

References

Almond, P. J., Cameto, R., Johnstone, C. J., Laitusis, C., Lazarus, S., Nagle, K., Parker, C. E., Roach, A. T., & Sato, E. (2009). White paper: Cognitive interview methods in reading test design and development for alternate assessments based on modified academic achievement standards (AA-MAS). Dover, NH: Measured Progress and Menlo Park, CA: SRI International.

American Educational Research Association (AERA), American Psychological Association (APA), & the National Council on Measurement in Education (NCME). (2014). S*tandards for educational and psychological testing*. Washington, DC: AERA.

Dynamic Learning Maps (2013, June 4). *The First Contact census student characteristics.* Lawrence, KS: University of Kansas, Center for Educational Testing and evaluation.

Gong, B., & Marion, S. (2006). *Dealing with flexibility in assessments for students with significant cognitive disabilities (Synthesis Report 60).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi:http://dx.doi.org/10.1037/0003-066X.50.9.741

Padilla, J., & Benitez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*, 136-144.

Towles-Reeves, E., Kearns, J., Kleinert, H., & Kleinert, J. (2009). An analysis of the learning characteristics of students taking alternate assessments based on alternate achievement standards. *Journal of Special Education, 42*(4), 241-254.

Willis, G. B. (2005). Cognitive interviewing: A tool for improving questionnaire design. Thousand Oaks, CA: Sage.

Table 1

Test Administrator Assistance with Navigation and Entry

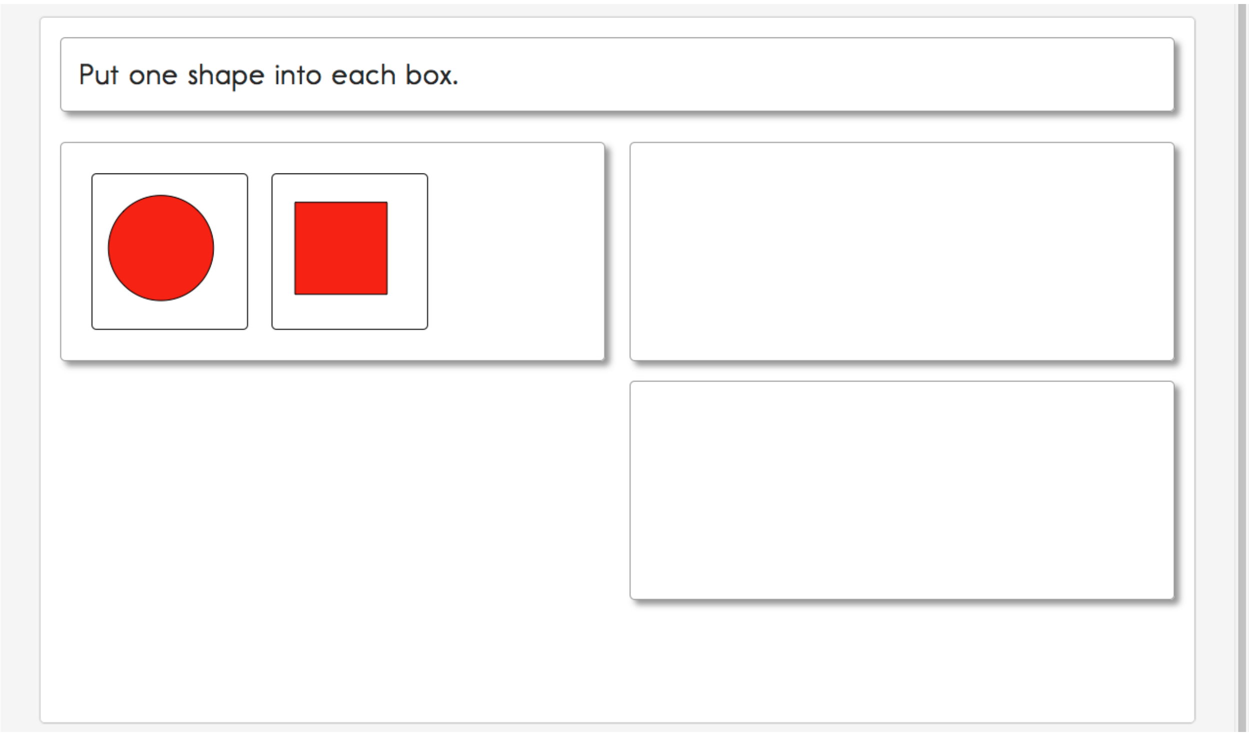
Test Administrator Action	ELA (*N* = 21)		Math (*N* = 15)	
	n	**%**	**n**	**%**
Test administrator navigated for the student	8	38	4	27
Test administrator physically assisted with navigation	2	10	--	--
Test administrator entered responses	6	29	4	27
Test administrator physically assisted with response entry	1	5	--	--

Table 2

Number of Students Using Writing Tools in Test Administration Observations of Writing Testlets (N = 22)

Writing Tool	**n***	**%**
Pen/Pencil and Paper	8	36
Traditional keyboard	6	27
Tablet keyboard	1	4
Other adapted keyboard	1	4
Alphabet chart/book	5	22
Letter dictation	1	4
Other Tool	3	14
No tool used	1	4

*Some students used multiple tools



Put one shape into each box.

Figure 1. Sample DD item from cognitive lab.