Received: 24 October 2016 | Revised: 21 December 2017 | Accepted: 22 December 2017

DOI: 10.1002/tea.21441

RESEARCH ARTICLE

WILEY | JRST

Articulating the validity evidence for a science alternate assessment

Lori Andersen | **Brooke L. Nash** | **Sue Bechard**

Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS), University of Kansas, Lawrence, Kanas

Correspondence
Lori Andersen, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS), University of Kansas, Lawrence, KS 66046.
Email: landersen@ku.edu

Abstract

Students with the most significant cognitive disabilities (SCD) are the 1% of the total student population who have a disability or multiple disabilities that significantly impact intellectual functioning and adaptive behaviors and who require individualized instruction and substantial supports. Historically, these students have received little instruction in science and the science assessments they have participated in have not included age-appropriate science content. Guided by a theory of action for a new assessment system, an eight-state consortium developed multidimensional alternate content standards and alternate assessments in science for students in three grade bands (3–5, 6–8, 9–12) that are linked to the Next Generation Science Standards (NGSS Lead States, 2013) and A Framework for K-12 Science Education (Framework; National Research Council, 2012). The great variability within the population of students with SCD necessitates variability in the assessment content, which creates inherent challenges in establishing technical quality. To address this issue, a primary feature of this assessment system is the use of hypothetical cognitive models to provide a structure for variability in assessed content. System features and subsequent validity studies were guided by a theory of action that explains how the proposed claims about score interpretation and use depend on specific assumptions about the assessment, as well as precursors to the assessment. This paper describes evidence for the main claim that test scores represent what students know and can do. We present validity evidence for the assumptions about the assessment and its precursors, related to this main claim. The assessment was administered to over 21,000 students in eight states in 2015–2016. We present selected evidence from system components, procedural evidence, and validity studies. We evaluate the validity argument and demonstrate how it supports the claim about score interpretation and use.

KEYWORDS
alternate science content standards, alternate assessment based on alternate achievement standards, large-scale assessment, low-incidence disabilities, students with significant cognitive disabilities, validity

1 | INTRODUCTION

The substantial diversity within the population of students with significant cognitive disabilities (SCD) necessitates variability in the content of alternate assessments, which has created inherent challenges in establishing the technical quality of these assessments. The multidimensionality of the Next Generation Science Standards (NGSS; NGSS Lead States, 2013) further compounds challenges for the validation of science alternate assessments, which are now based on more complex content standards than prior assessments. Guided by a theory of action for a new assessment system, an eight-state consortium developed multidimensional alternate content standards and alternate assessments in science for students with SCD in three grade bands (3–5, 6–8, 9–12) that are linked to the NGSS (NGSS Lead States, 2013). System features and subsequent validity studies were guided by a theory of action that explains how the proposed claims about score interpretation and use depend on specific assumptions about the assessment, as well as precursors to the assessment. This paper describes evidence for the main claim that test scores represent what students with SCD know and can do. We present data from validity studies designed to gather evidence for the assumptions about the assessment and its precursors, related to this main claim. We present selected evidence from system components, procedural evidence, and validity studies. We evaluate the validity argument and demonstrate how it supports the claim about score interpretation and use.

To understand the issues involved in validating an alternate assessment requires knowledge of the population of students with SCD, who comprise about 10% of the population of students with disabilities, or about 1% of the overall student population. The students in this highly heterogeneous population have a disability or multiple disabilities that significantly impact intellectual functioning and adaptive behaviors and require individualized instruction and substantial supports (Dynamic Learning Maps® [DLM], 2013). Historically, these students have received little instruction in science, despite the emphasis on *science for all* in science education policy documents over the past 20 years (e.g., National Research Council, 1996, 2012). In addition, alternate science content standards and assessments for students with SCD have either omitted many of the science concepts included for their general education peers, or focused on content designed for much younger students (Courtade, Spooner, & Browder, 2007; Karvonen et al., 2011). These differences comprise an equity issue regarding access to science for students with SCD. As of May 2017, the NGSS (NGSS Lead States, 2013) have been adopted in 18 states, and about a dozen more states are using state-developed standards based on the NGSS or the *Framework* (Loewus, 2017). This has resulted in the recent development of new general education assessments and new alternate assessments in science for many states.

Alternate assessments based on alternate achievement standards (AA-AAS) are designed to provide students with SCD opportunities to demonstrate understanding of academic content, as they are unable to participate in the regular grade-level assessments even with accommodations (ED, 2005). AA-AAS in science have a short history, beginning in 2007, and developers have worked to adhere to the same *Standards for Educational and Psychological Testing* as general education assessments (American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), 2014). Due to the need for

flexibility in assessed content, presentation, expression, and engagement for students with SCD, development and evaluation of AAs has required careful deliberation on what technical adequacy means for these assessments. This paper describes a new alternate assessment in science administered once in each of three grade bands (i.e., 3–5, 6–8, and 9–12) that incorporates multidimensional standards based on the NGSS performance expectations and examples of evidence used to support the validity argument. This science alternate assessment is the product of an eight-state consortium that administers assessments to over 21,000 students with SCD, which allowed for a more robust development process than what has previously been possible with state-specific science alternate assessments.

2 | LITERATURE REVIEW

The inclusion of all students in accountability systems was intended to raise expectations for students with SCD to learn and achieve in academics as they had mainly experienced instruction in functional skills (ED, 2003; 2015). However, in 2006, Towles-Reeves and Kearns found that teachers rated the state and district tests as having the lowest influence on what they taught, citing the students' Individualized Education Programs (IEPs) as having the most influence. A study conducted in Georgia indicated that teachers were providing academic instruction that was linked to the state grade level standards, yet little to no linkage was demonstrated between the IEP and the state standards, with most IEPs containing more functional than academic goals and objectives (Roden, 2011). These findings support that there are large disparities in the academic content that is taught to students with SCD and that much of this content is not linked to state standards. In contrast to these negative findings, Marion and Pellegrino (2006) stated that early alternate assessments had a positive outcome of informing educators that the capabilities of students with SCD were much greater than previously expected. Nonetheless, variations in assessed content are the biggest challenge to validity evaluations for alternate assessments (Gong & Marion, 2006).

Large-scale alternate assessments required for accountability purposes in science have been administered since 2007–2008, but as of 2015, have remained state-specific (Rogers, Thurlow, & Lazarus, 2015) with considerable variations in format, content, and administration procedures. Formats in current science AA-AAS include item-based, portfolio, and teacher observation checklists, as well as combinations of these. In 2015, more than half of states used professionally designed item-based assessments, which included selected-response, typically multiple choice, constructed-response, or performance tasks, while about one-third of states used portfolios in which teachers designed their own assessments or selected from a task bank (Rogers et al., 2015). One of the problems with the portfolio assessments used in some states, is that the assessments were not always comparable, making it impossible to compare scores of one portfolio to another (Wei, Pecheone, & Wilczak, 2014). This lack of comparability is a source of challenge in validity evaluation efforts.

In terms of content variations, the number and substance of the academic standards covered on the AA-AAS in many states are different from those on the general assessment, which led Rogers et al. (2015) to conclude that some students with SCD may not have access to rigorous grade-appropriate content and to recommend that states' alternate assessments should cover the same standards as their general assessments. Administration variations include differences in the intensity of teacher supports or scaffolding allowed, the artifacts required as evidence of student performance, and accessibility options available to students with SCD (Rogers et al., 2015).

These variations in AA-AAS design, administration, assessed content, and determinations of proficiency have interfered with efforts to compile a body of evidence to appropriately support validity arguments. In the early stages of the standards and assessment peer review, many states struggled with aligning the alternate assessment to academic content (ED, 2008). Alternate assessments often

inappropriately linked functional skills to the grade level content. "In 2005-06, over 30 states had not yet demonstrated that the alternate assessments based on alternate achievement standards meet the technical quality and alignment requirements in the Department's Peer Review Guidance" (p. 3). States also faced several challenges in documenting the validity and reliability of alternate assessment including: the heterogeneity of the group of students with SCD being assessed and how they demonstrated knowledge and skills, the relatively small numbers of students with SCD tested, the flexible assessment formats, administration, or experiences for alternate assessments (ED, 2008).

In a recent report, Thurlow and Wu (2016) found that while participation rates for AA-AAS are fairly consistent across states (slightly higher than 1% of the total population of students or about 10% of all students with disabilities), proficiency rates were extremely variable across states. Most states had quite high rates of students with SCD in the AA-AAS deemed proficient or above, with some as high as 90%. They concluded that at least part of the variability in performance rates is due to the differences in the states' AA-AAS themselves including where cut scores were set (Thurlow & Wu, 2016).

Challenges that AA-AAS developers face in providing evidence of the technical quality of alternate assessments, particularly in science, have set the stage for developing a new assessment system that allows for necessary flexibility in content and administration while providing a level of standardization that makes for a meaningful evaluation of technical adequacy than what has been historically possible. The assessment that is the subject of the validity evaluation in this paper represents efforts to design an assessment system that addresses many of the challenges previously mentioned, and facilitates the collection of evidence needed for validity evaluation.

Efforts to balance necessary flexibility for students with SCD and the standardization needed for validity evaluations are ongoing. Gong and Marion (2006) first described the challenges that AA-AAS faced regarding validity evaluations, explaining the flexibility in the design and administration of these assessments limits many of the available techniques for evaluating their technical characteristics. For example, teachers of students with SCD often create unique learning and assessment objectives for each student, which makes comparison of results impossible (Gong & Marion, 2006). In terms of flexible administration, students with SCD may require non-standardized customized response options, such as those provided by augmentative and alternative communication systems (AAC) and the additional time needed to implement them. These intended variations create tremendous challenges in validity evaluations of AA-AAS.

Validity is defined in the *Standards* as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11). The *Standards* provide "criteria for the development and evaluation of tests and testing practices" and "guidelines for assessing the validity of interpretations of test scores for the intended test uses" (p. 1). Although different sources of evidence are related to certain aspects of validity (e.g., evidence from an external review of items supports alignment with standards), validity itself is a unitary construct that is evaluated with a validity argument that brings together all sources of evidence coherently. According to Kane (2006), validation requires an interpretive argument and a validity argument. The interpretive argument, or theory of action, describes the set of claims, inferences, and assumptions that will be evaluated in validation, which is an explicit statement of the reasoning used in the interpretation and use of test scores (Kane, 2013a). The validity argument is an evaluation of the interpretive argument. Validity studies are conducted to gather evidence across five classifications: test content, response process, internal structure, relationships to other variables, and consequences of testing (AERA et al., 2014). A validity argument then integrates these various strands of evidence into a coherent argument, and makes explicit links between evidence, assumptions, and claims to show how the components of an assessment system function to produce desired and intended outcomes. The argument is then evaluated based on how well it is supported by the body of evidence.

Although assessment developers present validity evidence and arguments (e.g., Gotwals & Songer, 2013; Kampa & Koller, 2016; Opfer, Nehm, & Ha, 2012), few provide explicit theories of action that describe how a summative assessment functions within an educational system (Goldstein & Behuniak, 2011; Perie & Forte, 2011; Quenemoen, 2008; Reeves & Marbach-Ad, 2016). Such theories of action are particularly important for assessments of students with SCD because validity arguments need to evaluate the plausibility of precursors and assumptions about students with SCD that contribute to score interpretation, such as their opportunities to learn science or abilities to interact with the assessment system (Marion & Pellegrino, 2006; Perie & Forte, 2011). On the other hand, for interpretations that are limited to students' knowledge of a construct, interpretive arguments that solely consist of a theory that defines the assessed construct may be sufficient (Kane, 2013b).

The variability in the many AA-AAS led to a discussion on how to evaluate the technical quality of the assessments. The intended variability in assessment and learning objectives for students with SCD has a long history and is necessary to meet their needs, however, there are some new ways to deal with this flexibility while ensuring assessment results are comparable (e.g., Gong & Marion, 2006; Marion & Pellegrino, 2006). Prior to NCLB (ED, 2003), students with SCD were only held to personalized annual expectations via their IEP goals and objectives. The purpose of the law was to ensure that students with SCD are fully included in State accountability systems and have access to challenging instruction linked to State content standards (ED, 2005). Large-scale alternate assessments became the vehicle to achieve this goal. As technical evaluations of general education assessments are typically dependent on obtaining large sample sizes from standardized assessments to evaluate such things as item quality and test reliability, small sample sizes and intended variability within AA-AAS have prevented the use of many standard validity evaluation techniques in evaluations of AA-AAS (Marion & Pellegrino, 2006). The formation of alternate assessment consortia has allowed the aggregation of student data across states with common academic content standards and assessments, which enabled the use of a greater variety of evaluation methods.

Addressing other intended variability issues requires innovations in assessment design and evaluation techniques. For example, intended flexibility in assessment targets can be addressed by establishing construct comparability through a content map that shows a developmental sequence from less to more complex that could be used across assessments (i.e., a cognition model; Gong & Marion 2006; Marion and Pellegrino, 2006). Gong and Marion (2006) also recommended two processes that could ensure the validity of subsequent construct comparisons; (1) a content alignment process and (2) a cognitive analysis process. Similarly, Marion and Pellegrino (2006) argued for organizing validity and technical evaluations of alternate assessments around the assessment triangle with an emphasis on the three vertices (i.e., cognition, observation, and interpretation) and how they interact with each other and with validity. The assessment system presented in this article incorporates these recommendations. The assessment system design includes a cognition model. The validity evaluation includes content alignment and cognitive analysis processes. The basis of the assessment system and validity evaluation is the theory of action, which is essentially a series of if-then statements that make the interactions among the three vertices and validity explicit.

Another issue in validity evaluations of AA-AAS is how to document the validity evaluation; a few key articles and book chapters provide guidance for this documentation. Marion and Pellegrino (2006) presented an approach for organizing the technical documentation of an alternate assessment system that: (1) describes the assessment system, the population of students with SCD who participate in the assessment, content of the assessment, test development procedures, item analyses including differential item functioning (DIF), alignment, administration and training, scoring, characterizing errors associated with test scores, standard setting, and reporting and (2) introduces the validity framework and argument, presents empirical evidence across the five classifications, and presents the validity

evaluation. Marion and Perie (2009) demonstrated how to create a theory of action for an AA-AAS that is the basis of an interpretive argument, which then guides the design of appropriate validity studies, the generation of a complete and coherent validity argument for each claim in the theory of action, and a plan to evaluate validity. One example of this approach in action is presented by Goldstein and Behuniak (2011). They organized potential sources of evidence and demonstrated a validity argument for a large-scale alternate assessment, listing potential sources of evidence after articulating the purposes and assumptions of the state's skills checklist, organized by the five classifications of evidence. The authors concluded that the validity argument provides an organizational structure for the ongoing evaluation of a testing program and an "appropriate level of validity evidence should—at the very least—address all of the dimensions explicated in the Standards" (p. 188). For the purposes of this paper, the dimensions recommended by the *Standards* are used as the organizing structure. The content within this structure follows the recommendations of the literature. For a complete summary of the full body of evidence supporting the DLM science assessment system, see the *2015–2016 Technical Manual – Science* (DLM Consortium, 2017).

While examples of the validity of alternate assessments are scarce, there is even less information regarding the validation of science alternate assessments. We conducted a literature search on validation of alternate assessments that yielded a few studies, but none were for science alternate assessments. We compared the publicly available technical documentation for several states' 2015 science AA-AAS to the AERA et al. (2014) guidelines, restricting our review to states that use the selected-response format that is used in our assessments. We identified states that used selected response items from Rogers et al. (2015) review and conducted a web search to locate technical manuals; six technical manuals were located and reviewed. Overall, it was noted that several manuals included evidence across the five classifications specified by AERA et al. (2014), however, few included explicit theories of action or interpretive arguments (Kane, 2006) that articulated inferences and assumptions.

3 | PURPOSE

The purpose of this paper is to present evidence across the five classifications of validity evidence and an evaluation of the validity argument for an eight-state consortium's science alternate assessment based on propositions in the theory of action. The evidence is intended to provide examples of what could be used to support the use and interpretations of results from science alternate assessments. The example evidence provided in this paper does not represent the full body of validity evidence that can be generated and is limited to evidence related to the main claim that scores represent what students with SCD know and can do is presented, however, the space limitations of this manuscript prevent us from thoroughly describing all the ways that intended variability is managed in the assessment system. For more detailed and additional information on technical characteristics and validity evidence for the DLM® Science Alternate Assessment System, see the *2015–2016 Science Technical Manual* (DLM Consortium, 2017).

Validity evidence comes from system components, procedural evidence, and validity studies. To that end, the following sections provide: (1) an overview of the science assessment system including the overarching validity framework from which is was developed, (2) descriptions of the validity studies, including participants, methods, and results, which are organized by sections for each of the five classifications of validity evidence, and (3) an evaluation of the validity evidence in whole as it relates to the overall framework. Validity evidence presented in this paper is limited to evidence that supports the claim that scores represent what students with SCD know and can do.

3.1 | Overview of the DLM science alternate assessment system

One goal of the DLM Science Alternate Assessment System was to address the issue of balancing the need for flexibility and the need for standardization. In the previous section, the major issue with regard to flexibility was identified as the assessment objectives for students with SCD, which have historically been individualized to meet their needs. In this section, we describe system features that are relevant to understanding how this balance is achieved. To retain intended flexibility, alternate content standards called Essential Elements (EEs) were developed that have three levels, called linkage levels. The relationships between the standards, EEs, linkage levels, and testlets (i.e., set of related items) is shown in Figure 1. Each EE is linked to one standard and has three linkage levels. The three linkage levels are, in effect, a hypothetical cognitive model that describes a progression of learning steps from the lowest level (i.e., initial level) to the highest level (i.e., target level). The linkage levels are designed to provide multiple levels of access to the same disciplinary core idea (DCI) and science or engineering practice that is the focus of the EE. The creation of multilevel EEs based on a cognitive model allows students with SCD to be matched to appropriate content, while facilitating comparisons of scores and follows the recommendations of Gong and Marion (2006) and Marion and Pellegrino (2006) for ensuring construct comparability.

Other features of the assessment system are relevant to aspects of the validity argument, particularly with regard to maximizing accessibility. The science assessments are designed to be *instructionally relevant*, meaning that the content models good instruction (Kingston et al., 2017). For DLM, one of the ways instructional relevance is achieved is by delivering assessments as a series of testlets, each of which contains three to five related items that share a common stimulus, known as an engagement activity, which represents a common instructional context. Each science testlet begins with a nonscored engagement activity to increase access for this population by setting the context, activating prior knowledge, and increasing student interest. In general, engagement activities at the target linkage level provided contexts that were most conducive to including multidimensional items. Initial level testlets contain the least complex contexts, are administered offline by the test administrator, and involve students with SCD responding using picture response cards. Some engagement activities are science

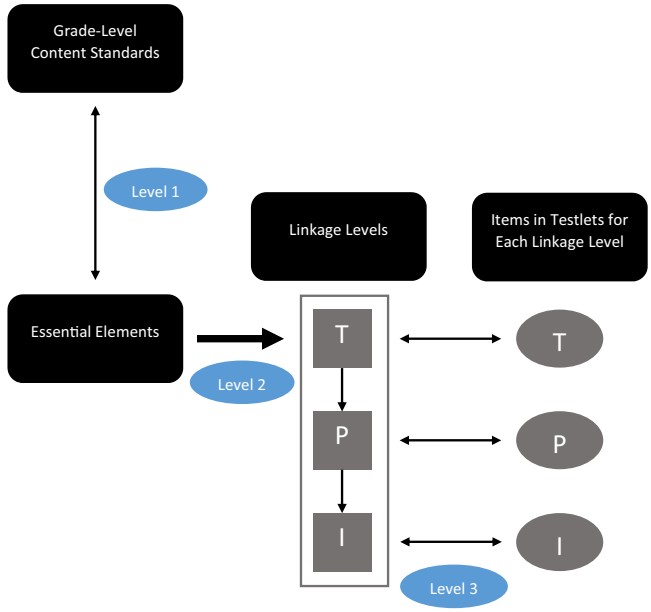

FIGURE 1 Relationships between standards, essential elements, linkage levels, and testlets [Color figure can be viewed at wileyonlinelibrary.com]

stories that describe a hypothetical student engaging in a science investigation. Testlet developers write science stories such that the student can use the science and engineering practices (SEPs) to demonstrate knowledge of DCIs that have been broken down into more manageable chunks. One effect of this chunking is that some items in testlets are unidimensional, and assess single practices or concepts to build a logical order of test questions within a testlet. Testlets begin with engagement activities that focus on a particular context for the content, followed by a series of questions about that context.

As seen in Figure 1, items in testlets are written to align to one of the three linkage levels for each EE. Students with SCD take one testlet for each EE at the linkage level that matches their skill level. They are assigned to linkage levels based on their communication and science skills, as indicated by teacher ratings. The assessment is adaptive, meaning that linkage levels of delivered testlets adjusts upward or downward based on performance on the prior testlet.

Assessments are delivered via an online platform. This allows for the opportunity to utilize more accessibility features than what has historically been possible with offline alternate assessments. For example, system supported accessibility features such as online text magnification makes it easier and more efficient for students with low vision and require large print to readily access the content. Within the online platform, available item types include single select response as well technology enabled items such as multiple choice multiple select, select text, matching lines and drag-and-drop items. However, only single select response items were used for science in order to avoid any potential additional cognitive load that the other item types may produce. There are two administration modes for testlet delivery: computer-delivered or teacher-delivered. For science, target and precursor level testlets are computer-delivered and are intended for students to interact directly with the computer. Single select response items for these testlets consist of three response options (either in text or images). Initial level testlets are teacher-administered and are intended for teachers to follow onscreen directions for engaging with the student and subsequently recording student responses into the system. Single response items for initial level testlets include five response options that reflect all possible student responses.

The content of each assessment covers a breadth and depth of science content at a complexity that is appropriate for students with SCD. Assessments have been created for each of three grade bands (i.e., grades 3–5, 6–8, and 9–12) and each grade band blueprint covers three science domains: life science, physical science, and earth and space science, while also including 10 DCIs and seven SEPs. Nine EEs are assessed at each grade band with one testlet for each linkage level of each standard. As a whole, each testlet is multidimensional and assesses both the DCI and SEP addressed by the linkage level.

3.2 | Validity evidence

Design features of the assessment system were carefully planned based on the theory of action for the science assessment. The theory of action (Figure 2) is a logic model for the assessment system that explains how goals will be met. Creating a theory of action begins with identification of critical problems that characterize alternate assessments in order to design a system that can mitigate these issues, many of which were identified in the literature review. Four propositions about score interpretation and use are the claims of the validity argument. The theory of action explicates how the first proposition depends on the qualities of the assessment (i.e., Assessment Assumptions in Figure 2) as well as precursors to the assessment (i.e., Precursor Assumptions in Figure 2). Relevant qualities of the assessment include considerations such as alignment of test content to standards, freedom from construct irrelevant variance, appropriateness of content, and test administration fidelity. Precursors include a wide range of conditions that must exist for the assessment to function as designed, including

Precursor Assumptions	Assessment Assumptions	Score Interpretation and Use Propositions	Goals
1. Alternate content standards, the Essential Elements, provide grade level access to NGSS and prepare students for college, career, and citizenship	**1. Testlets presented to the student align to the Essential Element and are free from construct irrelevant variance**	**1. Scores represent what students know and can do**	1. Students with SCD are able to show what they know and can do through the end of year assessment tasks
2. The system used to deliver assessments is designed to maximize accessibility	2. The end of year assessments have been designed to allow students to demonstrate their knowledge and skills in relation to academic expectations	2. Achievement level descriptors provide useful information about student achievement	2. Parents, teachers, and students have high expectations for students' academic achievement
3. The linkage levels represent the Essential Elements at appropriate access points for students with SCD	3. The combination of testlets administered at the end of the year measure knowledge and skills at the appropriate breadth, depth, and complexity of the content	3. Inferences regarding student achievement, progress, and growth can be drawn at the Domain level	3. Students achieve increasingly higher academic expectations
4. Educators understand the personal needs and preferences of their students and correctly document the students' needs within the assessment system	**4. Teachers administer the end of year assessments with fidelity so that students can respond to the items as intended**	**4. Assessment scores provide information that can be used to guide instructional decisions**	4. Trajectory of student growth in academic knowledge and skills is improved
5. Teachers provide instruction aligned with Essential Elements and at a level of complexity that provides an appropriate level of challenge			
6. Parents and teachers have high expectations regarding what students are able to achieve			
7. Students know how to interact with the assessment system			
UNINTENDED CONSEQUENCES Negative unintended consequences are minimized			

Note: Evidence for bolded items is presented in the article. Evidence for non-bolded items can be found in the Technical Manual.

FIGURE 2 Theory of action for science assessment design

establishing appropriate EEs, accessibility of the assessment system, appropriateness of content for students with SCD, correct use of system features during administration, provision of instruction that is standards-aligned, and their abilities to interact with the assessment system. In this way, the theory of action can be used to develop an interpretive argument for the assessment (Marion & Perie, 2009). The theory of action guides the design of validity studies because it specifies the data from assessment development that will support the propositions and the assumptions about the system upon which the validity argument relies. Evidence was identified across the five classifications of evidence and connected to the assumptions and propositions in the theory of action (Table 1). In the next five sections, the validity studies are presented, organized by the five classifications of evidence.

3.3 | Evidence based on test content

The Standards (AERA et al., 2014) explain that "important validity evidence can be obtained from an analysis of the relationship between the content of a test and the construct it is intended to measure" (p. 14). That is, evidence based on test content supports that the test measures what it purports to measure, which rests heavily on the representation of the content standards, the item and test construction process that ensures items both align to those content standards and avoid extraneous variables (i.e., construct irrelevant variance), as well as the evaluation of test items to support assumptions about item construction. As such, evidence based on test content of the DLM science alternate assessment is provided at three levels of the test development process: (1) development of the EEs, (2) development of items and testlets, and (3) empirical evaluation of item quality. This evidence supports several assumptions in the theory of action (Figure 2 and Table 1).

3.3.1 | Process to develop EEs

The EEs and corresponding linkage level statements were developed in a four-step iterative process with the goal of developing alternate content standards that would accurately reflect the knowledge,

TABLE 1 Sources of validity evidence by classification with correspondence to theory of action assumptions and propositions

Classification	Sources of evidence	Precursor Assumption 1	Precursor Assumption 3	Precursor Assumption 7	Assessment Assumption 1	Assessment Assumption 4	Score Interpretation Proposition 1	Score Interpretation Proposition 4
Test content (TC)	TC.1 Process to develop EEs	X	X					
	TC.2 External evaluation of EE alignment	X	X		X			
	TC.3 Process to develop items and testlets		X		X			
	TC.4 External review of testlets and items				X			
	TC.5 External evaluation of item and testlet alignment				X			
	TC.6 Pilot and Field testing							
Response processes (RP)	RP.1. Test administration observations			X		X		
	RP.2. Test administrator feedback survey			X				
Internal structure (IS)	IS.1. Differential item functioning (DIF) analyses				X			
Relations with other variables (OV)	OV.1. Correlations to scores on other assessments				X			
	OV.2. Correlations to demographic characteristics				X			
Consequences of testing (CT)	CT.1. Interpretation and use of score reports						X	X

skills, and understandings that were appropriately challenging grade-level and NGSS-aligned targets for students with SCD. Andersen and Nash (2016) described the process of EE development in greater detail and a brief overview is provided in this section. Several panels of experts were involved in the development process involving iterative review and feedback. First, content from the NGSS was selected for EE development based on a crosswalk of science consortium states' previous alternate content standards, demonstration of strong progressions across grade levels, and relative importance for students with SCD to be prepared for college, career, and community life. From there, EEs were drafted and the first panel of experts reviewed the draft standards with respect to several features including fidelity to the NGSS grade-level performance expectations and vertical alignment across grades. The second panel of experts reviewed the draft descriptions from the first panel and critiqued each EE based on a standardized checklist which again highlighted the goals of the EEs. Feedback and suggestions were used to make changes and improve clarity of the descriptions.

The third step in the development process was conducted at the state level within each state participating in the science consortium. Using the draft EEs from the previous iteration, a training video and guiding review questions, states facilitated internal reviews and compiled feedback into a spreadsheet to return to the consortium. The feedback was used to make edits for a final set of EEs in science. Finally, the state members convened for a final review of the EEs and corresponding linkage levels. The states voted to accept the set of descriptions that resulted from the iterative development process.

3.3.2 | External evaluation of EE alignment

While the link between the EEs and the NGSS grade level performance expectations was made explicit throughout the development process, this relationship was further evaluated as part of an externally conducted alignment study. Panelists in the study reviewed the EEs according to how well they represented the intended NGSS standard: (1) content alignment (DCI and SEP), (2) categories (domain, DCI, and topic), and (3) cognitive process dimension (a taxonomy for learning) Panelists' ratings were aggregated and evaluated against the following three criteria: (1) 90% or more of the EE ratings were rated as "partially" or "fully aligned" to the NGSS standard, (2) EEs matched the domain, DCI, and topic of the corresponding NGSS standard, and (3) 75% or more of the EE ratings were at the same or lower cognitive process dimension as the NGSS standard.

Results from panelists' ratings[1] showed that all EEs to aligned with the content of the associated NGSS standard. Across all grade bands, all EEs were found to adequately represent the intended NGSS domain, DCI and topic categories. Finally, more than 75% of the panelists' ratings indicated that the elementary, high school, and biology EEs were found to assess the same or lower cognitive process dimension as the standard, while 33% ($n = 3$) of the middle school EEs were found to reflect a higher cognitive process dimension than the standard. As the EEs are aligned to the NGSS grade-level content standards but at a reduced depth, breadth and complexity, the cognitive process dimension findings for elementary, high school and biology are expected. For EEs that were rated at a higher cognitive dimension than the NGSS, including those at the middle school level, follow-up analyses are planned to evaluate the external ratings and determine next steps.

3.3.3 | Process to develop items and testlets

A variant of evidence-centered design (ECD; Mislevy, Steinberg, & Almond, 1999) was used to develop test items in an effort to ensure alignment of items to the EEs. ECD provides a conceptual framework for designing, developing, and administering educational assessments generally (Mislevy et al., 1999) while more recent research has explored its use for alternate assessment (DeBarger et al.,

2011; Flowers et al., 2015). In any case, the use of an ECD framework in developing large-scale assessments is beneficial in substantiating the validity argument for score use and interpretation in that ECD requires test developers to explicitly link the inferences that they want to make about students' skills and understandings to the test items that are intended to provide evidence of those skills and understandings (DLM Consortium, 2017).

To this end, Essential Element Concept Maps (EECMs) were developed following an adapted version of the DeBarger ECD template, in order to provide item writers with specific information and guidance regarding the content of the EEs. These graphic organizers provide an explicit link between the conceptual understandings of the EEs and the test content. The template was designed for clarity and ease of use and included a definition of the content and science practices for the EE and corresponding linkage levels, key concepts and vocabulary at each level, common misconceptions, common questions to ask, prerequisite and requisite skills needed, and any accessibility issues related to particular concepts and tasks (Bechard & Sheinker, 2012). Item writers used the EECMs as guides to developing content-aligned and accessible items. Groups of items, known as testlets, were written to each of the three linkage levels available for each EE such that a testlet measured only one linkage level. A series of development steps occur prior to external review, including editorial review, review by test development staff, as well as a content and special education review by K-12 special educators.

3.3.4 | External review of testlets and items

The next step in the item development process was external review, which used content and special education experts, who were not part of item writing, to evaluate items with respect to three types of review: content, accessibility, and bias and sensitivity. Panels were created for each review type and each panel was given specific training and direction for completing their review. Specific criteria for evaluating both items and testlets as a whole were provided. The criteria for evaluating items and testlets from a content perspective focused on alignment to the EE, content accuracy, technical accuracy (e.g., only one correct answer option), and quality of format, layout, and graphics. The criteria for evaluating content from an accessibility perspective focused on the use of text and graphics that minimized unnecessary complexity, inference, or working memory. Finally, the bias and sensitivity criteria evaluated the content on the scope of the content ranging beyond the intended target, representativeness in race, ethnicity, gender, disability, and family composition, language that does not promote stereotypes or controversy, disadvantage a subgroup of people, or cause extreme emotional responses. Across all grades and rounds of reviews, only 0–4% of items sent through external review were identified as not meeting criteria by external reviewers. Testlets and items that were flagged were examined by the content team for revision or rejection and revisions were made as needed to address reviewer concerns. Out of 642 items and 202 testlets that were externally reviewed, the science test development team made a total of 85 minor revisions to items and 52 minor revisions to testlets in response to feedback from external review. No major revisions were needed.

3.3.5 | External evaluation of item and testlet alignment

While the link between the test items and the EEs was made explicit through the item development process via EECMs and verified via the external review process, this relationship was evaluated as part of the externally conducted alignment study (DLM Science Consortium, 2017). Similar to the evaluation of EEs, items and testlets were evaluated with respect to how well they aligned to the intended EE linkage level: (1) content, (2) categories (domain, DCI, and topic), and (3) intended cognitive process dimension. Panelists' ratings were again aggregated and evaluated against the following three criteria[2]:

(1) 90% or more of the item ratings were rated as "partially" or "fully aligned," (2) testlets matched the domain, DCI, and topic content, and (3) 75% or more of the target level item ratings[3] were at the same or lower cognitive process dimension.

Results from panelists' ratings showed that the majority of testlets were found to fully cover the associated EE linkage level content with all grade bands exceeding the 90% threshold. Furthermore, all testlets were rated as adequately representing the intended domain, DCI and topic categories. Finally, more than 75% of the panelists' ratings indicated that the elementary, middle school, and unique high school and biology testlets were found to assess the same or lower cognitive process dimension as the EE. However, of the testlets that were written to EEs that were common to both the high school and end-of-course biology test blueprints, 65% were written at a higher cognitive process dimension than the EE. Additional evaluations of the cognitive process dimensions for all items is planned (DLM Consortium, 2017).

3.3.6 | Pilot and field testing

While the EE and item development processes described above lend substantial evidence to the claim that the test content aligns to the intended constructs, additional evaluation of test items is necessary for validating the assumptions made during item construction and helping to ensure that construct irrelevant variance is not interfering with the intended alignment. Prior to the operational administration of the science assessment (i.e., for large-scale accountability purposes), all items were pilot or field tested and evaluated for technical quality.

A pilot test was conducted in spring 2015 for the purpose of evaluating the format and content of the new science items and testlets. Approximately 1,605 students with SCD in grades 3–12 across the five U.S. states who were members of the science consortium participated in the pilot test. Additionally, a field test was subsequently conducted in fall 2015 to evaluate edited content from the pilot test as well as new content not yet tested, and to gather cross-linkage level data in order to evaluate intended differentiation of cognitive complexity across linkage levels. More than 5,613 students with SCD in grades 3–12 across eight states, who were either members of or who had interest in becoming a member of the science consortium, participated in the field test. Table 2 below summarizes student participation in both the pilot and field tests by grade band and Table 2 displays student demographic information. In both pilot and field tests student participation was relatively evenly distributed across grade bands, the majority were male, white, not of Hispanic ethnicity, and did not participate in ESOL programs. Primary disability was not a required field during either administration.

3.3.7 | Pilot test results

During the pilot test, one testlet consisting of three to four items was administered for each Essential Element linkage level on the test blueprint for a total of 251 items across the three grade bands and biology. The percentage of students who correctly answered an item was used as the measure of item difficulty (i.e., item *p*-value). As the majority of items consisted of three answer options, items were expected to have a *p*-value of .35 or greater (i.e., greater than 33% chance of randomly selecting the correct option). Items with a *p*-value of less than 0.35 were interpreted to either have a content or accessibility issue and were reviewed by the test developers and content experts. Overall, 38 (15%) of the items were flagged as having a *p*-value less than .35. Of the flagged items, nine (26%) were deleted from the potential pool of items and 28 (74%) items were rewritten to address the content or accessibility issue. A pattern was noted in the rejected testlets: five of the six rejected testlets (83%) were at the precursor level and one (17%) was at the initial level.

TABLE 2 Participants by demographic ($n = 2{,}546$)

	Pilot test		Field test	
Demographic	***n***	**%**	***n***	**%**
Gender				
Female	568	35.3	1,978	35.2
Primary disability				
Autism	254	15.8	372	6.6
Documented disability	0	0.0	165	2.9
Intellectual disability	435	27.1	615	11.0
Multiple disabilities	90	5.6	156	2.8
Other health impairment	98	6.1	86	1.5
Specific learning disability	62	3.9	20	0.4
Missing	614	38.2	4,129	73.6
Race				
White	1,182	73.6	4,176	74.4
African American	169	10.5	1,056	18.8
Asian	55	3.4	114	2.0
American Indian	95	5.9	95	1.7
Alaska Native	29	1.8	19	0.3
Two or more races	3	0.2	126	2.2
Native Hawaiian or Pacific Islander	74	4.6	16	0.3
Missing			11	0.2
Hispanic ethnicity				
Yes	156	9.7	322	5.7
Missing	576	35.8	0	0.0
ESOL participation				
ESOL eligible/monitored student	79	4.9	105	1.9

3.3.8 | Field test results

During the field test, students were administered three testlets each at one of three adjacent EE linkage levels. The same flagging criteria for item difficulty was used to identify potentially problematic items. Of 259 items administered during the field test, 74 (27%) items were flagged for review, eight (11%) of which were deleted from the pool and 50 (68%) were revised.

Overall, the data collected from the pilot and field tests provided test developers empirical evidence of item quality. Insights from the review process of flagged items included adding more context to presentation of testlets, reduce text complexity particularly at the initial linkage level, and reevaluate the use of more difficult vocabulary when not required by the specific science concept assessed.

3.4 | Evidence based on response process

Examination of the response processes of tested students provides evidence about the match between the cognitive processes engaged in by test takers and the claims about the tested construct (AERA et al., 2014). Evidence based on response process includes information about implementation of the

assessment from both the student and test administrator perspectives. Assessments that are administered with fidelity to the prescribed procedures allow students the opportunity to respond to the test based on their knowledge, skills and understandings. Response process validity evidence for the DLM science alternate assessment is provided in two ways: (1) test administration observations, and (2) a test administrator feedback survey. This evidence supports two assumptions in the theory of action (Table 1).

3.4.1 | Test administration observations

Test administration observations were conducted in two states during 2015–2016. Project staff or state and local education agency staff conducted the observations using a standardized protocol. Observers collect data about how students and teachers interact with the assessment system including level of engagement, navigation, and any differences in administration (DLM Consortium, 2017). Of 37 observations, 29 (78%) were of computer-delivered testlets. Due to space limitations, results from teacher-administered test observations are not provided here but can be found in the *2015–2016 Technical Manual – Science* (DLM Consortium, 2017).

The test administration observations were in part intended to evaluate the position that educators should allow students to respond to and engage with the system as independently as possible. That is, test administrator actions during the assessment should either directly support or not interfere with the student's opportunity to respond independently where possible. Evidence of test administrators' actions is summarized in Table 3, with actions categorized as supporting, neutral, or non-supporting of students' ability to engage independently. Overall, test administrators engaged in more supportive or neutral actions than non-supportive actions suggesting that teachers did not interfere with the student response process in ways that would introduce construct irrelevant variance.

Aside from test administrator actions, the assessment system should be designed such that students should be able to respond and engage in construct relevant ways regardless of any sensory, mobility, health, communication, or behavioral constraint. As displayed in Table 4, observations of the students taking computer-delivered testlets showed that 41.4% independently responded to items, and 13.8% used eye gaze as a form of independent answer selection. Furthermore, of the 37 observations collected, 36 (97%) of the students completed the testlet demonstrating that students were able to complete testlets regardless of disability.

TABLE 3 Test administrator actions during computer-delivered testlets ($N = 29$)

Evidence	Action	*n*	%
Supporting	Used verbal prompts to direct the student's attention	10	34.5
	Clarified directions	10	34.5
Neutral	Navigated one or more screens for the student	19	65.5
	Repeated question(s) before student responded	16	55.2
	Defined vocabulary used in the testlet	1	3.4
	Repeated question(s) after student responded	4	13.8
	Asked the student to clarify one or more responses	0	NA
Non-supporting	Used physical prompts	5	17.2
	Reduced number of choices available to student	0	NA
	Total Actions	65	100.0

TABLE 4 Student actions during computer-delivered testlets ($N = 29$)

Action	*n*	%
Navigated the screens independently	10	34.5
Navigated the screens with verbal prompts	5	17.2
Selected answers independently	12	41.4
Selected answers with verbal prompts	12	41.4
Indicated answers using eye gaze	4	13.8
Indicated answers using materials outside of KITE Client	4	13.8
Skipped one or more items	2	6.9
Used manipulatives	2	6.9
Total Actions	51	100.0

Note. Respondent could select multiple responses to this question.

3.4.2 | Test administrator feedback survey

Validity evidence based on response process was also collected through a test administrator feedback survey which was administered during the spring 2015–2016 operational test administration. Survey questions were aimed at test administrators' perceptions of the students' ability to respond as intended, free of barriers, as well as their own perceptions about administering testlets. Test administrators responded to three statements about the student for which they administered a test to using a 4-point Likert-type scale (strongly disagree, disagree, agree, strongly agree): (1) the student responded to items to the best of his/her knowledge and ability, (2) the student was able to respond regardless of his/her disability, behavior, or health concerns, and (3) the student had access to all necessary supports to participate (DLM Consortium, 2017). As shown in Table 5 below, the majority of test administrators agreed or strongly agreed to all three statements suggesting that there was a match between the cognitive processes used by students when responding to test items and those intended by the construct of the assessment.

TABLE 5 Test administrator feedback on students' assessment experiences ($n = 2, 267$)

	Strongly disagree		Disagree		Agree		Strongly agree	
Statement	*n*	%	*n*	%	*n*	%	*n*	%
Student responded to items to the best of his/her knowledge and ability	237	10.4	298	13.1	1,171	51.4	570	25.0
Student was able to respond regardless of his/her disability, behavior, or health concerns	400	17.6	402	17.7	1,113	49.1	352	15.5
Student had access to all necessary supports to participate	125	5.5	183	8.1	1,315	58.0	644	28.4

3.5 | Evidence based on internal structure

Analyses that examine the internal structure of the assessment determine how well "the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA et al., 2014, p. 16). There are typically three facets of internal structure that may be considered when gathering validity evidence for an assessment: dimensionality (i.e., the number of constructs measured by the test items), measurement invariance (i.e., the degree to which items measure the same construct(s) across different groups of test takers), and reliability (i.e., the consistency of measurement across repeated test administrations; Rios & Wells, 2014). Evidence from this classification supports assessment assumption one in the theory of action (Table 1). The following section is limited to the aspect of measurement invariance. Specifically, items were evaluated to determine if they function differently based on subgroups of students taking the science assessment.

DIF analyses are used to identify construct irrelevant variance whereby test items are performing systematically different for identifiable subgroups of students (AERA et al., 2014). For example, DIF may be detected if a test item uses terminology or a social context that is uncommon or unknown in a specific geographic region. If the intended construct measured by the item was not related to the terminology or context used, then differences in student performance on the item would be considered construct irrelevant variance. To analyze DIF, data from the spring 2016 operational administration were collected. However, due to sample size constraints, the initial DIF analyses only examined male and female subgroups. In total, 300 science items were eligible for inclusion in the analysis: 82 items in each of the elementary and middle school grade bands, and 136 items in the high school grade band. Sample sizes ranged from 276 to 4,134 students per item. Logistic regression analyses were conducted to predict the probability of a correct response given group membership (i.e., male or female) and total score.[4] If systematic differences exist between males and females after accounting for total score, it is known as uniform DIF. An interaction term was also included to determine if any systematic difference between males and females differed by total score (non-uniform DIF); for example, if low scoring males were advantaged by the item but high scoring males were disadvantaged and vice versa for females.

Results of the DIF analyses showed that 34 of the 300 items were flagged for evidence of uniform DIF. However, using Jodoin and Gierl's (2001) threshold values for distinguishing the magnitude of the effect (i.e., practical significance) which suggests that items with an effect size less than 0.035 have negligible DIF, all of the flagged items were negligible. Similar results were found for the combined model that also evaluated non-uniform DIF; 34 items were flagged but all had negligible effect sizes. Table 6 summarizes the number of items flagged by grade band for evidence of uniform DIF and in the combined model that detects uniform or non-uniform DIF.

Overall, the low flagging rates and negligible effect sizes indicate that the science items are not systematically advantaging students based on gender. As more data are collected and sample size requirements are met for evaluating other subgroups of students, additional DIF studies can further support validity evidence based on internal structure.

TABLE 6 Items flagged for evidence of DIF

Grade Band	Items flagged for uniform DIF	Items flagged in combined model	Total items	Number of moderate or large effect sizes
Elementary	9	10	82	0
Middle	11	14	82	0
High	14	10	136	0

3.6 | Evidence based on relations to other variables

Analyses should also provide evidence that the test is both related to external variables that, in theory, should be related (i.e., convergent validity) and is unrelated to variables that are not intended to be part of the test construct (i.e., divergent validity). In other words, "analyses of the relationship of test scores to variables external to the test provide another important source of validity evidence" (AERA et al., 2014, p. 16). Evidence from this classification is related to assessment assumption one (Table 1).

After the first operational testing year in science, external validity evidence was limited to two types of correlational analyses (DLM Consortium, 2017). First, inter-correlations were calculated between total scores on all content areas. While relationships across content areas can indicate how consistently students perform across the different constructs of interest, these constructs are intended to be different (and therefore assessed separately), and therefore, only moderate relationships were expected (DLM Consortium, 2017). Second, Pearson correlation coefficients between student demographic characteristics and total score were calculated to provide a form of discriminant validity evidence. As relationships are not expected to exist between how a student performs on the test and the student's demographic characteristics, the correlations should be close to zero. The correlation coefficients were determined between content areas and with demographic characteristics that met sample size requirements. Students' scores on the science assessment were moderately and positively related to performance on the ELA ($r = .57$) and mathematics ($r = .59$) assessments, and were not related to students' gender ($r = .03$), race ($r = .03$), or Hispanic ethnicity ($r = .04$).

3.7 | Evidence based on consequences of testing

Of particular importance to large-scale assessment used for accountability purposes is the accumulation of evidence based on testing consequences used to support its many purposes. Validity evidence must include the evaluation of the overall "soundness of these proposed interpretations for their intended uses" (AERA et al., 2014, p. 19). One critical component of this evaluation is ensuring sound score report interpretations both in terms of how to use and not use test scores.

The design of the score reports was a result of a series of studies that were originally conducted by the consortium for English language arts and mathematics. The science consortium subsequently adopted the template. The studies provided insight and guidance in the design of a score report that was intended to support sound interpretations and use (Clark et al., 2015). The studies involved focus groups that provided insight into parents' perceptions of alternate assessments and use of test results from such assessments which helped guide the types of information that would be helpful to parents on a score report as well as focus groups for reacting to report prototypes. Individual and paired interviews of educators helped identify misconceptions, gaps in explanatory information, and use patterns. With each study, the score report template was refined and edited for additional clarity, more comprehensive information/explanation, and ease of use.

Results indicated that teachers' explanations of the score reports to parents varied in terms of the specific parts of the reports they chose to focus on. Some teachers focused on the more fine-grained test results in their descriptions to parents and others chose to focus on the overall results. Teachers also found ways to connect the results seen in the reports to the types of information parents were used to seeing in other score reports as part of their interpretations (DLM Consortium, 2017).

Results from the studies further suggested that teachers used one part of the score report, known as the learning profile, the most for instructional planning purposes. Teachers reported using the score report for IEP planning, specifically to develop statements about students' present performance level as well as for setting annual goals (DLM Consortium, 2017).

3.8 | Evaluation of validity evidence

As explained previously, the theory of action (Figure 2) guided the gathering of validity evidence because it identified the propositions and the assumptions about the system. The plausibility of the proposed interpretations and uses of test scores is evaluated from the validity argument, which integrates the evidence from these different sources. A complete validity argument provides evidence for all of the inferences and assumptions and rules out alternatives. If the evidence supports that assumptions about the precursors and the assessment are plausible, then score interpretation and use can be considered valid. Table 1 shows sources of evidence that support the theory of action, organized by classification of evidence and corresponding assumptions and propositions.

We will examine the evidence for the first proposition about score interpretation and use, that scores represent what students know and can do. The majority of the validity studies presented in this paper support this proposition, which depends on five assumptions (Table 1). Each assumption was investigated via multiple sources of validity evidence. For example, the assumption about testlet alignment with EEs and freedom from construct irrelevant variance (Assessment Assumption 1) is supported by seven different sources of evidence, including a wide variety of procedural (e.g., development process for the EEs, process for item development) and evaluative evidence (e.g., external alignment study results, external review, pilot and field testing) across four of the evidence source classifications (Table 1). The findings in the test content classification rule out alternative assumptions (e.g., construct irrelevant variance or misalignment) and support the proposed interpretation and use of scores.

Different classifications of evidence address different assumptions and propositions (Table 1). For example, the evidence in the response process classification rules out alternate assumptions about student interactions with the system and the fidelity of assessment administration. Under the classification of internal structure, the DIF analyses rule out some alternate assumptions about measurement invariance across groups of male and female test takers. In the category of relations with other variables, the correlations to scores on other assessments and to demographic characteristics rule out the alternate assumptions that assessment scores have relationships with variables that should not, in theory, be related. Altogether, the evidence supports that the assumptions listed in Table 1 are plausible by refuting alternate assumptions, and that the proposed interpretation and use of test scores is plausible.

The validity argument supports the theory of action and interpretive argument, but does not establish them beyond doubt (Kane, 2013a). Validation is an ongoing, continuing process and evidence of alternate assumptions should lead to a reevaluation of the theory of action. For example, evidence gathered to date has not ruled out every possible alternate assumption because the nature of the data limited the examination of measurement invariance to male and female groups and prevented comparisons across other possible groups of test takers. Future evidence gathering will examine measurement invariance across other groups.

The validity argument presented in this paper provides an example of the approach recommended in the Standards (AERA et al., 2014) and the literature (e.g., Goldstein & Behuniak 2011; Kane 2006; Marion & Pellegrino 2006; Marion & Perie, 2009). We have specified both an interpretive argument that explains the interpretations and uses of test scores with the set of inferences and assumptions that clarify the reasoning for the interpretations (in the theory of action), and a validity argument that evaluates the interpretive argument.

4 | DISCUSSION AND IMPLICATIONS

To meet the persistent challenge of managing the variability in assessed content that is necessary to accommodate the wide variability within the population of students with SCD (Gong & Marion,

2006; Marion & Pellegrino, 2006) required a new approach to assessment design. This science alternate assessment system design addressed an identified challenge in alternate assessment validity evaluation by providing flexibility in the assessment objectives through the use of cognitive models. The sets of linkage levels within EEs facilitate comparisons of assessment scores across students with SCD and had implications regarding the validity evidence that was gathered. At the same time, the introduction of multidimensional standards such as the NGSS also required new approaches that can validly measure student performance in multiple dimensions (National Research Council, 2014). Testlets in this alternate assessment target student performance in two dimensions of the NGSS (i.e., DCIs and SEPs) simultaneously. These are two features that make the DLM Science Alternate Assessment System unique among contemporary alternate assessments. Iterative development that results from ongoing argument-based validity research enables solutions to such challenges to be refined over time. For example, currently student performance is reported at the domain level (e.g., life science), but future iterations may use refined multidimensional testlet designs to enable more detailed score reporting.

An important aspect of validity evaluations for AA-AAS and of implications for this study relates to the consequences of testing. The increased focus on science in AA-AAS is likely to affect curriculum and teaching for students with SCD. However, it is not yet known how this assessment will affect teachers' expectations and teaching plans for students with SCD in science, given prior findings that found IEPs to have more influence than tests (Roden, 2011; Towles-Reeves & Kearns, 2006). Although it is difficult to measure this effect in the first year of the assessment, we plan to collect data on students' opportunity to learn science every year as part of the validity argument (i.e., consequences of testing). The data from the first year of assessment provide a baseline of comparison that indicates a relatively low opportunity to learn science and little experience with some SEPs for many students with SCD. However, it is likely that low opportunity to learn is related to special education teachers' lack of training to teach science, which is an area of concern for science educators.

Previous validity evaluations of AA-AAS were limited in scope, breadth, and depth because of challenges presented by the nature of the data (e.g., disparities in assessed content, small numbers of participants, variations in administration; Gong & Marion, 2006; Marion & Pellegrino, 2006). The design of the DLM science alternate assessment removes many challenges by: providing a common set of alternate content standards, aggregating participants across states for larger sample sizes, and providing test administration guidelines that define boundaries for flexible test administration to ensure greater fidelity of implementation. These design choices allowed analyses of validity evidence that had not previously been possible for AA-AAS.

This study is important to the fields of science education and special education because studies documenting next generation science alternate assessments have not yet been published. We have presented the results of initial efforts to develop a validity argument for alternate science assessment based on new alternate science content standards (i.e., EEs). Preliminary evidence supports that the new alternate standards and assessments are valid. Implications of this work for science education include that the results of the alternate assessments and development of EEs for science may help increase expectations for the science abilities of students with SCD and that they may be taught science that is based on EEs and linked to grade-level NGSS standards. In particular, evidence from teacher surveys indicates a greater need for preservice training and professional development in science content and pedagogy for teachers of the students with SCD. This work adds to the body of evidence regarding the potential for students with SCD to learn science content that is based on the science *Framework*. This study has implications for equity issues related to the goal of helping all learners achieve science literacy.

ACKNOWLEDGMENTS

We acknowledge Dr. Meagan Karvonen's leadership of the Dynamic Learning Maps® project and the work of Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS) staff members who contributed to the development of the Dynamic Learning Maps® Science Alternate Assessment System.

ENDNOTES

[1] Panelists' ratings were analyzed by project staff for consensus ratings.

[2] Items were rated against several additional criteria that are not included here for the sake of simplicity.

[3] This comparison is only examined at the target linkage level because the target linkage level directly assesses the EE content.

[4] Total score in this context in not equivalent to the traditional definition of total number of items answered correctly or a scaled score. For more detail regarding the scoring model, see the *2015–2016 Science Technical Manual*, chapter 5.

ORCID

Lori Andersen http://orcid.org/0000-0001-6671-7037

REFERENCES

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

Andersen, L., & Nash, B. (2016). Making science accessible to students with significant cognitive disabilities. *Journal of Science Education for Students with Disabilities*, *19*, 3.

Bechard, S., & Sheinker, A. (2012). *Basic framework for item writers using evidence-centered design (ECD)*. Lawrence, KS: University of Kansas.

Clark, A., Karvonen, M., Kingston, N., Anderson, G., & Wells-Moreaux, S. (2015). *Designing alternate assessment score reports that maximize instructional impact*. Paper presented at the Annual meeting of the National Council on Measurement in Education, Chicago, IL.

Courtade, G. R., Spooner, F., & Browder, D. M. (2007). Review of studies with students with significant cognitive disabilities which link to science standards. *Research and Practice for Persons with Severe Disabilities*, *32*, 43–49. https://doi.org/10.2511/rpsd.32.1.43

DeBarger, A. H., Seeratan, K., Cameto, R., Haertel, G., Knokey, A-M., & Morrison, K. (2011). *Alternate assessment design–mathematics* (Technical Report 9: Implementing evidence-centered design to develop assessments for students with significant cognitive disabilities: Guidelines for creating design patterns and development specifications and exemplar task templates for mathematics). Washington, DC: SRI International.

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

Dynamic Learning Maps® Consortium (DLM). (2017, June). *2015–2016 Technical Manual – Science*. Lawrence, KS: Center for Educational Testing and Evaluation, University of Kansas.

Flowers, C., Turner, C., Herrera, B., Towles-Reeves, L., Davidson, A., & Hagge, S. (2015). *Developing a large-scale assessment using components of evidence-centered design: Did it work*? Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

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

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

Gotwals, A. W., & Songer, N. B. (2013). Validity evidence for learning progression-based assessment items that fuse core disciplinary ideas and science practices. *Journal of Research in Science Teaching*, *50*, 597–626. https://doi.org/10.1002/tea.21083

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*, 329–349.

Kampa, N., & Koller, O. (2016). German national proficiency scales in biology: Internal structure, relations to general cognitive abilities and verbal skills. *Science Education*, *100*, 903–922. https://doi.org/10.1002/sce.21227

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17–64). Westport, CT: Praeger.

Kane, M. (2013a). The argument-based approach to validation. *School Psychology Review*, *42*, 448–457.

Kane, M. (2013b). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73.

Karvonen, M., Wakeman, S. Y., Browder, D. M., Rogers, M. A. S., & Flowers, C. (2011). *Academic curriculum for students with significant cognitive disabilities: Special education teacher perspectives a decade after IDEA 1997*. Retrieved from ERIC database (ED521407).

Kingston, N. M., Karvonen, M., Thompson, J. R., Wehmeyer, M. L., & Shogren, K. A. (2017). Fostering inclusion of students with significant cognitive disabilities by using learning map models and map-based assessments. *Inclusion*, *5*, 110–120.

Loewus, L. (2017, May). Next generation science slowly takes shape. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2017/05/24/next-generation-science-tests-slowly-take-shape.html

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. https://doi.org/10.111/j.1745-3992.2006.00078.x

Marion, S. F., & Perie, M. (2009). An introduction to validity arguments for alternate assessments. In W. D. Shafer & R. W. Lissitz (Eds.), *Alternate assessments based on alternate achievement standards: Policy, practice and potential* (pp. 113–126). Baltimore, MD: Brookes.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999, December 4). *Evidence-centered assessment design. Educational testing service*. Retrieved from http://www.education.umd.edu/EDMS/mislevy/papers/ECD_overview.html

National Research Council. (1996). *National science education standards*. Washington, DC: The National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

National Research Council. (2014). Developing assessments for the next generation science standards. Committee on developing assessments of science proficiency in K-12. Board on testing and assessment and board on science education. In J. W. Pellegrino, M. R. Wilson, J. A. Koenig & A. S. Beatty (Eds.). Division of behavioral and social sciences and education. Washington, DC: The National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.

Opfer, J. E., Nehm, R. H., & Ha, M. (2012). Cognitive foundations for science assessment design: Knowing what students know about evolution. *Journal of Research in Science Teaching*, *49*, 744–777. https://doi.org/10.1002/tea.21028

Quenemoen, R. (2008). *A brief history of alternate assessments based on alternate achievement standards* (Synthesis Report 68). Minneapolis, MN: National Center on Educational Outcomes, University of Minnesota.

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 margins: Challenges, strategies, and techniques* (pp. 335–378). Charlotte, NC: Information Age Publishing.

Reeves, T. D., & Marbach-Ad, G. (2016). Contemporary test validity in theory and practice: A primer for discipline-based education researchers. *CBE-Life Sciences Education*, *15*, rm1. https://doi.org.10.1187/cbe.15-08-0183

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, *26*, 108–116. https://doi.org.10.7334/psicothema2013.260.

Roden, M. R. (2011). *The impact of alternate assessment on teaching and learning for students with significant cognitive disabilities (Order No. 3479265)*. Available from ProQuest Dissertations & Theses Global. (901259237). Retrieved from https://search.proquest.com/docview/901259237?accountid=14556

Rogers, C. M., Thurlow, M. L., & Lazarus, S. S. (2015). *Science alternate assessments based on alternate achievement standards (AA-AAS) during school year 2014–2015*. (Synthesis Report 99). Minneapolis, MN: National Center on Educational Outcomes: University of Minnesota.

Thurlow, M., & Wu, Y.-C. (2016). *2013–2014 APR snapshot #12: AA-AAS participation and performance*. Minneapolis, MN: National Center on Educational Outcomes: University of Minnesota.

Towles-Reeves, E., & Kearns, J. F. (2006). *Alternate assessment impact survey (AAIS) report. National Alternate Assessment Center, University of Kentucky. Lexington, KY*. Retrieved from: http://www.naacpartners.org/publications/researchReports/20120.pdf

U.S. Department of Education [ED]. (2003, December). *Improving the academic achievement of the disadvantaged; Final Rule34 CFR Part 200, Title I*. Washington DC: Author.

U.S. Department of Education [ED]. (2005, August). *Alternate achievement standards for students with the most significant cognitive disabilities: Non-regulatory guidance*. Washington DC: Author.

U.S. Department of Education [ED]. (2008). *National Technical Advisory Council: Validity evidence for alternate assessments based on modified achievement standards*. Retrieved from: www2.ed.gov/about/bdscomm/list/ntac/aas.doc

U.S. Department of Education [ED]. (2015). *Every Student Succeeds Act: Assessments under Title I, Part A & Title I, Part B: Summary of Final Regulations*. Retrieved from: https://www2.ed.gov/policy/elsec/leg/essa/essaassessmentfactsheet1207.pdf

How to cite this article: Andersen L, Nash BL, Bechard S. Articulating the validity evidence for a science alternate assessment. *J Res Sci Teach.* 2018;55:826–848. https://doi.org/10.1002/tea.21441