Constructing and Evaluating a Validation Argument for a Next-Generation Alternate Assessment

Amy K. Clark and Meagan Karvonen

University of Kansas

Author Note

Amy K. Clark, ATLAS, University of Kansas, 1122 West Campus Road, Lawrence, KS, 66045, akclark@ku.edu. ORCiD: 0000-0002-5804-8336

Meagan Karvonen, ATLAS, University of Kansas, 1122 West Campus Road, Lawrence, KS, 66045, karvonen@ku.edu. ORCiD: 0000-0003-2071-2673

This work was supported in part by the U.S. Department of Education, Office of Special Education Programs, under Grant 84.373 100001. The views expressed herein are solely those of the authors, and no official endorsement by the U.S. Department of Education should be inferred.

Correspondence concerning this manuscript should be addressed to Amy K. Clark, ATLAS, University of Kansas, 1122 W. Campus Road, Lawrence, KS, 66045. Email: akclark@ku.edu

The authors would like to acknowledge the Dynamic Learning Maps Governance Board and Technical Advisory Committee members for their contributions to the validity argument.

The Version of Record of this manuscript has been published and is available in Educational Assessment January-March 2020

https://www.tandfonline.com/doi/full/10.1080/10627197.2019.1702463

Abstract

Alternate assessments based on alternate achievement standards (AA-AAS) have historically lacked broad validity evidence and an overall evaluation of the extent to which evidence supports intended uses of results. An expanding body of validation literature, the funding of two AA-AAS consortia, and advances in computer-based assessment have supported improvements in AA-AAS validation. This paper describes the validation approach used with the Dynamic Learning Maps® alternate assessment system, including development of the theory of action, claims, and interpretive argument; examples of evidence collected; and evaluation of the evidence in light of the maturity of the assessment system. We focus especially on claims and sources of evidence unique to AA-AAS and especially the Dynamic Learning Maps system design. We synthesize the evidence to evaluate the degree to which it supports the intended uses of assessment results for the targeted population. Considerations are presented for subsequent data collection efforts.

Keywords: validation, assessment, accountability testing, diagnostic assessment, large-scale assessment

Constructing and Evaluating a Validation Argument for a Next-Generation Alternate Assessment

Alternate assessments based on alternate achievement standards (AA-AAS) provide the means for students with the most significant cognitive disabilities to participate in large-scale assessment, be held to high expectations for achieving grade-level academic content, and have their results included in accountability systems. Alternate assessments have evolved considerably in the last two decades, beginning as primarily assessments of functional skills with or without links to academics (Thompson & Thurlow, 2001) and, in response to the No Child Left Behind Act (2002), transitioning to grade-level-aligned assessments based on alternate academic achievement standards. Historically, AA-AAS were portfolio- or performance-based assessments, checklists, or inventories (Altman et al., 2010; Thompson & Thurlow, 2003). Under the Every Student Succeeds Act (ESSA; 2015–2016), the next generation of AA-AAS is aligned to more-challenging college- and career-readiness standards and is expected to meet high standards for technical adequacy (U.S. Department of Education, 2018).

Early AA-AAS systems generated some published validity studies, typically focusing on a single type of evidence. For example, content-related evidence included content analysis (e.g., Johnson & Arnold, 2004), alignment studies (Flowers, Browder, & Ahlgrim-Delzell, 2006; Roach, Elliott, & Webb, 2005), and evaluation of the effect of opportunity to learn on AA-AAS outcomes (Karvonen & Huynh, 2007; Roach & Elliott, 2006). As AA-AAS matured, guidance on interpretive arguments and synthesis of validity evidence emerged (Goldstein & Behuniak, 2011; Marion, 2010; Marion & Perie, 2009; Perie & Forte, 2012). However, these resources were mostly high-level recommendations (e.g., how to involve stakeholders in developing a theory of action) with limited examples of claims and evidence. The literature provides few examples of multiple sources of AA-AAS validity evidence summarized or evaluated for operational measures (Elliott, Compton, & Roach, 2007; Goldstein & Behuniak, 2012; Hager & Slocum, 2008; Johnson & Arnold, 2004). Given their relatively short history as a form of large-scale assessment for a complex and heterogeneous population, AA-AAS may be more susceptible to inadequate validation than other large-scale assessments are.

Advances in assessment design and technology in the last decade have changed how students, including the approximately 1% of students with the most significant cognitive disabilities, are assessed. In 2010, two consortia, the National Center and State Collaborative and the Dynamic Learning Maps Consortium, were awarded grants to create next-generation AA-AAS. These multistate projects addressed some pragmatic challenges for AA-AAS, such as limited population sizes that prevent the use of statistical techniques common in large-scale assessment. The new AA-AAS used more technically rigorous test development and administration procedures typically found in large-scale applications and benefitted from the emerging guidance on AA-AAS validation. This paper describes the validation work for the Dynamic Learning Maps® (DLM) Alternate Assessment System, including a summary of the approach to validation, sources of evidence, and validity evaluation in light of intended uses of assessment results.

The Dynamic Learning Maps Alternate Assessment System

The DLM Consortium develops and administers alternate assessments to nearly 90,000 students with significant cognitive disabilities in 19 states. DLM assessments measure alternate academic achievement standards in grades 3 through high school for students who cannot meaningfully access general education assessments, even with accommodations. Two assessment models are available: Integrated and Year-End. This paper focuses on the Integrated model, which provides summative results based on evidence collected from instructionally embedded and spring assessments administered to approximately 14,000 students annually.

Purpose

DLM assessments are designed to measure what all students with the most significant cognitive disabilities know and can do relative to grade-level alternate achievement expectations in English language arts (ELA) and mathematics[1]. Using a diagnostic approach, results provide fine-grained information about student mastery of specific skills in addition to overall achievement levels traditionally reported for AA-AAS. The DLM Consortium governance board, which consists of state education agency members from each state, defined the intended uses of assessment results. Instructionally embedded results are intended for instructional planning, monitoring, and adjustment, while summative results are intended for reporting student achievement of grade-level standards to a variety of audiences; inclusion in state accountability models to evaluate school and district performance; and planning instructional priorities and program improvement for the following school year. Because the second and third uses are more dependent on state practices, in this manuscript we focus on consortium-level evidence for the first two uses.

Intended Examinees

DLM assessments are intended for students with the most significant cognitive disabilities. While local teams responsible for individualized education programs (IEPs) ultimately determine whether a student is eligible to take the assessment, the DLM Consortium uses agreed-upon eligibility criteria: (a) the student has a significant cognitive disability (a

[1] Science assessments are also available but due to differences in design and administration are not included here.

definition that is not limited to specific disability labels); (b) the student is primarily instructed using the DLM alternate content standards; and (c) the student requires extensive, direct, and individualized instruction and substantial support to achieve measureable gains in the grade- and age-appropriate curriculum.

Census data on the DLM examinee population reveal that students are extremely heterogeneous (Nash, Clark, & Karvonen, 2016). The primary disability labels for most students are intellectual disability (44%), autism (23%), and multiple disabilities (14%). Most (68%) are instructed primarily in classrooms separate from their grade-level peers, and nearly 60% read at or below a first-grade level. Students substantially vary in academic skills, communication systems, and sensory characteristics. For example, while 76% of students use speech for expressive communication, those who instead use sign language or symbols tend to use only one or two at a time. Among students who do not yet communicate with speech, sign, or augmentative and alternative communication systems, nearly half (48%) use conventional gestures or vocalizations to communicate intentionally, and 38% exhibit behaviors that are not intentionally communicative but may be interpreted by others as such. Many DLM examinees also have sensory and physical challenges that must be addressed for effective assessment. In the same census study, teachers reported that 19% of students use an augmentative and alternative communication device, 7% are blind or have low vision, and 5% are deaf or hard of hearing. One-third (33%) have a health or care issue that interferes with instruction or assessment.

Targeted Constructs

DLM assessments measure alternate content standards, called *Essential Elements*, which are aligned to the grade-level college- and career-readiness standards but at reduced depth, breadth, and complexity. Each subject features an underlying learning map model that includes thousands of nodes, or skills, and connections between them, that lead up to high school college- and career-readiness expectations. Each grade-level Essential Element is measured at five levels, called *linkage levels*, to provide all students with access to grade-level academic content. Each linkage level contains items that measure one or more nodes in the underlying map structure. The Target linkage level, which consists of nodes aligned to the grade-level Essential Element, is preceded by three precursor levels and extended by a successor level. As an example, for the seventh grade Essential Element *Determine two or more central ideas in a text*, the levels are as follows: Initial Precursor – *Can pair an object with a picture*; Distal Precursor – *Can identify the concrete details mentioned in informational texts*; Proximal Precursor – *Can identify the main idea for a paragraph in an informational text that lacks an explicit statement of the topic*; Target – *Can determine more than one main idea in an informational text*; Successor – *Can summarize the information in a familiar text.*

Assessment Design and Administration

The test pool is composed of short *testlets*, which consist of three to nine items that measure a linkage level for an Essential Element. Most items are multiple choice, but additional item formats are used on a limited basis, including multi-select multiple-choice and technology-enhanced items for matching, text selection, and sorting. All testlets are designed to reduce cognitive load and minimize barriers for the student population.

Teachers select Essential Elements from among a set of blueprint requirements. For each *conceptual area,* or collection of related Essential Elements, the blueprint stipulates the minimum number of Essential Elements that must be assessed. For example, in third-grade ELA, for the conceptual area “Determine critical elements of text,” the teacher selects for each student at least three Essential Elements from among the eight available, including at least one Essential

Element on reading informational texts and one on reading literature. Each student is expected to test at one linkage level per selected Essential Element. The system recommends the linkage level for an Essential Element based on teacher responses to a survey about the student's prior academic skills. Teachers may accept or reject the system-recommended linkage level and are encouraged to base the decision on specific instructional goals or additional information they have about the student.

During the instructionally embedded testing window, teachers administer testlets on a schedule they select, as long as each student meets all blueprint requirements. Teachers are encouraged to assess after instruction on a linkage level. The spring assessment measures student learning on a subset of the Essential Elements sampled from the previously completed blueprint, with each testlet's linkage level assigned by the system according to student performance on the previous testlet.

Scoring and Reporting

Given the underlying map structure and the desire to provide fine-grained reporting, the assessment is scored using a diagnostic model (see Chapter 5 of DLM Consortium, 2018) with mastery classifications calculated for each linkage level for every tested Essential Element. To aid interpretation by a range of stakeholders, we refer to linkage levels as *skills*. Because results are based on dichotomous skill-mastery decisions rather than a raw or scale score on a unidimensional trait, we use the term *results* (instead of scores) in score reports, interpretive materials, and throughout this paper.

Summative results are based on all student responses from the entire academic year and are summarized in individual student score reports as (a) the set of mastered skills for each assessed Essential Element, (b) the percentage of skills mastered in each conceptual area, and (c) an overall achievement level for the subject (i.e., emerging, approaching, at target, and advanced; see Clark, Nash, Karvonen, & Kingston, 2017 for method). During the instructionally embedded window, teachers are encouraged to generate on-demand progress reports that show mastery information for skills tested to date.

Developing the Interpretive Argument

The DLM Consortium adopted an argument-based approach to validity (Kane, 2006) and followed recommendations from the validity literature grounded in post-NCLB alternate assessment designs (e.g., Marion, 2010; Perie & Forte, 2012). The consortium started by developing a theory of action, with claims grouped into precursors, assessment characteristics, score interpretations and uses, and ultimate goals. The theory of action includes claims that would routinely be expected of large-scale academic assessment but also claims unique to the DLM system including assessment design, the emphasis on using results throughout the year to drive instruction, and the examinee population. Each claim has several associated propositions.

Once the consortium agreed on the claims and propositions, staff identified one or more studies or sources of evidence that could be used to evaluate each proposition. They also labeled the anticipated evidence according to the categories articulated by the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014).

In total, the DLM validity argument has 21 claims and 36 propositions. Such a large scope would be challenging to describe comprehensively in this paper. Instead, we focus on a subset of claims related to the intended use of results to guide instructional decisions. Figure 1 illustrates the logical relationships between these claims. Some claims are commonly found in theories of action for AA-AAS (e.g., assessment items are aligned to content standards; teachers administer assessments with fidelity), while others are unique to the DLM assessment system (e.g., several claims related to accessibility, one about accuracy of map pathways). Each claim has several associated propositions that must be evaluated. For example, the claim “Educators understand their students’ personal needs and preferences and correctly document those within the assessment system” has the following underlying propositions:

1. Supports that students need are available in the assessment system.
2. Educators select preferences to reflect what students typically use during instruction.
3. System-documented needs and preferences correspond with those found in the student’s IEP.
4. Educators accurately enter student needs and preferences into the system.

Propositions 2 and 3 may appear on the surface to be tangential. However, they are widely recommended and observable indicators of educators’ understanding of students’ needs and preferences.

Illustration of Validity Evidence

In this section, we describe selected evidence gathered to evaluate propositions underlying the claims included in Figure 1. We emphasize evidence that may be unique to the DLM system because of the intended examinees, content structure of the map, and nature of online assessment administration. Evidence is organized based on the assessment system’s developmental phase (i.e., design, development, and operational stages).

Design Phase

Accessibility. The consortium’s model for accessibility by design includes accessible content, accessible technology, and a personal learning profile that drives customization. Early in the design phase, we conducted a census survey to understand the characteristics of students who would be eligible for the assessments (Nash et al., 2016). Teachers completed the First Contact survey for each student who met eligibility criteria. The results informed development of many aspects of the assessment system.

Student access to content is one of the most critical assumptions evaluated in the validity argument. The goal for the DLM system was to ensure that each student would be presented with items that balanced rigor and access; the content should not be so advanced that the student could not demonstrate knowledge, nor should they be so easy that the student could not show knowledge of more complex content. This goal drove decisions about which nodes should be assessed at each linkage level and about the design of algorithms to recommend or assign linkage levels based on teacher responses to several First Contact survey items.

First Contact census results also influenced decisions about how students would interact with the content. For example, because of the range of communication modes in the population, students may indicate their responses to items in any response mode without it being recorded as an accessibility support. Information about students' cognitive, sensory, and motor skills, and the devices they use to interact with computers and instructional materials, drove decisions about design of the online student interface. The available accessibility supports are also based on First Contact results. For example, because students in this population who are also deaf or hard of hearing have idiosyncratic communication systems, teachers may use sign interpretation (rather than scripted American Sign Language or Signed Exact English) as an accessibility support to provide all students access to the content.

Content structure. The map structure was designed so that all students have an entry point for demonstrating their knowledge, skills, and understandings relative to grade-level expectations. Through an iterative process, map development teams synthesized available research to develop multidimensional representations of each subject consisting of map nodes and their connections (DLM Consortium, 2016). Cognitive scientists, content experts, population experts, and educators reviewed preliminary map structures. Reviewers evaluated maps for node grain size, cognitive complexity, ordering of connections, and accessibility for the range of examinees, including alternate pathways that reflect the varied ways students may acquire skills (e.g., writing skills for students with mobility challenges). The map review process (see Swinburne Romine, Andersen, Schuster, & Karvonen, 2018) yielded procedural and empirical evidence of the structure of knowledge, skills, and understandings measured by the assessments.

Essential Elements were written to reflect academic expectations for students with the most significant cognitive disabilities and align with grade-level college- and career-readiness standards. Blueprints were designed to have broad content coverage, reflect consistent and connected expectations across grades, and limit testing burden for students. Advice from the consortium governance board, external contractors, and state-selected content and population experts, along with evidence from map development, informed the final blueprints.

Professional development. The DLM Consortium offers both training on assessment administration and professional development to support teachers in delivering standards-aligned instruction. Test administrator training is required annually, and teachers must past post-tests to have access to DLM assessments. Training covers procedures such as test security but also the assessment design and expectations for blueprint coverage. DLM's instructional professional development departs from historic academic instructional approaches for this population, which have been fragmented and focused on behavioral approaches to teaching discrete skills. Consistent with the consortium's overall approach to advancing student learning, instructional professional development is designed to help teachers provide instruction that crosses groups of related Essential Elements and support students' conceptual understanding of interrelated academic content across grades. More than 50 modules are available to support instruction in ELA and mathematics.

Development Phase

Procedural evidence from test development. Test development procedures followed guidance provided in the *Standards* (AERA et al., 2014) for constructing assessments that adequately measure student knowledge, skills, and understandings. The DLM approach integrates evidence-centered design and Universal Design for Learning (UDL) principles (Bechard et al., in press). A consortium-developed graphic organizer called an Essential Element Concept Map serves as the basis for testlet development. It describes the nodes each linkage level measures, key vocabulary, definitions, misconceptions, prerequisite skills, cognitive process dimensions, and accessibility requirements. Item writers also used DLM Core Vocabulary (DLM Professional Development Team, 2013), a list of words frequently used in academic curricula and communication devices for this population, to minimize barriers to student demonstration of knowledge that may be introduced by the use of complex, construct-irrelevant vocabulary in assessment items.

Consortium state partners recruited item writers with expertise in the subject area and/or instruction for students with the most significant cognitive disabilities. Item writers were trained in UDL, accessibility, bias and sensitivity, and general item-writing guidelines. Item writers positively rated the item-writing training, process, and products. After the most recent event, all agreed or strongly agreed that other teachers would find the testlets instructionally relevant.

To reduce potential sources of construct-irrelevant variance, both internal project staff and external panels of educators from consortium states reviewed testlets for content, accessibility, and bias and sensitivity before field testing. Reviewers evaluated items and testlets as a whole using criteria related to content (e.g., alignment, cognitive process dimension), accessibility (e.g., clarity and appropriateness of images and graphics, minimizing barriers to students with specific needs), and bias and sensitivity (e.g., requirements for prior knowledge outside bounds of targeted content, fair representation of diversity; Clark, Beitling, Bell, & Karvonen, 2016). Test development teams considered reviewer recommendations when deciding whether to revise or reject items and testlets. Across grades, subjects, and pools, the vast majority of items were accepted after external review; testlets are typically recommended for rejection at a rate of 5% or lower. Following external review, pilot and field testing provided additional content-related evidence, including evidence of linkage-level ordering and low item-flagging rates (Clark, Kingston, Templin, & Pardos, 2014; DLM Consortium, 2016).

The test development process was also designed to minimize response barriers and promote construct-relevant interactions with items. Item-writer training and practice activities included discussion of how students might demonstrate the targeted knowledge, skills, and understandings and how to produce testlets that would be accessible to the largest number of students, avoiding barriers that may limit students’ demonstration of conceptual understanding. Ensuring item writers and external reviewers came from multiple states minimized the likelihood of disadvantaging students based on regional or cultural content in testlets. External reviewers judged whether testlets were reasonably free of barriers for students with limited working memory, communication disorders, and/or limited implicit understanding of others’ intentions and emotions.

Response process. Cognitive labs were conducted to evaluate the extent that students interact with assessment content as intended and to evaluate barriers to the intended response process caused by construct-irrelevant testlet features or item response demands (Swinburne Romine, Karvonen, & Clark, 2015). Because the DLM system is composed of some testlets administered to students via computer and others administered through teacher-student interaction, cognitive labs were conducted with both students and teachers. Student labs evaluated students' abilities to handle the response demands of computer-administered items. Students preferred drag-and-drop sorting and multi-select multiple-choice item types to click-to-place (a sorting item type that works with communication switches); however, the multi-select multiple-choice items posed conceptual challenge for many students. Test administrator labs evaluated teacher-administered testlets, including the clarity of educator instructions and the degree to which teachers could identify and select response options that corresponded with student behaviors across various student response modes. While participants were able to accurately indicate student responses, many recommended simplifying the educator directions to make them easier to follow. Additional labs are planned to evaluate refinements to educator directions and supports for test administrators.

Assessment assignment logic. At the development phase, we also evaluated the appropriateness of using First Contact survey responses to recommend and assign the linkage level for testlets. We used teacher responses to items about students' existing academic skills to assign students to one of four complexity bands. During the pilot administration, students from all complexity bands completed fixed form assessments spanning three linkage levels. Consistent with intended design, the percentage of students providing correct responses increased over complexity bands, and decreased as linkage level increased. In a hierarchical ordinal logistic regression for each subject, student complexity band was a significant predictor of the probability of success at the linkage level (Clark et al., 2014). These findings provide some support for using complexity band to recommend or assign linkage levels for assessment at an appropriate level of challenge for each student.

Reporting. Because DLM assessment results are intended to be instructionally useful, reports must clearly summarize student knowledge and skills at a grain size that supports planning, monitoring, and adjustment. After completing a needs assessment to understand parents' information needs and historic challenges with AA-AAS reports (Nitsch, 2013), we conducted parent and teacher focus groups to inform report design and interpretability (Clark, Karvonen, Kingston, Anderson, & Wells-Moreaux, 2015). Score-report prototypes were refined between rounds of review. We also created resources at this stage to support teachers' accurate interpretation and use of the reports.

Operational Phase

Accessibility. At the operational phase, we evaluate accessibility through multiple data sources related to implementation and teacher perceptions. Teachers documented students' accessibility supports in the Personal Needs and Preferences profile. The most commonly selected supports included human read aloud (88%), test administrator response entry (49%), and individualized manipulatives (37%) (DLM Consortium, 2017).

An annual teacher survey yielded additional evidence of system accessibility (DLM Consortium, 2017). Most teachers reported that students effectively used accessibility supports (93%) and that the supports were similar to those used during instruction (93%). Most teachers agreed that students had access to necessary supports (94%), that students responded to the best of their ability (87%), and that students responded regardless of health, behavior, or disability concerns (81%).

Test-administration observations also provided accessibility evidence. Project staff and state and local education agency staff annually conduct observations to evaluate whether students interact with the system as intended and respond to items irrespective of sensory, mobility, health, communication, or behavioral constraints. The goal is to annually observe sessions for the full range of students eligible for the assessment and across subjects, states, and testing windows. Consistent with intended flexibility, observers note whether students responded to tasks using verbal, gesture, and eye-gaze response modes. The protocol includes a section for recording circumstances when test administrators experience difficulty with accessibility supports but that section is rarely used (DLM Consortium, 2017).

Response-time data provide additional evidence for whether students respond as intended. Lengthier response times may indicate challenges in using the system or accessing the content. Typical testlet response time is around five minutes, with mathematics generally taking less time than ELA (DLM Consortium, 2016). These time spans are within expected limits set by the consortium governance board.

High expectations. Because of the historically limited academic curricula for this population and because the DLM assessment system is based on more-challenging content standards, the DLM validity argument includes a claim that teachers have high expectations for students' academic achievement. We use an annual teacher survey to collect longitudinal evidence of teacher expectations. In the most recent administration, teachers typically agreed or strongly agreed that content reflected high expectations for their students (82.0%), measured important academic skills (71.0%), and was similar to instructional activities used in the classroom (70.5%; DLM Consortium, 2018).

We collect additional evidence of teacher expectations by reviewing teacher choice of linkage levels during the instructionally embedded assessment window. Teachers typically used the system-recommended level (79% of the time). When the teacher adjusted the linkage level, it was typically down one level (12%) and often after first administering a testlet at the system-recommended level.

Instruction. We collect evidence of propositions related to teachers' approach to instruction from opportunity-to-learn data and evaluation of professional development modules. A tacit assumption in large-scale assessment is that achievement reflects student knowledge after a full year of instruction and that all students have equal and full opportunity to learn the scope of standards assessed. Given the evidence that opportunity to learn is limited and variable for students taking AA-AAS (Karvonen, Flowers, & Wakeman, 2013), this claim is explicit in the DLM validity argument. Evidence is obtained from the annual teacher survey, in which respondents indicated the number of testlets for which content matched the student's instruction (DLM Consortium, 2017). Across subjects, teachers reported that for most students (60%) the content of *most* or *all* testlets matched students' instruction.

Teachers also indicated the number of hours of instructional time spent per week per conceptual area. In ELA, the most frequent response was more than 20 hours; results varied more in mathematics. Sixty-four percent of students received fewer than 20 hours of academic instruction across all subjects per week. Instructional time responses were correlated with linkage-level mastery results by conceptual area. Values ranged from .22 to .40, with the strongest relationship observed for writing. Moderate values were expected and are likely a result of factors including variation in student population, instructional practice, and breadth of required Essential Elements for each conceptual area. For instance, a student may spend a large amount of instructional time on a conceptual area but only demonstrate mastery at the lowest linkage level because their growth is incremental or health issues reduced instructional time.

The DLM professional development system provides modules to support instructional practice (DLM Consortium, 2017). Over 102,000 self-directed modules were completed by the end of 2016–2017. In postmodule survey responses, teachers agreed with the importance of the content for the population of students and that they intended to apply the information to their instructional practice.

Testlet pool. To evaluate the extent that results reflect accurate information about student knowledge and skills, testlets were evaluated for alignment to linkage levels, the grade-level Essential Element, and college- and career-readiness standards; consistency of measurement at the linkage level; and difficulty.

Alignment. An external alignment study evaluated the relationships for (a) college- and career-readiness standards and Essential Elements, (b) the Essential Element and the Target linkage level, (c) linkage-level ordering, and (d) linkage levels to assessment items (DLM Consortium, 2016). Panelists provided positive ratings for content and performance centrality between grade-level content standards and Essential Elements and between nodes and items. Panels also evaluated correctness in the ordering of linkage levels. In limited instances, findings prompted a reevaluation of the best match of some Essential Elements to specific college- and career-readiness standards and items for which panelists' rating of the cognitive process dimension did not match the dimension identified by the item writer.

Consistency of measurement. Our approach to quantifying consistency of measurement is based on the DLM system's unique design, administration, and scoring approach. Consistent with the diagnostic assessment literature (e.g., Johnson & Sinharay, 2018; Roussos et al., 2007; Wang, Song, Chen, Meng, & Ding, 2015), we report attribute-level reliability for DLM assessments. Test-retest reliability is calculated by simulating a second administration based on students' known mastery status for each linkage level and the calibrated model parameters. Estimated mastery values are compared to true values and reliability results are reported as the tetrachoric correlations, correct classification rates, and correct classification kappas between true and estimated results. Evidence of measurement consistency is provided for each scoring level including linkage-level mastery status; the number of linkage levels mastered per Essential Element, conceptual area, and subject; and the overall achievement level (Thompson, Clark, & Nash, in press).

Attribute-level (i.e., linkage-level) reliability was calculated for 1,275 linkage levels across 255 Essential Elements in all grades and both subjects (DLM Consortium, 2018). There was strong consistency between true and estimated mastery status, with median values of .85 for Cohen's kappa, .95 for correct classification rate, and .98 for the tetrachoric correlations. While these values may be higher than those observed for some traditionally scored assessments, research indicates that diagnostic models demonstrate greater reliability with fewer items (e.g., Templin & Bradshaw, 2013) because consistency is evaluated for discrete categories rather than values on a continuous scale.

Difficulty. With multiple linkage levels available per Essential Element, the goal is to present each student with items that are of appropriate difficulty to maximize information that can be used to determine the student's mastery status for the linkage level. Annual evaluation of the operational pool includes a review of the percent of correct responses for each item; in the most recent pool, 88.4% of items had a *p* value ≥ .40. Additionally, item *p* values were compared for all items measuring the same linkage level; 94.6% of items had a standardized difference value within two standard deviations of the linkage-level mean *p* value (DLM Consortium, 2018). Further evidence that items are of intended difficulty is demonstrated by observed low rates of adaptation between testlets during the spring window, when adaptation is driven by thresholds of 35% and 80% correct on a testlet (DLM Consortium, 2016).

Bias. Items should perform equivalently across student groups. Operational items are evaluated for evidence of differential item functioning (DIF) for subgroups as sufficient data are available. To date, DIF analyses have been possible only for gender subgroups. In the most recent analysis, uniform and nonuniform DIF were evaluated via logistic regression for 4,171 items. Using the Jodoin and Gierl (2001) effect-size classification criteria, one item (0.02%) was flagged for uniform DIF and 15 items (0.4%) were flagged for nonuniform DIF with a moderate effect-size change. Test development teams review flagged items for potential content-related explanations of the statistical results. In this case, the teams identified two of 16 items with potential content-based explanations that the items favored one gender group (DLM Consortium, 2017). As subgroup sample sizes increase and operational pools are refreshed, additional subgroup analyses will be conducted.

Fidelity of administration. As described previously, the DLM system includes several options for flexibility during administration. The expectation is that test administrators will maintain fidelity to the intended process with as much standardization as possible. General guidance on allowable practices is provided in test administration and accessibility manuals and in required test administrator training. Short, testlet-specific documents provided when each testlet is assigned further support teachers' readiness to administer testlets with integrity by specifying needed materials, suggested manipulative substitutions, alternate text that may be needed for students with visual impairments, and any accessibility supports prohibited for the specific testlet due to the construct being measured.

Evidence of implementation fidelity is collected from multiple sources. In the most recent annual teacher survey, most respondents reported confidence in their ability to deliver computer-based and teacher-administered testlets (DLM Consortium, 2017). Test-administration observations indicated that test administrators accurately captured student responses (DLM Consortium, 2017). Additional evidence for fidelity of response entry is collected from scoring of writing samples. Writing testlets require the student to complete a writing product outside the system, which test administrators immediately score for low-inference features such as syntax and orthography by selecting the response option(s) that best matches the product (e.g., student wrote a complete sentence). When additional teachers scored the same writing samples, there was evidence of strong interrater agreement (DLM Consortium, 2017).

Blueprint coverage. Because of the Integrated model's flexible blueprint design, we evaluate the extent to which students are administered a combination of testlets that meet blueprint requirements. This evidence helps us evaluate whether assessment results are based on the intended breadth of content. The most recently available data indicate most students took a combination of testlets that met the exact blueprint requirements (49%–58% per grade and subject); a smaller portion exceeded the blueprint requirements (19%–22%). However, a portion of students (20%–33%) did not meet coverage expectations (DLM Consortium, 2018). While some students may not meet coverage requirements due to extenuating circumstances (e.g., chronic illness), most students are expected to meet all blueprint-coverage requirements. The Integrated model design includes a spring administration to aid in meeting blueprint-coverage requirements when students have gaps following instructionally embedded assessment. However, the spring testing window covers only a subset of Essential Elements and cannot guarantee that students with very few instructionally embedded assessments will cover all blueprint requirements.

Reporting. Results are intended to provide teachers with information they can use for instructional decision-making. A series of score-report interviews and focus groups was conducted over several years. Its purpose was to collect data on teacher interpretation of results, hypothetical uses, evaluation of the effectiveness of a training video for improving teacher understanding of report contents, and actual use of results (Clark, Karvonen, Swinburne Romine, & Kingston, 2018; Karvonen, Clark, & Kingston, 2016; Karvonen, Swinburne Romine, Clark, Brussow, & Kingston, 2017). Across studies, most teachers accurately read and interpreted student score reports. In some instances, teachers incorrectly interpreted the percentage of skills mastered by conceptual area as a percentage of correct item responses or percentage of trials. Teachers generally reported finding fine-grained mastery information most useful for instructional planning, goal setting, and organizing student groupings. Teachers also indicated a desire for additional interpretation materials to support their use of reports.

Evidence of progress report utility was collected from the teacher survey. Most teachers (70%) reported accessing a student progress report. The most commonly reported uses of progress reports were to document a student's progress on current IEP goals (58%), share results with parents (54%), and plan a student's next IEP (51%; DLM Consortium, 2017). In future studies we will further evaluate the utility of accessing and using progress reports and the impact of progress reports on instructional planning throughout the instructionally embedded window.

Evaluation of Evidence

Comprehensive validation includes an examination of the entire body of evidence and an evaluative statement regarding the extent to which uses are supported (Haertel, 1999; Kane, 2006, 2013). However, stakeholders do not have access to all evidence simultaneously, so evaluation is an ongoing, dynamic process (Marion, 2010). The DLM Alternate Assessment

System is still relatively new; the evidence presented here draws from the first years of operational administration. Again, rather than being comprehensive, the evaluation is limited to evidence presented earlier and in light of intended interpretations and uses in Figure 1.

Precursors

The logical relationships presented in Figure 1 and described throughout this manuscript begin with precursor claims related to accessibility. Evidence collected during the design, development, and operational administration phases support the claim that the system maximizes accessibility. There is also some evidence that teachers understand their students' personal needs and preferences and that students know how to interact with the system. Options for flexibility and a range of accessibility supports are available and used. Teachers indicated that most students had access to necessary supports, that the assessment supports were similar to those used during instruction, and that students responded to the best of their ability regardless of health, behavior, or disability concerns. So far, most of this evidence comes from observation and teacher self-report. We have not yet collected evidence to evaluate consistency of use across testlets or to ensure supports selected on the Personal Needs and Preferences profile were actually provided during testlet administration. Ongoing technology-system enhancements will soon allow teachers to immediately indicate supports used after each testlet to provide more-accurate information about support use linked to specific testlets.

While survey responses provide evidence of access for most students, a small proportion of students may still encounter barriers during assessment that could affect their ability to demonstrate what they know and can do. More research is needed to determine if the reported barriers were caused by gaps between students' accessibility needs and currently available supports, if challenges occurred because students were assessed during times when they were not receptive to using supports that were offered (e.g., during behavioral difficulties), or if survey responses were signs of other issues such as teacher misconceptions. Perceived barriers may also be related to discrepancies between supports students need during teacher-delivered instruction and sppports needed for online assessment. Additional research will help identify steps the consortium needs to take to narrow accessibility gaps, including improvements in teacher training or software enhancements to promote compatibility with current assistive devices.

Evidence collected during all three assessment developmental phases supports the claim that Essential Elements provide students with access to challenging grade-level content. Procedural and empirical evidence indicate that Essential Elements align to college- and career-readiness standards in general education. The underlying map structure supports content at five linkage levels for every Essential Element. This evidence for the content structure supports the bidirectional pathway between Essential Elements providing grade-level access and key stakeholders (i.e., parents and teachers) having high expectations for what students with significant cognitive disabilities can achieve. However, while teachers generally indicate that the DLM assessments measure important academic content and reflect rigorous academic expectations, some teachers disagree. We do not yet have data to evaluate the reason for the responses. One plausible alternative hypothesis is that some teachers find the expectations in the Essential Elements too low for their students. Another likely hypothesis is that, for some teachers, the disagreement reflects a long-standing curricular philosophy that prioritizes functional skills over academics and limits students' access to instruction based on challenging grade-level content. We also have not yet been able to collect evidence to evaluate how parents perceive the expectations or how those expectations influence teachers' instructional decisions.

Claims regarding map structure, Essential Elements, and high expectations connect to and support the claim that teachers provide instruction aligned with Essential Elements at a level appropriate for each student. Evidence for the claim that professional development strengthens educators' knowledge and skills also supports delivery of aligned instruction. While educators using the DLM professional development modules reported satisfaction with the modules' content and application to instructional practice, both the proportion of educators using the modules and the number of modules they complete could be improved. Because professional development supports many other claims in the validity argument, it is an area the consortium continues to prioritize for ongoing improvements, including new instructional resources and new methods to improve the reach of existing resources.

Another precursor in Figure 1 is the claim that map pathways accurately describe the development of knowledge and skills. The current map structure is supported by a robust review of published literature on content acquisition and development and by multiple rounds of internal and external review. While we have conducted some empirical evaluation of map structure and linkage-level ordering, the consortium's research agenda prioritizes additional data needed to support model-based methods for evaluating node connections and granularity.

Many of the precursor claims are prerequisites for teachers to deliver aligned instruction. While available evidence indicates students taking DLM assessments have more academic instructional time than their peers did less than a decade ago (Karvonen, Wakeman, Browder, Rogers, & Flowers, 2011), limited instructional time itself raises questions about students' opportunities to learn. Current evidence about instructional content also suggests that students have varying levels of opportunity to learn the full breadth of grade-level content, especially in mathematics. While opportunity-to-learn findings may reflect differences between how content is taught offline versus how it is assessed in online DLM assessments, responses indicated a need for further evidence of student opportunity to learn the full breadth of tested content.

Assessment Characteristics

Evidence from each of these precursors provides connections to claims about the assessment. Accessibility, map structure, and Essential Elements all support the evidence-based approach to test development. Procedural and empirical evidence support the claim that testlets align to the Essential Elements. Construct-irrelevant variance is minimized by applying lessons learned from the extensive First Contact survey data set to assessment design and development processes. Potential sources of construct-irrelevant variance are checked through observations and cognitive labs. These processes yield low rates of testlets recommended for rejection following external review and low flagging rates for DIF across gender subgroups. Additional studies are planned to expand DIF analyses to include additional subgroups and to identify potential item misfit within the diagnostic scoring model.

Administration-fidelity evidence collected to date provides some support for the claim that teachers administer assessments as intended. Writing-sample scoring agreement and test-administration observations provide some evidence that, for teacher-administered testlets, teacher responses reflect the knowledge and skills students demonstrated. Because these studies provide evidence for only a subset of test-administration events, we have taken steps to increase the number of observations collected in future years to provide more-robust information about the fidelity of administration.

Critical to the claim that the combination of testlets measures student knowledge and skills at the appropriate breadth, depth, and complexity are two sources of evidence: blueprint coverage and teacher selection of linkage levels. Each year, most students meet or exceed blueprint requirements. However, a subset of students does not meet all requirements, introducing potential fairness and construct-representation concerns. For students who do not meet all requirements, assessment results may not reflect the full breadth of what they know and can do. While teachers may still find their results useful in informing instructional decisions, the validity of inferences made from their overall academic achievement may not be supported. To close the coverage gaps, the consortium created a blueprint-coverage extract and made it available in the assessment system on demand so local staff may monitor completion. State education agency staff also used coverage data aggregated by district to identify sites that needed targeted technical assistance on blueprint-coverage expectations. Coverage will continue to be evaluated annually.

One of the early concerns with the Integrated model design was that teachers might ignore the system recommendations and choose low linkage levels in an attempt to “game the system” and hold students to low expectations. Yet operational data indicate most teachers accept system-recommended linkage levels for testlets. When the linkage level was adjusted, the most common adjustment was down just one level, and the adjustment often occurred after the student attempted the testlet at the recommended level. Further studies will need to include direct evidence of the rationales for linkage-level choices to evaluate how teachers’ expectations influence those choices.

Score Interpretation and Use

All of this evidence combines to support the claim that results represent what students know and can do. Reliability evidence demonstrates consistency in linkage-level mastery results. Additional psychometric evidence that supports this claim is summarized in the annual technical manual updates (DLM Consortium, 2016; 2017; 2018).

Annual score-report interpretation and use studies provide some support for the claims that results provide instructionally useful information and that teachers make sound instructional decisions based on results of the assessments. Teachers indicated that summative score reports provide useful information that informs IEP goals, instructional plans, and student groupings. Because score-report evidence was collected from a sample of teachers, results may not fully represent the broader population of teachers administering assessments. Score-report evidence does not currently tell us about the prevalence of teachers' optimal use of results for instructional planning, monitoring, or adjustment. More evaluation is needed on the use of progress reports and the extent to which professional development supports the soundness of teachers' instructional decisions based on DLM results.

Overall, and in light of the changes in academic expectations and assessment-system design from states' previous AA-AAS to the DLM system, the body of evidence summarized here reasonably supports teacher use of DLM assessment results to inform instructional practice. Additional studies will inform ongoing system maintenance and improvement over time. As additional data are collected, they will be incorporated into the validity argument, and the evidence will be reevaluated to determine the extent that intended uses are supported.

Discussion

DLM assessments were developed after a decade of advancements in AA-AAS. Starting an innovative AA-AAS system positioned the DLM Consortium to carefully consider the types of claims and evidence that reflect the consortium's philosophies and goals and the system design. Validity evidence blended traditional (e.g., bias and sensitivity review) and unique (e.g., learning map model review, reliability method) approaches. We drew from existing methods wherever possible, but design decisions were preceded by deliberation about fidelity to the overall assessment system design and the theory of action. The DLM Technical Advisory

Committee, whose members have diverse expertise in diagnostic classification modeling and other psychometric models, large-scale operational assessment, accountability, accessibility, and state policy, was a valuable resource in thinking about how to adapt and design appropriate studies.

As summarized in the preceding section, evidence generally supported the intended uses. This confirmatory approach is not unusual at the early phases of design and development (Marion, 2010). However, we were careful to look beyond confirmatory evidence (Kane, 2006). We reviewed validity evidence with an eye toward potential sources of invalidity, which we described in the previous section.

As noted by others (e.g., Marion & Perie, 2009), states must weigh a variety of factors when prioritizing validity studies. The DLM Consortium used a comprehensive approach to identify intended uses and needed evidence. Within this comprehensive approach, we prioritized studies according to immediacy of need and availability of data. With a rapid iteration cycle needed to design and deliver a system by the end of the grant, we prioritized studies that informed the design (e.g., cognitive labs to evaluate technology-enhanced items) and studies that likely were needed for U.S. Department of Education peer review. It is not surprising that the largest body of available evidence was related to content. For a new, large-scale academic achievement test, content-related evidence is of paramount importance, and lack of content-related evidence presents a threat to interpretation and uses of results. Perhaps more unusual for a new AA-AAS, we also emphasized interpretation and consequences from the beginning of the project, starting with a needs assessment and continuing research with annual studies on interpretation and uses. In a diagnostic model-based system with fine-grained score reports that differ from traditional summative score reports, and given the history of limited perceived value of AA-AAS results, we believed it important to attend to interpretation and uses across all phases of assessment system development.

Additional validity evidence is collected annually and reported in each technical manual update (DLM Consortium, 2018). We share findings with the DLM Technical Advisory Committee and the Governance Board, including interpretations and proposed next steps, for system improvement and future validity studies. We will collect additional evidence as the system expands over time. For example, as students' opportunity to learn improves and students become increasingly familiar with online assessments, these impacts need to be evaluated. Moreover, as states respond to the ESSA requirement to cap AA-AAS participation at 1% of the population, we will monitor the First Contact data for evidence of changes in the characteristics of students who remain eligible for DLM assessments. The availability of data over multiple administration years also creates opportunities to evaluate new assumptions, replicate study findings, and observe trends over time. As partner states' needs shift and the assessment evolves in response to collected evidence, the validity argument must be reviewed and refined.

References

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

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

Bechard, S., Clark, A. K., Swinburne Romine, R., Karvonen, M., Kingston, N., & Erickson, K. (in press). Use of evidence-centered design to develop learning maps-based assessments. *International Journal of Testing.*

Clark, A., Beitling, B., Bell, B., & Karvonen, M. (2016). *Results from external review during the 2015–2016 academic year* (Technical Report No. 16-05). Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation. Retrieved from https://dynamiclearningmaps.org/sites/default/files/documents/publication/External_Review_Report_2015-2016_Technical%20Report_16-05.pdf

Clark, A., Kingston, N., Templin, J., & Pardos, Z. (2014). *Summary of results from the fall 2013 pilot administration of the Dynamic Learning Maps Alternate Assessment System* (Technical Report No. 14-01). Lawrence: University of Kansas, Center for Educational Testing and Evaluation. Retrieved from https://dynamiclearningmaps.org/sites/default/files/documents/publication/pilot_summary_of_findings.pdf

Clark, A. K., Karvonen, M., Kingston, N., Anderson, G., & Wells-Moreaux, S. (2015, April). *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.

Clark, A. K., Karvonen, M., Swinburne Romine, R., & Kingston, N. (2018, April). *Teacher use of score reports for instructional decision-making.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Clark, A. K., Nash, B., Karvonen, M., & Kingston, N. (2017). Condensed mastery profile method for setting standards for diagnostic assessment systems. *Educational Measurement: Issues and Practice, 36*(4), 5–15. doi:10.1111/emip.12162

Dynamic Learning Maps Consortium. (2016). *2014–2015 Technical manual – integrated model.* Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation. Retrieved from https://dynamiclearningmaps.org/about/research/publications

Dynamic Learning Maps Consortium. (2017). *2016–2017 Technical manual update – integrated model.* Lawrence, KS: University of Kansas, ATLAS. Retrieved from https://dynamiclearningmaps.org/about/research/publications

Dynamic Learning Maps Consortium. (2018). *2017–2018 Technical manual update – integrated model.* Lawrence, KS: University of Kansas, ATLAS. Retrieved from https://dynamiclearningmaps.org/about/research/publications

Dynamic Learning Maps Professional Development Team. (2013). *The Dynamic Learning Maps core vocabulary*. Chapel Hill, NC: Author. Retrieved from https://www.med.unc.edu/ahs/clds/files/2018/09/vocabulary-overview.pdf

Elliott, S. N., Compton, E., & Roach, A. T. (2007). Building validity evidence for scores on a state-wide alternate assessment: A contrasting groups, multimethod approach. *Educational Measurement: Issues and Practice, 26*(2), 30–43. Retrieved from https://doi.org/10.1111/j.1745-3992.2007.00092.x

Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177 (2015–2016).

Flowers, C., Browder, D., & Ahlgrim-Delzell, L. (2006). An analysis of three states' alignment between language arts and mathematics standards and alternate assessments. *Exceptional Children, 72,* 201–215. doi:10.1177/001440290607200205

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

Goldstein, J., & Behuniak, P. (2012). Assessing students with significant cognitive disabilities on academic content. *The Journal of Special Education, 46,* 117–127. doi:10.1177/0022466910379156

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

Hager, K., & Slocum, T. (2008). Utah's alternate assessment: Evidence regarding six aspects of validity. *Education and Training in Developmental Disabilities, 43,* 144–161. Retrieved from http://daddcec.org/Portals/0/CEC/Autism_Disabilities/Research/Publications/Education_Training_Development_Disabilities/2008v43_Journals/ETDD_200806v43n2p144-161_Utahs_Alternate_Assessment_Evidence_Regarding_Six_Aspects_Validity.pdf

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power raters using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education, 14,* 329–349. doi: 10.1207/S15324818AME1404_2

Johnson, E., & Arnold, N. (2004). Validating an alternate assessment. *Remedial and Special*

Education, 25, 266–275. doi: 10.1177/07419325040250050101

Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement, 55,* 635–664. doi:10.1111/jedm.12196

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50,* 1–73. doi:10.1111/jedm.12000

Karvonen, M., Clark, A. K., & Kingston, N. (2016, April). *Designing alternate assessment score reports: Implications for instructional planning.* Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C.

Karvonen, M., Flowers, C., & Wakeman, S. (2013). Factors associated with access to the general curriculum for students with intellectual disability. *Current Issues in Education, 16*(3), 10. Retrieved from https://cie.asu.edu/ojs/index.php/cieatasu/article/view/1309/542

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

Karvonen, M., Swinburne Romine, R., Clark, A. K., Brussow, J., & Kingston, N. (2017, April). *Promoting accurate score report interpretation and use for instructional planning.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

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)

Marion, S. F. (2010). Constructing a validity argument for alternate assessments based on modified achievement standards. In M. Perie (Ed.), *Teaching and assessing low-achieving students with disabilities: A guide to alternate assessment based on modified achievement standards* (pp. 247–268). Baltimore, MD: Brookes.

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

Nash, B., Clark, A. K., & Karvonen, M. (2016). *First contact: A census report on the characteristics of students eligible to take alternate assessments* (Technical Report No. 16-01). Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation. Retrieved from https://dynamiclearningmaps.org/about/research/publications

Nitsch, C. (2013). *Dynamic Learning Maps: The Arc parent focus groups.* Washington, DC: The Arc. Retrieved from http://dynamiclearningmaps.org/about/research/publications

No Child Left Behind Act of 2001, P.L. 107-110, 20 U.S.C. § 6301 et seq. (2002).

Perie, M., & Forte, E. (2012). Developing a validity argument for assessing students in the margin. In M. Russell & M. Kavanaugh (Eds.), *Assessing students in the margin: Challenges, strategies, and techniques* (pp. 335–378). Charlotte, NC: Information Age Publishing.

Roach, A., & Elliott, S. (2006). The influence of access to general education curriculum on alternate assessment performance of students with significant cognitive disabilities. *Educational Evaluation and Policy Analysis, 28,* 181–194. doi: 10.3102/01623737028002181

Roach, A. T., Elliott, S. N., & Webb, N. L. (2005). Alignment of an alternate assessment with state academic standards: Evidence for the content validity of the Wisconsin Alternate Assessment. *The Journal of Special Education, 38,* 218–231. doi: 10.1177/00224669050380040301

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. (2007). The Fusion Model skills diagnosis system. In J. Leighton, & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*. New York, NY: Cambridge University Press.

Swinburne Romine, R., Andersen, L., Schuster, J., & Karvonen, M. (2018). *Developing and evaluating learning map models in science: Evidence from the I-SMART project*. Lawrence, KS: University of Kansas, ATLAS. Retrieved from https://ismart.works/sites/default/files/ documents/Publications/I-SMART_Goal_1_Technical_Report_FINAL.pdf

Swinburne Romine, R., Karvonen, M., & Clark, A. (2015, April). *Gathering evidence of response processes for alternate assessments (AA-AAS)*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Templin, J., & Bradshaw, L. J. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30,* 251–275. doi:10.1007/s00357-013-9129-4

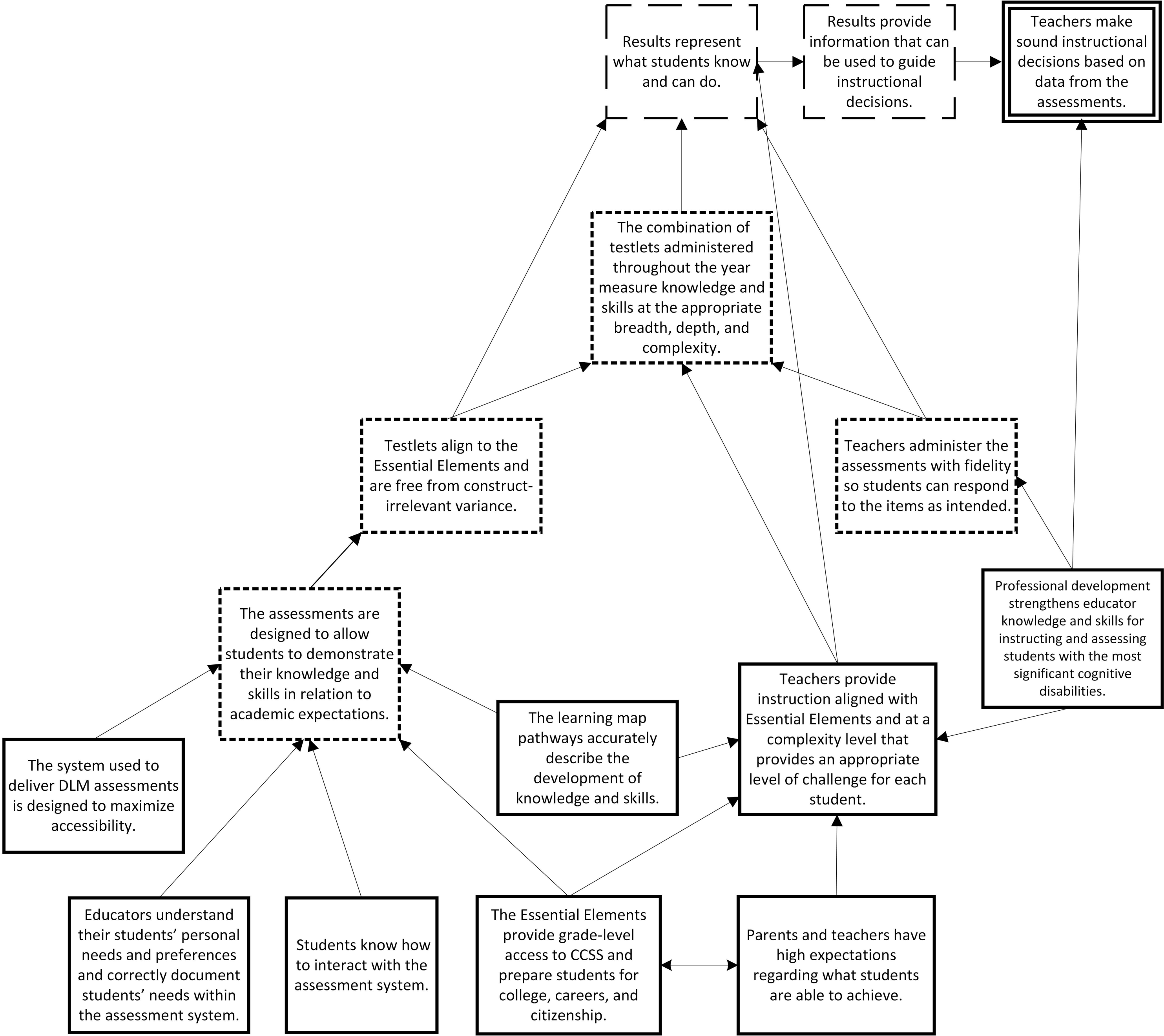
Thompson, S., & Thurlow, M. (2003). *2003 state special education outcomes: Marching on*. Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved from https://nceo.info/Resources/publications/OnlinePubs/2003StateReport.htm

Thompson, S. J., & Thurlow, M. L. (2001). *2001 state special education outcomes: A report on state activities at the beginning of a new decade*. Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved from https://nceo.info/Resources/ publications/OnlinePubs/2001StateReport.html

Thompson, W. J., Clark, A. K., & Nash, B. (in press). Measuring the reliability of diagnostic mastery classifications at multiple levels of reporting. *Applied Measurement in Education*.

U.S. Department of Education, Office of Elementary and Secondary Education. (2018). *A state's guide to the U.S. Department of Education's assessment peer review process.* Washington, DC: Author. Retrieved from https://www2.ed.gov/admins/lead/account/saa.html

Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement, 52,* 457–476. doi:10.1111/jedm.12096



Results represent what students know and can do.

Results provide information that can be used to guide instructional decisions.

Teachers make sound instructional decisions based on data from the assessments.

The combination of testlets administered throughout the year measure knowledge and skills at the appropriate breadth, depth, and complexity.

Testlets align to the Essential Elements and are free from construct-irrelevant variance.

Teachers administer the assessments with fidelity so students can respond to the items as intended.

The assessments are designed to allow students to demonstrate their knowledge and skills in relation to academic expectations.

Professional development strengthens educator knowledge and skills for instructing and assessing students with the most significant cognitive disabilities.

The learning map pathways accurately describe the development of knowledge and skills.

Teachers provide instruction aligned with Essential Elements and at a complexity level that provides an appropriate level of challenge for each student.

The system used to deliver DLM assessments is designed to maximize accessibility.

Educators understand their students' personal needs and preferences and correctly document students' needs within the assessment system.

Students know how to interact with the assessment system.

The Essential Elements provide grade-level access to CCSS and prepare students for college, careers, and citizenship.

Parents and teachers have high expectations regarding what students are able to achieve.