Visualizing Validity Evidence: Considering Strength of Evidence Following Disrupted Administration

Amy K. Clark, Megan Mulvihill, Jennifer Kobrin, W. Jake Thompson

ATLAS: University of Kansas

Author Note

Paper presented at the 2022 annual meeting of the National Council on Measurement in Education, San Diego. Correspondence concerning this paper should be addressed to Amy Clark, ATLAS, University of Kansas, 1122 West Campus Road, Lawrence, KS, 66045; akclark@ku.edu.

Abstract

Validity arguments consider the strength of evidence for intended interpretations and uses. When instruction and assessment administration are disrupted, for instance during pandemic conditions, test developers must consider the strength of validity evidence and whether intended uses are supported. We demonstrate a preliminary validity visualization method for evaluating relative strength of validity evidence and share considerations for other programs evaluating the strength of their programmatic evidence in disrupted years.

Keywords: validation, large-scale assessment, strength of evidence, visualization

Visualizing Validity Evidence: Considering Strength of Evidence Following Disrupted Administration

The Standards for Educational and Psychological Testing (the *Standards*; AERA et al., 2014) and other seminal works (e.g., Kane, 2006) emphasize the criticality of collecting validity evidence for assessments. Testing organizations rely on such evidence to determine the extent to which the intended interpretations and uses of assessment results are supported. However, technical documentation often describes validity evidence using a narrative format, which can be dense and complex, making it challenging to quickly evaluate the strength of the validity argument.

The ability of testing organizations to collect and evaluate validity evidence is also impacted by the dynamic and ongoing process of validation. Validity arguments must encompass each intended use of an assessment, and existing validity arguments must be reviewed and amended to account for changes or advancements in research and practice (AERA et al., 2014; Cizek, 2020). Any significant change in the purpose, design, content, administration, scoring, or outcomes of a test requires a careful reconsideration of the validity argument (U.S. Department of Education, 2018). As stated concisely by Jacobson and Dubravka (2019), “In light of any change, the troubling question always is: To what extent are the validation studies already conducted still relevant and to what extent do they need to be redone or understood in a different way given the change in use or interpretation?” (p. 14).

The onset of the COVID-19 pandemic in 2020 and the subsequent disruption of the 2020-2021 academic year highlighted the necessity for testing organizations to be able to re-evaluate existing validity arguments quickly and efficiently. According to the National Academy of Education (2021), the COVID-19 pandemic likely impacted the conditions, contexts, legal requirements, content, mode, and length of test administration and classroom instruction. The conditions of the pandemic likely had negative impacts on students’ academic performance and rates of participation in educational testing (Dadey et al., 2021; Sireci & Suarez-Alvarez, 2022). In response, the United States Department of Education offered temporary waivers for summative assessments, permitted new flexible administration procedures, and provided guidance that the focus of testing in the immediate future would be to provide information about student performance and resource allocation rather than to support program accountability (Rosenblum, 2021).

With these unprecedented changes to the administration and use of educational tests, Jacobson and Dubravka’s “troubling question” was brought to the forefront. Under a worst-case scenario (Clark et al., 2021), all claims related to assessment administration, instruction, and scoring could be at risk. The Center for Assessment (Dadey et al., 2021) urged state education agencies and other testing organizations to weigh validation considerations and impacts. The DLM Consortium, which develops the

Dynamic Learning Maps Alternate Assessment System, worked proactively to identify claims in the validity argument most likely to be impacted by the disruption (Clark et al., 2021). As part of this process, we developed an approach for visualizing validity evidence that would allow us to identify claims in the validity argument that were potentially impacted by pandemic conditions and whether intended interpretations and uses of the assessment were supported and defensible.

The following sections of this paper provide a brief introduction to argument-based validity theory. We then describe existing methods employed by testing organizations to present and evaluate validity arguments for large-scale summative assessments. Drawing on these prior conceptions of evaluating validity evidence, we demonstrate our process for developing and applying an approach for visualizing evidence for the validity argument for Dynamic Learning Maps alternate assessments in light of pandemic disruptions. We conclude this paper by discussing the implications of using this method to evaluate validity evidence and considerations for other programs that may adopt such an approach.

Argument-Based Validity Theory

Modern validity theory supports a validity argument framework, often consisting of an interpretation and use argument and validity evidence (Kane, 2016, 2020). The argument typically contains a series of propositions, assumptions, inferences, and warrants (collectively referred to as "claims" in this paper[1]) about an assessment that justifies its intended use and interpretation. Validity evidence can include empirical data, procedural evidence, or logical reasoning that evaluate the assessment claims. The strength of a validity argument relies on the structure of claims in the argument, which ultimately directs the collection and interpretation of validity evidence (Carney et al., 2019).

Validity evidence for evaluating claims may span the five sources of evidence identified in the *Standards* (AERA et al., 2014). The amount and type of evidence necessary to support a claim depends on factors such as the type or importance of the claim, the purpose of testing, and any available prior evidence (Chapelle, 2021). Claims that are unique or integral to the argument will require evidence in greater quantities or of higher quality; the amount of evidence is generally proportional to the importance of the claim or the potential risks to the test-taker (Kane 2013; Cizek, 2020). In some cases, a strong logical argument for a claim can reduce the need for empirical evidence (Kane, 2013). The overall requirement for evidence may also be lowered when evidence is difficult or expensive to obtain (AERA et al., 2014).

[1] see Carney et al., 2019 for a brief discussion about inconsistent terminology in validity arguments.

Evaluating Validity Evidence

Just as there is no single set of criteria for developing validity arguments (Lavery et al., 2020), it is impractical to set specific quantitative or qualitative standards for evaluating them (Cizek, 2016). Instead, test developers and testing organizations must rely on professional judgement to evaluate validity arguments (AERA et al., 2014). When evaluating validity arguments, the *Standards* indicate that a strong validity argument will (1) include multiple types of evidence from multiple sources, (2) link each piece of evidence to a specific claim, (3) consider evidence that contradicts claims, (4) rule out alternate explanations for intended outcomes, (5) provide detailed descriptions of how evidence was obtained, and (6) identify any portions of the validity argument that may differ from operational use at the local level (AERA et al., 2014).

Cizek (2020) provides guidance for making evaluation judgements about evidence of measurement (i.e., score accuracy) and justification claims (i.e., intended use and interpretation). When evaluating evidence for measurement claims, Cizek recommends considering the purposes of testing, the quantity of available evidence, the quality or relevance of evidence, the resources available for designing and carrying out validity studies, and the potential burden to individuals and organizations involved in the validation process. When evaluating evidence for justification claims, evaluators should consider the overall evaluation of measurement validity (based on the previous set of considerations), the appropriateness of the resource allocation for collecting validity evidence, the potential burden to individuals and organizations involved in the validation process, the input and incorporation of relevant stakeholders and values in validation, the need for testing and any alternatives to testing that may produce the same or better outcomes, and the possible intended and unintended positive or negative consequences of testing. Wools, Eggen, and Sanders (2010) identify three criteria, phrased as questions, that are applied to evaluate validity arguments: (1) Does the interpretive argument address the correct inferences and assumptions? (2) Are the inferences justified? (3) Is the validity argument as a whole plausible?

Test developers and testing organizations often apply their own criteria and processes for evaluating validity evidence. For K-12 assessment programs, validity evidence is additionally evaluated as part of federal peer review (USDE, 2018). We describe three example approaches to validity arguments and evaluative judgements from publicly available materials for three large-scale educational assessments: the Multi-State Alternate Assessment, the Smarter Balanced Assessment System, and the Dynamic Learning Maps Alternate Assessment. Examples of the validity arguments and evaluation criteria for each assessment system are presented in Appendices A through C.

The Multi-State Alternate Assessment

The Multi-State Alternate Assessment (MSAA, 2021) applies a contingent validity argument based on Kane's framework (2016, 2020). The MSAA interpretation and use argument includes four primary claims: the primary intended interpretation of the assessment scores and three primary intended uses. Subsumed under each of these primary claims is a multi-level set of sub-claims, which work together to support the four primary claims. Validity evidence is collected and presented in a narrative format for each claim or set of related claims. Each claim-evidence argument is evaluated based on the relevance, applicability, and completeness of the available evidence and assigned a rating of complete evidence, moderate to substantial evidence, limited evidence, or no evidence. A summary table lists each sub-claim and the rating assigned in the validity argument (see Appendix A).

The Smarter Balanced Assessment System

The Smarter Balanced Assessment System (2020, 2021) is a system comprised of both summative assessments, for which there are seven primary purposes, and interim assessments, which have four primary purposes. A prescriptive validity argument is presented in the technical documentation, which includes bullet-style lists of evidence based on four of the five sources described in the *Standards* (i.e., content, response process, internal structure, and other variables) for each intended purpose. Unlike the MSAA, the Smarter Balanced Assessment System does not provide evaluative judgements for validity evidence. Instead, it concludes the validity chapters of its technical reports with the statement that "Much of the information in this technical report supports the validity of the Smarter Balanced [summative or interim] assessment for one or more of its purposes." It is, therefore, up to the user of the test to review the individual pieces of evidence for each claim and determine its overall strength and/or applicability (see Appendix B).

The Dynamic Learning Maps (DLM) Alternate Assessment System

Validation of the DLM assessment is guided by a theory of action (Clark & Karvonen, 2020; 2021). Technical documentation for the DLM assessment (DLM Consortium, 2016, 2019) presents a narrative description of validity evidence within the prescriptive framework of the five sources of evidence as described in the *Standards* (AERA et al., 2014). The validity evidence is summarized in a table by source of evidence to depict where the evidence is described in the technical manual. The technical manual (DLM Consortium, 2016) also provides users with an overall summary and evaluation of the support for each primary claim (see Appendix C).

Visualizing Validity Evidence

Each of these organizations presents and describes validity evidence with narrative and/or list format. MSAA applies a rating scale to evaluate the relevance of evidence to each claim, and DLM provides a brief descriptive overall evaluation of the evidence for each primary claim. Only two organizations, MSAA and DLM, make use of tables that demonstrate validity evidence for the assessment.

The MSAA summary table (MSAA, 2021, p.93-94; see Figure A.3. in Appendix A of this paper) contains each of the four primary claims (the intended interpretation and use) of the validity argument, along with the subclaims associated with each primary claim. Each subclaim is rated according to the results of the validity evaluation of the evidence available for the claim (i.e., *no evidence, limited evidence, moderate to substantial evidence, or complete evidence;* see Appendix A for definitions). While this method is useful in describing the relative strength of evidence for each subclaim and primary claim, it does not indicate which claims rely on the same evidence or the total amount of evidence available for any given claim or subclaim.

Technical documentation for the DLM assessments includes a list of available evidence for the assessment and assigns it a unique identifier (e.g., 3.1, 3.2) based on the chapter of the manual in which the evidence is described. These identifiers are summarized in a table of the four primary scoring claims for the DLM assessment. However, this table does not describe the relevance, quality, or quantity of the accumulated evidence for each claim.

There are other methods used to describe and, in some cases, visualize validity evidence. For example, Hatala et al. (2015) created a table listing each individual source of evidence for the Objective Structured Assessment of Technical Skills, then provided brief description of how each source contributed evidence for claims about scoring, generalization, extrapolation, and decisions. A similar method was employed by Camara et al. (2019), who listed a series of four primary claims in a table, each with a set of associated subclaims, and then completed columns to identify sources of validity evidence (e.g., evidence based on content), the type of evidence (e.g., the test development process), and the citation for each piece of evidence. Each of the table summaries developed by Hatala et al. (2015) and Camara et al. (2019) present visual summaries of validity evidence that include information about the claim to be supported, the specific evidence to support the claim, and the direct source and citation of the evidence.

Key to note among all the methods of visualizing validity evidence that we have described is that no method has been developed to visually demonstrate variations in validity evidence or the strength of a validity argument. The existing methods of preparing summary tables would require users to carefully compare arguments over time to identify changes.

We introduce a preliminary method for visualizing validity evidence that extends existing practices of summarizing evidence. This method is adaptable over time and can be used to evaluate the changes in validity arguments as new evidence, counter evidence, and assessment practices are developed. In situations such as the COVID-19 pandemic, which caused disruptions to test administration, the approach for visualizing validity evidence can help test developers and testing organizations evaluate the claims of a validity argument that are most likely to be impacted and determine the extent to which intended uses are likely to be supported.

Methods

Study Context

We used the Dynamic Learning Maps (DLM) alternate assessment system to demonstrate the validity visualization approach. DLM is an operational assessment system used in over 20 states for state accountability purposes (DLM Consortium, 2019). DLM assessments are administered to students with the most significant cognitive disabilities, the ~1% of students who take alternate assessments because even with accommodations, general assessments are not appropriate. DLM assessments measure student achievement on academic content standards that are of reduced breadth and complexity of state college- and career-ready standards. To provide all students with access to grade-level academic content, each standard is available at five levels. These levels range from early foundational representations, precursors to the grade-level expectation, the grade-level expectation, and a successor skill for students who can show additional learning.

DLM assessments are scored using diagnostic classification modeling (e.g., Thompson, 2019). Score reporting for DLM assessments provides results at two levels. A Learning Profile summarizes fine grained mastery information for each standard at the five levels available for assessment (Figure 1). A Performance Profile summarizes overall performance in the subject (Figure 2). Fine-grained mastery information is intended to inform instructional planning, monitoring, and adjustment, while performance level results are intended to communicate performance in the subject to a variety of audiences and for inclusion in state accountability models.

Figure 1. *Learning Profile Portion of Score Reports*

REPORT DATE: 01-19-2022
SUBJECT: English language arts
GRADE: 10

Individual Student End-of-Year Report
Learning Profile 2021-2022

DYNAMIC® LEARNING MAPS

Dynamic Learning

NAME: Student DLM
DISTRICT: DLM District
SCHOOL: DLM School

DISTRICT ID: DLM District
STATE: DLM State
STATE ID: DLM State ID

Student's performance in 10th grade English language arts Essential Elements is summarized below. This information is based on all of the DLM tests Student took during the 2021–2022 school year. Grade 10 had 19 Essential Elements in 4 Areas available for instruction during the 2021–2022 school year. The minimum required number of Essential Elements for testing in 10th grade was 10. Student was tested on 11 Essential Elements in 4 of the 4 Areas.

Demonstrating mastery of a Level during the assessment assumes mastery of all prior Levels in the Essential Element. This table describes what skills your child demonstrated in the assessment and how those skills compare to grade level expectations.

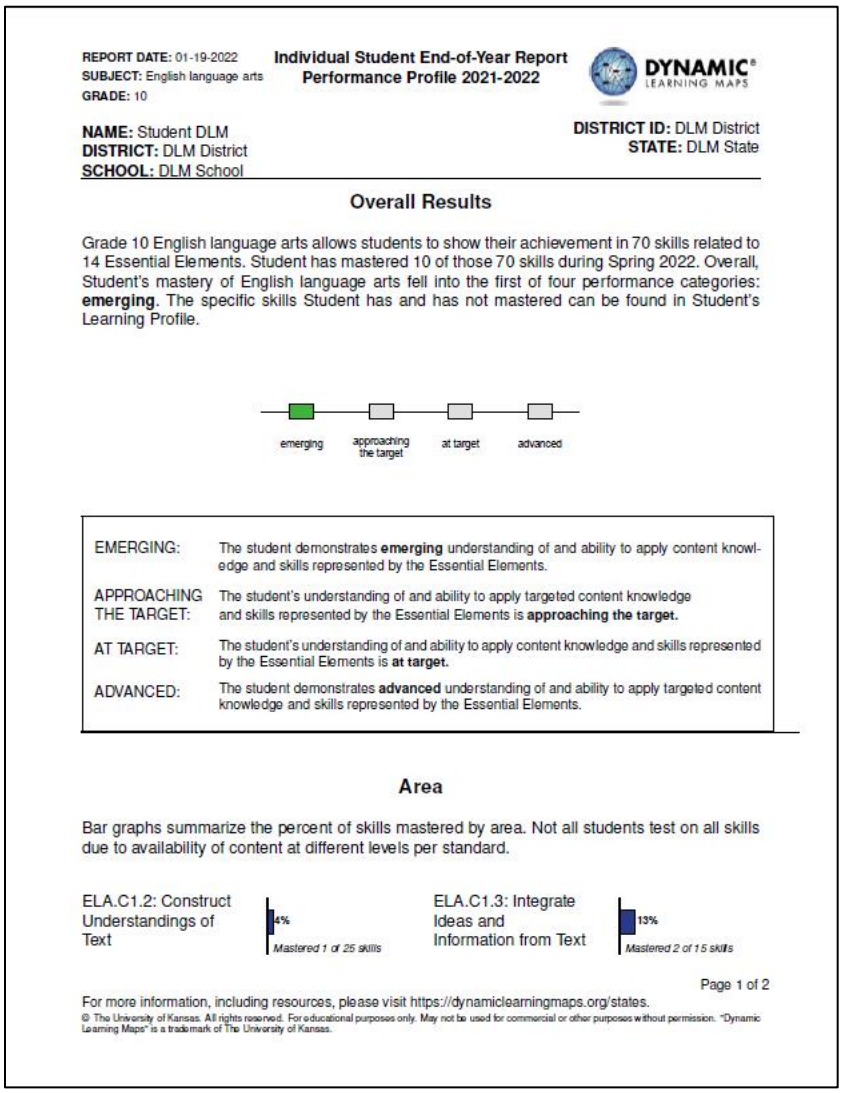
		Level Mastery				
Area	Essential Element	1	2	3	4 (Target)	5
ELA.C1.2	ELA.EE.RL.9-10.1	Identify concrete details in a familiar story	Answer questions by referring to a text	Cite textual evidence for explicit information in text	Discriminate between explicit and implicit citations	Determine a narrative's explicit meaning
ELA.C1.2	ELA.EE.RL.9-10.2	Identify the forward sequence in a familiar routine	Identify main idea	Identify details related to the theme of a story	Recount events contributing to the theme using details	Recount main events related to the theme
ELA.C1.2	ELA.EE.RL.9-10.4	Identify descriptive words	Identify the words or phrases to complete a literal sentence	Determine the meaning of idioms and figures of speech	Determine the meaning of words and phrases	Determine the meaning and impact of words and phrases
ELA.C1.2	ELA.EE.RI.9-10.1	Identify concrete details in a familiar informational text	Identify concrete details in an informational text	Cite textual evidence for inferred information	Discriminate between citations for explicit and inferred information	Cite evidence for a text's specific meaning

Levels mastered this year | No evidence of mastery on this Essential Element | Essential Element not tested

This report is intended to serve as one source of evidence in an instructional planning process. Results combine all item responses from the full academic year. Because your child may demonstrate knowledge and skills differently across settings, the estimated mastery results shown here may not fully represent what your child knows and can do. For more information, including resources, please visit https://dynamiclearningmaps.org/states.

Page 1 of 4

Figure 2. *Performance Profile Portion of Score Reports*



REPORT DATE: 01-19-2022
SUBJECT: English language arts
GRADE: 10

Individual Student End-of-Year Report
Performance Profile 2021-2022

DYNAMIC LEARNING MAPS

NAME: Student DLM
DISTRICT: DLM District
SCHOOL: DLM School

DISTRICT ID: DLM District
STATE: DLM State

Overall Results

Grade 10 English language arts allows students to show their achievement in 70 skills related to 14 Essential Elements. Student has mastered 10 of those 70 skills during Spring 2022. Overall, Student's mastery of English language arts fell into the first of four performance categories: **emerging**. The specific skills Student has and has not mastered can be found in Student's Learning Profile.

emerging | approaching the target | at target | advanced

EMERGING:	The student demonstrates **emerging** understanding of and ability to apply content knowledge and skills represented by the Essential Elements.
APPROACHING THE TARGET:	The student's understanding of and ability to apply targeted content knowledge and skills represented by the Essential Elements is **approaching the target.**
AT TARGET:	The student's understanding of and ability to apply content knowledge and skills represented by the Essential Elements is **at target.**
ADVANCED:	The student demonstrates **advanced** understanding of and ability to apply targeted content knowledge and skills represented by the Essential Elements.

Area

Bar graphs summarize the percent of skills mastered by area. Not all students test on all skills due to availability of content at different levels per standard.

ELA.C1.2: Construct Understandings of Text — 4% — *Mastered 1 of 25 skills*

ELA.C1.3: Integrate Ideas and Information from Text — 13% — *Mastered 2 of 15 skills*

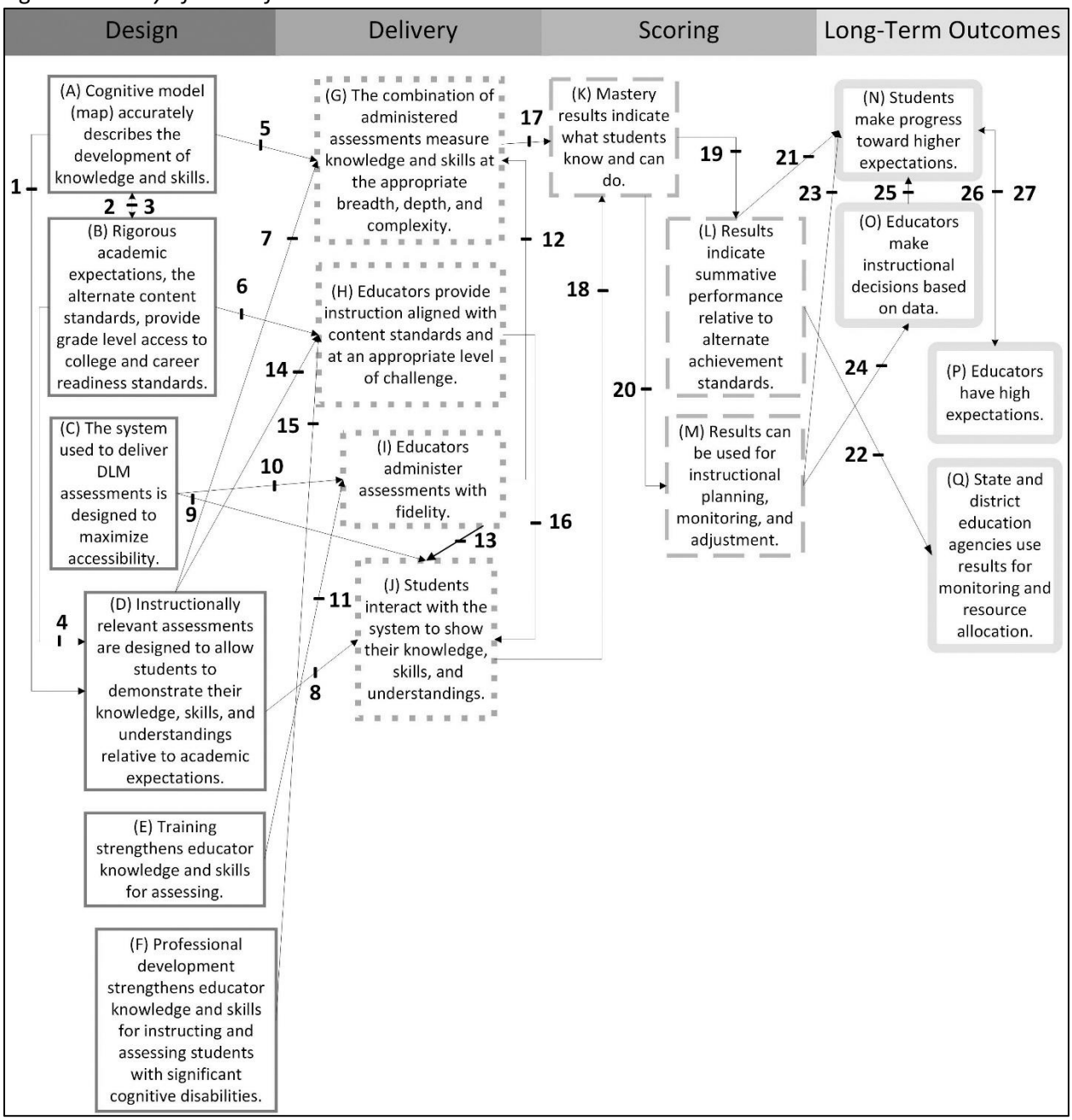
Page 1 of 2

For more information, including resources, please visit https://dynamiclearningmaps.org/states.

DLM assessments use a theory of action validity approach (Clark & Karvonen, 2021). The theory of action specifies claims for assessment design, delivery, scoring, and long-term outcomes (see Figure 3). Arrows between claims indicate the hypothesized chain of reasoning for how the long-term outcomes are ultimately achieved. Each claim in the theory of action has a set of underlying propositions, for which evidence is collected to evaluate the extent to which the proposition, and ultimately the claim, is defensible and supported. As an example, for Claim A, the cognitive model accurately describes the development of knowledge and skills, the underlying propositions include (1) nodes (i.e., knowledge, skills, or understandings) in the cognitive model are specified at the appropriate level of granularity and are distinct from other nodes; (2) nodes are sequenced according to acquisition

order; (3) nodes that are measured by the different levels of DLM assessments are correctly prioritized and are adequately spaced within breadth of the full map. Evidence is collected to evaluate each proposition (e.g., evidence of correct ordering; Thompson & Nash, 2022).

Figure 3. *Theory of Action for DLM Assessments*



Note: Letters indicate the claim. Numbers indicate relationships between claims.

COVID Context

Following the waiver of assessments in spring 2020 (*Recommended Waiver Authority Under Section 3511(d)(4) of Division A of the Coronavirus Aid, Relief, and Economic Security Act ("CARES ACT")*, 2020), the DLM Consortium convened a stakeholder group of state education agencies and DLM staff to consider potential impacts of the pandemic during the 2020-2021 academic year. The group determined five possible scenarios for instruction and assessment based on known circumstances in DLM states (Clark et al., 2021). The scenarios ranged from school resuming normally, school resuming with varying levels of disruption, or testing being halted again. For each scenario, the group determined which claims in the theory of action were likely to be impacted (Table 1) and any potential implications for score reporting if claims were impacted, given their intended uses (Table 2). For instance, under the scenario of school resuming normally (scenario 1), stakeholders would receive the standard score reports (i.e., both the Learning Profile and Performance Profile). However, under the scenario that school resumes but with multiple disruptions (scenario 3), delivery claims related to instruction and assessment administration might be impacted, which could necessitate the need for modified reporting that includes caveat language on score reports to support appropriate interpretation.

Table 1. *Validity Risks for Delivery and Scoring Claims*

	Claim						
Scenario	(G) Depth, breadth, complexity	(H) Instruction	(I) Fidelity	(J) Students use system	(K) Mastery	(L) Achievement levels	(M) Instruction Use
1	●	●	●	●	●	●	●
2	◒	⍰	●	●	●*	●*	●*
3	◒	⍰	◒	◒	◒	◒	◒
4	◒	⍰	◒	◒	◒	◒	◒
5	○	⍰	◒	◒	◒	○	◒

Note. ● = no risk, ◒ = partial risk, ○ = at risk, ⍰ = unknown risk.

* Conditional on the amount of quality instruction received.

1 = normal, 2 = alternate scheduling, 3 = multiple disruptions, 4 = entirely virtual, 5 = testing halted.

Table 2. *Level of Reporting by Scenario*

Level of reporting	Scenario				
	1 Normal	2 Alternative scheduling	3 Multiple disruptions	4 Entirely virtual	5 Testing halted
Overall achievement	●	◒	◒	◒	○
Fine-grained mastery	●	◒	◒	◒	◒

Note. ● = standard, ◒ = potentially modified, ○ = not provided.

Based on feedback from state education agencies, it was determined that in most instances instruction and assessment administration during 2020-2021 had some level of disruption (scenarios 2-4), but some students did resume school normally and attend for the full year. As such, a determination was needed for whether score reports should be modified to indicate potential impacts of the pandemic on instruction and assessment administration. We used the validity visualization approach to support our evaluation of actual impacts on claims on the theory of action.

Procedures

We first reviewed technical documentation and summarized sources of evidence collected through 2018-2019 (the last full administration year) for each theory of action claim (A-Q in Figure 3). We organized sources of evidence as procedural or empirical. For this initial exploration, we summed the evidence to create a visual display of evidence collected by claim. We used Excel color scales to indicate relative strength and weakness of the evidence. Areas with darker shading indicated the most evidence and areas shaded white indicated the least evidence.

We next evaluated the strength of 2020-2021 evidence, during which instruction and assessment administration were impacted by the COVID-19 pandemic. We first considered the existing evidence by theory of action claim and determined whether it was impacted by 2020-2021 disruptions. For instance, the strength of evidence collected for Design claims was judged to be unimpacted by the pandemic (e.g., the underlying cognitive model and standards remained the same). These were coded 0. However, evidence was impacted for some Delivery, Scoring, and Outcome claims. For instance, while test administration observations are collected annually, in 2020-2021 observations were limited and likely nonrepresentative because some students participated in remote learning and many schools were closed to visitors. Instances in which the source of evidence was likely impacted were coded -1. Finally, we listed areas with new evidence, such as the availability of new instructional resources intended to

support remote instruction practices and coded it +1. We summed the evidence by claim to determine the total impact.

Results

Table 3 lists an example source of procedural and empirical evidence for each claim. Procedural evidence ranges from description of system elements and methods to research literature. Empirical evidence includes data collected from a variety of sources, such as external ratings, survey results and focus group feedback. Note that the grain size of evidence varies across claims; each source was merely listed.

Table 3. Example Sources of Evidence by Claim

Claim	Procedural	Empirical
A: Cognitive model	Procedures for external review of map structure	Alignment: skill to item content
B: Content standards	Procedures for developing content standards	Alignment: alternate content standards to college and career ready standards
C: Accessible system	Description of development of accessible system components	Teacher survey responses on system accessibility
D: Assessments	Description of item writing process	Field test item statistics
E: Training	Description of facilitated and self-directed modules	Teacher survey responses on training and resources
F: Professional development	Scope of modules	Module ratings
G: Depth, breadth, complexity	Description of pool depth	Blueprint coverage
H: Instruction	Description of instructional support resources	Opportunity to learn data
I: Fidelity	Test administration manual	Test administration observation data
J: Students use system	Description of student response procedures	Test administration observation data
K: Mastery	Diagnostic scoring method	Mastery data
L: Achievement results	Standard setting methods	Impact data
M: Instructional use	Description of progress report delivery in system	Teacher interview data
N: Student progress	Research literature for students making progress over time	Results over time
O: Instructional decisions	Research literature for supporting instructional decision making	Focus group responses
P: Higher expectations	Research literature for improving expectations	Postsecondary opportunities ratings data
Q: District Use	Description of district uses	District use data

Table 4 summarizes the counts of evidence by claim through 2019. Claims D, C, and I have darker shading and correspondingly the most evidence, while Claims O and P have lighter shading and the least evidence. Most evidence was for claims about the design of the assessment (*n* = 102); the amount of evidence decreased for claims about delivery (*n* = 54), scoring (*n* = 25) and outcomes (*n* = 19). This gradual decrease in the quantity of evidence is consistent with the expectation (AERA et al., 2014; Cizek, 2020; Kane, 2020) that test developers provide most of the evidence for technical aspects of the assessment and local testing organizations provide most of the evidence for the local use and interpretation of an assessment (i.e., outcomes). The amount of procedural and empirical evidence was about equal within each cluster of claims (design, delivery, scoring, and outcomes), though overall there was more procedural evidence (*n* = 110) than empirical evidence (*n* = 90). This is also consistent with common practice, as most evidence collected by test developers relies on the processes and procedures for test design and development (e.g., item development, bias/sensitivity reviews, etc.), rather than empirical analysis or experimentation (Cizek, 2020), which may not always be feasible in educational contexts.

Table 4. *Sources of Validity Evidence for System through 2019*

	Claim	Procedural	Empirical	Overall
Design	A: Cognitive model	5	8	13
	B: Rigorous academic expectations	10	4	14
	C: The system	8	11	19
	D: Instructionally relevant assessments	16	17	33
	E: Training	11	3	14
	F: Professional development	6	3	9
Delivery	G: Combination of assessments	8	6	14
	H: Provide instruction	6	4	10
	I: Administer with fidelity	11	7	18
	J: Show their knowledge	4	8	12
Scoring	K: Mastery results	4	4	8
	L: Alternate achievement standards	5	4	9
	M: Results can be used for planning	5	3	8
Outcomes	N: Students make progress	3	3	6
	O: Educators make decisions	1	2	3
	P: Educators have high expectations	1	2	3
	Q: State and district use	6	1	7

Note: The summed sources of evidence come from the expanded version of Table 1

Table 5 summarizes the impact to validity evidence during the disrupted 2020-2021 year. COVID-19 disruptions impacted the collection of validity evidence in different ways. New sources of evidence included instructional resources to support educators during pandemic conditions and updated guidance for administration, including for off-site in-person administration by trained test administrators. Some sources were impacted by representation challenges (e.g., teacher survey data) while other studies were limited or unable to be conducted (e.g., test administration observations). As might be anticipated, the strength of evidence for Delivery claims was most impacted during the 2020-2021 year.

Table 5. Sources of Evidence Likely Impacted by 2020-2021 Disruptions

	Claim	Procedural	Empirical	Overall
Design	A: Cognitive model	0	0	0
	B: Rigorous academic expectations	0	0	0
	C: The system	0	0	0
	D: Instructionally relevant assessments	0	0	0
	E: Training	0	0	0
	F: Professional development	1	0	1
Delivery	G: Combination of assessments	0	-2	-2
	H: Provide instruction	1	-3	-2
	I: Administer with fidelity	0	-4	-4
	J: Show their knowledge	-2	-3	-5
Scoring	K: Mastery results	0	-1	-1
	L: Alternate achievement standards	0	-2	-2
	M: Results can be used for planning	0	0	0
Outcomes	N: Students make progress	0	-2	-2
	O: Educators make decisions	0	0	0
	P: Educators have high expectations	0	0	0
	Q: State and district use	-1	0	-1

Because the theory of action represents a theory of change, impacts to Delivery claims impact subsequent claims as well. Directional arrows linking claims in the theory of action represent if-then statements. For instance, in the theory of action claim G points to claim M; if students are assessed on the appropriate depth, breadth, and complexity of assessments, then mastery results reflect what they know and can do. However, when instruction and assessment are disrupted and students are absent or remote and do not test on the full breadth of academic content, mastery results may not fully reflect student knowledge and skills.

Given known disruptions to instruction and assessment and the range of claims impacted by evidence collection challenges in 2020-2021, the decision was made to err on the side of caution and provide modified score reports that included caveat language to help support appropriate interpretation and use of results. This decision was made in light of guidance that reports themselves should contain information to support appropriate interpretation and use and the extent to which evidence supports those uses (e.g., AERA et al., 2014; Zenisky & Hambleton, 2013). Caveat language indicated: *The 2020–2021 academic year was significantly impacted by the COVID-19 pandemic. Results may reflect the unusual circumstances for instruction and assessment this year. Use results with caution.* Caveat language appeared at the top of both the Performance Profile and Learning Profile, along with all aggregated reports (class, school, district, and state) in red font.

Discussion

Validity is known for being a complex and evolving aspect of educational measurement (Kane, 2016; Russell, 2021). Being able to readily visualize the extent of validity evidence can support test developers in knowing relative strength of validity evidence by claim and the extent to which intended uses are supported. Consistent with the need for ongoing validation, this method can adapt to include new findings or account for changes to the intended use and interpretation of a test. Using this method, test developers and testing organizations can monitor validity evidence to determine how it varies over time or, as in the current context, in response to a disruption to test administration. In a typical administration year, the visualization can support prioritization of future operational studies. In disrupted years, like 2021, visualization can support evaluation of evidence and determining whether intended uses are supported in light of the disruptions.

Visualizing validity evidence may also support testing organizations and researchers when reporting validity evidence. Reviewers have found that only a portion of literature related to educational measurement reports evidence of validity, possibly due to the varied and inconsistent frameworks for validity arguments (Carney et al., 2019) or the large learning curve that education professionals must overcome to build expertise in validity theory (Lavery et al., 2020). Additionally, validity evidence for assessments may be difficult for users to find or not publicly accessible (Boyer & Landl, 2021). Incorporating visualization into annual technical reports could make it easier for stakeholders to identify the relative strengths and weaknesses in a validity argument.

We caution, as Kane (2013) discusses, that the amount of evidence collected by claim is expected to vary. Users should not assume all claims require the same amount of evidence. Test developers may also prioritize collecting evidence for claims at different stages. For DLM assessments,

Design claims (claims A-F) were prioritized, and long-term outcomes (claims N-Q) will be evaluated over time as change is enacted. Similarly, evidence collected in a disrupted administration year may prioritize evaluating population representation and equity (Ho, 2021), and some claims may be more impacted that others (e.g., delivery impacted more than design). Visualization approaches should similarly account for the presence of disconfirming evidence, since validation should not serve as a merely confirmationist exercise (Kane, 2006), and contradictory explanations should be explored (AERA et al., 2014). The 2021 evidence coding process we shared here accounted for evidence that was impacted by pandemic conditions by coding them -1; there are likely other approaches that could similarly reflect the range of evidence programs collect.

The preliminary version shared here served its purpose for evaluating 2021 impacts and evidence collection. There are myriad ways developers can expand this preliminary visualization approach to suit their programs at varying programmatic stages or when disruptions occur. Future iterations may expand the tables to reflect the *Standards'* five categories of evidence or weight evidence sources by contribution strength. Future iterations may also consider more sophisticated approaches for visualizing evidence, its relative strength, recency, or changes in evidence collection over time.

References

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

Boyer, M., & Landl, E. (2021, April). Interim assessment practices for students with disabilities (NCEO Brief #22). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes and National Center for the Improvement of Educational Assessment. https://files.eric.ed.gov/fulltext/ED613028.pdf

Camara, W., Mattern, K., Croft, M., Vispoel, S., & Nichols, P. (2019). A validity argument in support of the use of college admissions test scores for federal accountability. *Educational Measurement: Issues and Practice, 38.* https://doi.org/10.1111/emip.12293

Carney, M., Crawford, A., Siebert, C., Osguthorpe, R., & Thiede, K. (2019). Comparison of two approaches to interpretive use arguments. *Applied Measurement in Education, 32*(1), 10-22. https://doi.org/10.1080/08957347.2018.1544138

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment*. Thousand Oaks, CA: SAGE Publications, Inc.

Cizek, G. J. (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy & Practice, 23*(2), 212–225. https://doi.org/10.1080/0969594X.2015.1063479

Cizek, G. J. (2020). Validity: An integrated approach to test score meaning and use. New York, NY: Taylor & Francis Group.

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

Clark, A. K., & Karvonen, M. (2021). Instructionally embedded assessment: Theory of action for an innovative system. *Frontiers in Education, 6*(724938). https://doi.org/10.3389/feduc.2021.724938

Clark, A. K., Thompson, W. J., Kobrin, J., Kavitsky, E., & Karvonen, M. (2021). *The impact of COVID-19: Validity considerations and scoring and reporting in flexible scenarios* (Project Brief No. 21–01). University of Kansas, Accessible Teaching, Learning, and Assessment Systems (ATLAS). https://dynamiclearningmaps.org/sites/default/files/documents/publication/Project%20Brief%2021_01_COVID%20Impacts.pdf

Dadey, N., Keng, L., Boyer, M., & Marion, S. (2021). *Making sense of spring 2021 assessment results*. Dover, NH: The National Center for the Improvement of Educational Progress. https://www.nciea.org/sites/default/files/publications/CFA-MakingSenseSpring2021Assess-R2.pdf

Dynamic Learning Maps (DLM) Consortium. (2016). *2014-2015 Technical Manual—Integrated Model*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems (ATLAS). https://dynamiclearningmaps.org/sites/default/files/documents/publication/Technical_Manual_IM_2014-15.pdf

Dynamic Learning Maps Consortium. (2019). *2018-2019 Technical Manual Update—Integrated Model.* University of Kansas, Accessible Teaching, Learning, and Assessment Systems (ATLAS). https://dynamiclearningmaps.org/sites/default/files/documents/publication/2018-2019_IM_Technical_Manual_Update.pdf

Hatala, R., Cook, D. A., Brydges, R., & Hawkins, R. (2015). Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. Advances in health sciences education : theory and practice, 20(5), 1149–1175. https://doi.org/10.1007/s10459-015-9593-1

Ho, A. (2021). Three test-score metrics that all states should report in the COVID-19 affected spring of 2021. https://scholar.harvard.edu/files/andrewho/files/threemetrics.pdf

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.

Kane, M. (2013a). The argument-based approach to validation. *School Psychology Review, 42*(4), 448–457. https://doi-org/10.1080/02796015.2013.12087465

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

Kane, M. (2020). Validity studies commentary. *Educational Assessment, 25*(1). P. 83-89. https://doi.org/10.1080/10627197.2019.1702465

Kane, M. T. & Wools, S. (2019). Perspectives on the validity of classroom assessment. In Brookhart, S. M., & McMillan, J. H. (Eds.). *Classroom assessment and educational measurement*. New York, NY: Taylor & Francis Group.

Ketterlin-Geller, L. R., Perry, L., & Adams, E. (2019). Integrating validation arguments with the assessment triangle: A framework for operationalizing and instantiating validation. *Applied Measurement in Education, 32*(1), 60–76. https://doi.org/10.1080/08957347.2018.1544136

Lavery, M.R., Bostic, J.D., Kruse, L., Krupa, E.E. and Carney, M.B. (2020), Argumentation Surrounding Argument-Based Validation: A Systematic Review of Validation Methodology in Peer-Reviewed Articles. *Educational Measurement: Issues and Practice, 39*, 116-130. https://doi.org/10.1111/emip.12378

Multi-State Alternate Assessment. (2021). *Multi-State Alternate Assessment 2021 Technical Report.* Cognia. https://www.azed.gov/sites/default/files/2021/10/2020-21%20MSAA%20Technical%20Report_ADA.pdf

National Academy of Education. (2021). *Educational Assessments in the COVID-19 Era and Beyond (p. 22)*. National Academy of Education. https://naeducation.org/wp-content/uploads/2021/02/Educational-Assessments-in-the-COVID-19-Era-and-Beyond.pdf

Nichols, P. D., Gianopulos, G., (2021). Arguing about the effectiveness of assessments for the classroom. *Journal of Mathematical Behavior, 61.* https://doi.org/10.1016/j.jmathb.2020.100839

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Rosenblum, I. (2021, February 22). *Letter from the Assistant Secretary: An update on assessment, accountability, and reporting requirements for the 2020-2021 school year.* United States Department of Education, Office of Elementary and Secondary Education. https://www2.ed.gov/policy/elsec/guid/stateletters/dcl-assessments-and-acct-022221.pdf

Schilling, S. G., & Hill, H. C. (2007). Assessing measures of mathematical knowledge for teaching: A validity argument approach. Measurement: Interdisciplinary Research and Perspectives, 5(2–3), 70–80. https://doi.org.10.1080/15366360701486965

Schilling, S. & Hill, H. (2007). Assessing Measures of Mathematical Knowledge for Teaching: A Validity Argument Approach. *Measurement 5*, 70-80. https://doi.org.10.1080/15366360701486965

Sireci, S. G., Lim, H., Rodriguez, G., Banda, E., & Zenisky, A. (2018). *Evaluating criteria for validity evidence based on test content*. Annual meeting of the National Council on Measurement in Education, New York, NY.

Sireci, S. G., & Suarez-Alvarez, J. (2022). Deriving Decisions from Disrupted Data. *Educational Measurement: Issues and Practice, 41*(1), 23–27. https://doi.org/10.1111/emip.12499

Smarter Balanced Assessment Consortium. (2020, November). *2018-19 Summative Technical Report.* Retrieved from https://technicalreports.smarterbalanced.org/2018-19_summative-report/_book/

Smarter Balanced Assessment Consortium. (2021, August). *2020-21 Interim Technical Report.* Retrieved from https://technicalreports.smarterbalanced.org/interim_rep2021/_book/index.html

Thompson, W. J. (2019). Bayesian psychometrics for diagnostic assessments: A proof of concept (Research Report No. 19-01). Lawrence, KS: University of Kansas, Accessible Teaching, Learning, and Assessment Systems. https://dynamiclearningmaps.org/sites/default/files/documents/publication/Bayes_Proof_Concept_2019.pdf

Thompson, W. J., & Nash, B. (2022). A diagnostic framework for the empirical evaluation of learning maps. *Frontiers in Education, 6*. https://www.frontiersin.org/article/10.3389/feduc.2021.714736

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.* https://www2.ed.gov/admins/lead/account/saa/assessmentpeerreview.pdf

Whalen, A. (October 6, 2016). *Letter to Chief State School Officers Regarding Assessment Peer Review Outcomes.* United States Department of Education Office of Elementary and Secondary Education. https://www2.ed.gov/admins/lead/account/saa/dcletterassepeerreview1072016ltr.pdf

Wools, S., Eggen, T., & Sanders, P. (2010). Evaluation of validity and validation by means of the argument based approach. Giornale Italiano di Pedagogia Sperimentale, 8(1), 63-82. https://doi.org/10.3280/CAD2010-001007

Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice, 31*(2), 21–26. https://doi.org/10.1111/j.1745-3992.2012.00231.x

Appendix A

The Multi-State Alternate Assessment
Validity Evaluation

Figure A.1. Relevance criteria and outcomes for MSAA validity arguments

Complete Evidence	When all required pieces of relevant evidence are provided to support a validity argument
Moderate to Substantial Evidence	When several pieces of relevant evidence are provided, but not all required pieces of evidence are provided
Limited Evidence	When only one or two pieces of evidence are provided, where the evidence may be only marginally relevant or where more than 1–2 pieces of evidence are required
No Evidence	When no relevant evidence exists

Note: MSAA ratings reflect the applicability and completeness of evidence. Ratings do not reflect persuasiveness.
Excerpt from Multi-State Alternate Assessment. (2021). Multi-State Alternate Assessment 2021 Technical Report. Cognia. https://www.azed.gov/sites/default/files/2021/10/2020-21%20MSAA%20Technical%20Report_ADA.pdf

Figure A.2. Example of the narrative summary and evaluation of evidence for the MSAA

Assumption 4.1. Parents find MSAA scores and other information useful for understanding what their child knows and can do.

Element 4.1.1. Parents understand and interpret correctly MSAA scores and other information to understand what their child knows and can do.

Evidence: MSAA provides information to guide parents in interpreting and using MSAA scores and other information about their child's achievement and learning needs. For example, the Arizona Department of Education sends to districts a Parent Overview to accompany each child's Individual Score Report. The overviews are available online in both English and Spanish; see http://www.azed.gov/assessments/parents/. Similarly, the Maine Department of Education provides the Parent Overview of the MSAA Assessment System (see https://www.maine.gov/doe/sites/maine.gov.doe/files/inline-files/2016ParentOverview-allgradescombined.pdf).

Summary of evidence: Limited Evidence; an example of additional evidence could be a survey of parents to begin to understand the degree to which parents correctly understand and interpret MSAA scores and other MSAA-based information to understand what their child knows and can do.

Excerpt from Multi-State Alternate Assessment. (2021). Multi-State Alternate Assessment 2021 Technical Report. Cognia. https://www.azed.gov/sites/default/files/2021/10/2020-21%20MSAA%20Technical%20Report_ADA.pdf

Figure A.3. Excerpt of the summary table for the MSAA validity evaluation

Table 11.3 Status of Evidence for All Four SIUs, Assumptions, and Elements

Element	Relevance of the Evidence to the Element			
	No Evidence Exists Currently	Limited	Moderate to Substantial	Complete
Primary Intended Score Interpretation				
MSAA scores provide reliable and valid information about important knowledge and skills in grade-level numeracy and literacy that students with the most significant cognitive disabilities are attaining.				
1.1.1 MSAA content is aligned to the CCCs and grade-level standards.				X
1.1.2 MSAA items are aligned to the CCCs.				X
1.1.3 States have confirmed alignment of the MSAA to state content standards.				X
1.1.4 MSAA items are aligned to the PLDs.			X	
1.2.1. Items require application of the KSAs of the targeted construct.			X	
1.2.2. Items are accessible to all students.			X	
1.2.3. Appropriate accommodations are provided to meet student needs.			X	
1.2.4. Scoring rubrics focus on construct-relevant aspects of student responses.			X	
1.2.5. Scaffolding is not a source of construct-irrelevant variance.			X	
1.2.6. Item rendering does not interfere with student access to test content.			X	

Appendix B

The Smarter Balance Assessment System

Validity Argument

Figure B.1. Sources of Evidence for Smarter Balanced Summative Assessment Scores

Table 1.1: SMARTER BALANCED ASSESSMENT PURPOSES CROSS-CLASSIFIED BY SOURCES OF VALIDITY EVIDENCE

Purpose	Sources of Validity Evidence
1. Report achievement with respect to the CCSS as measured by the ELA/literacy and mathematics summative assessments in grades 3 to 8 and high school.	A. Test Content
	B. Internal Structure
	C. Response Processes
	D. Relation to Other Variables
2. Assess whether students prior to grade 11 have demonstrated sufficient academic proficiency in ELA/literacy and mathematics to be on track for achieving college and career readiness.	A. Test Content
	B. Internal Structure
	C. Response Processes
	D. Relation to Other Variables
3. Assess whether grade 11 students have sufficient academic proficiency in ELA/literacy and mathematics to be ready to take credit-bearing, transferable college courses after completing their high school coursework.	A. Test Content
	B. Internal Structure
	C. Response Processes
	D. Relation to Other Variables
4. Measure students' annual progress toward college and career readiness in ELA/literacy and mathematics.	A. Test Content
	B. Internal Structure
	C. Response Processes
	D. Relation to Other Variables
5. Inform how instruction can be improved at the classroom, school, district, and state levels.	A. Test Content
	B. Internal Structure
	C. Response Processes
6. Report students' ELA/literacy and mathematics proficiency for federal accountability purposes and potentially for state and local accountability systems.	A. Test Content
	B. Internal Structure
	C. Response Processes
7. Assess students' achievement in ELA/literacy and mathematics in a manner that is equitable for all students and targeted student groups.	A. Test Content
	B. Internal Structure
	C. Response Processes

Excerpt from Smarter Balanced Assessment Consortium. (2020, November). 2018-19 Summative Technical Report. Retrieved from https://technicalreports.smarterbalanced.org/2018-19_summative-report/_book/

Figure B.2. Example of a validity argument for the Smarter Balanced Summative Assessment

1.5.1 Validity Evidence Supporting Purpose 1

Purpose 1: Provide valid, reliable, and fair information about students' ELA/literacy and mathematics achievement with respect to those Common Core State Standards (CCSS) measured by the ELA/literacy and mathematics summative assessments in grades 3 to 8 and high school.

Source A: Test Content

In This Report:

- Chapter 4
  - Test blueprint, content specifications, and item specifications are aligned to the full breadth and depth of grade-level content, process skills, and associated cognitive complexity.
  - Blueprint fidelity studies are performed for each test administration for regular and accommodated populations. They are performed prior to test administration by simulation and following test administration using member data.
  - With very few exceptions, operational computer adaptive test events meet all blueprint constraints, both for the general student population and for students taking accommodated test forms.

List of Other Evidence Sources:

- Smarter Balanced Content Specifications (Smarter Balanced Assessment Consortium, 2017a,b)
- Evaluating the Content and Quality of Next Generation High School Assessments (Schultz, Michaels, Dvorak, & Wiley, 2016)
- Evaluating the Content and Quality of Next Generation Assessments (Doorey & Polikoff, 2016)
- Smarter Balanced Assessment Consortium: Alignment Study Report (HumRRO, November 2016)
- Evaluation of the Alignment Between the Common Core State Standards and the Smarter Balanced Assessment Consortium Summative Assessments for Grades 3, 6, and 7 in English Language Arts/Literacy and Mathematics – Final Report (WestEd Standards, Assessment, and Accountability Services Program, 2017)
- 2017–18 Smarter Balanced Summative CAT Simulation Results (American Institutes for Research, 2017)
- Blueprint Fidelity of the 2017–18 Summative Assessment (Smarter Balanced Assessment Consortium, 2019)

Note: This is a portion of the validity argument for Smarter Balanced primary purpose 1. The complete argument contains additional types of evidence and can be found in the Smarter Balanced 2018-2019 Summative Technical Report. The Smarter Balanced technical documentation does not provide an evaluation of the validity argument.

Excerpt from Excerpt from Smarter Balanced Assessment Consortium. (2020, November). 2018-19 Summative Technical Report. Retrieved from https://technicalreports.smarterbalanced.org/2018-19_summative-report/_book/

Appendix C

Dynamic Learning Maps Assessment (2014 – 2019)
Validity Argument

Figure C.1. Summary of Primary Claims and Quantity of New Evidence for the 2018-2019 DLM Assessment Systems

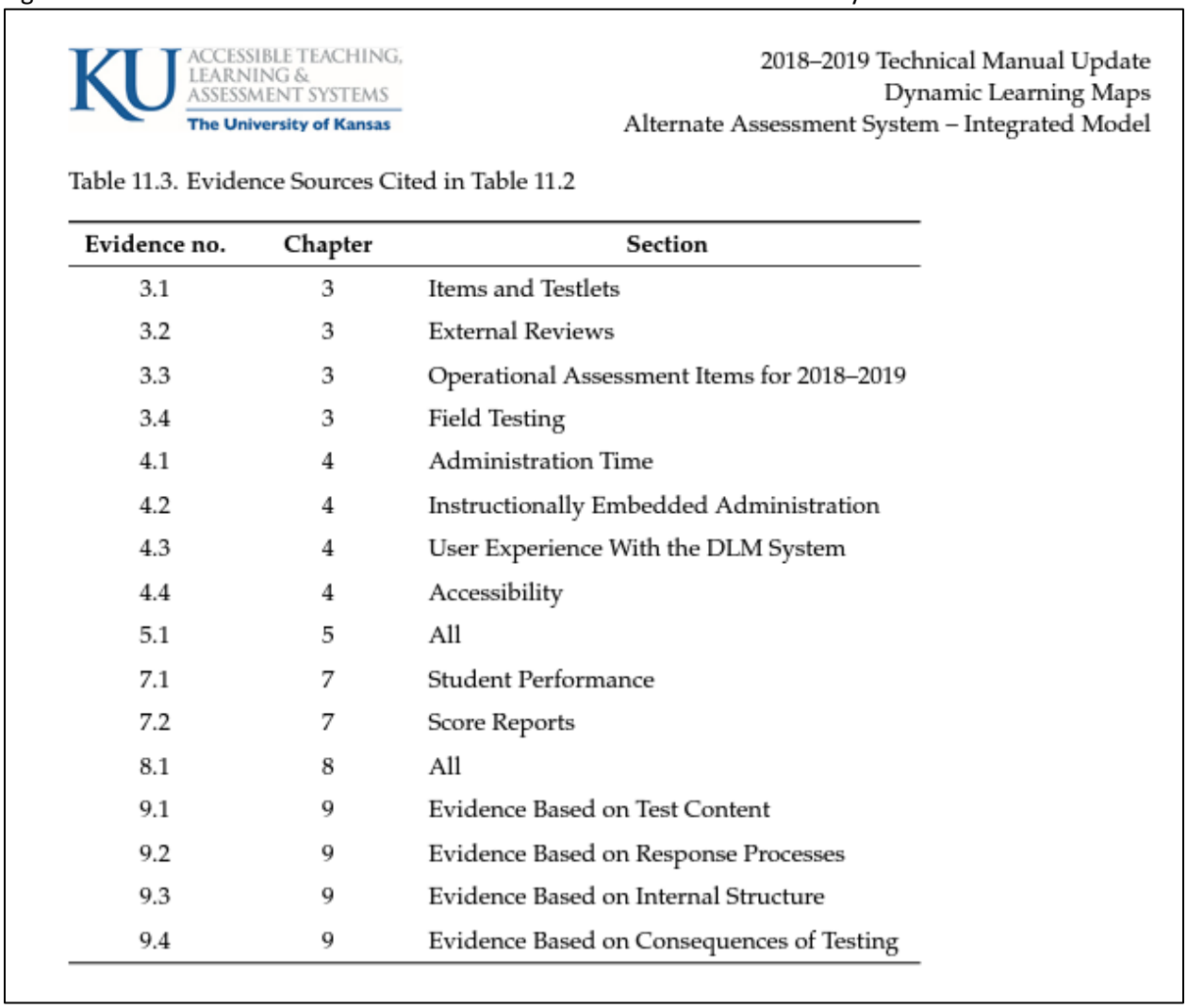
Table 11.2. DLM Alternate Assessment System Claims and Sources of Updated Evidence for 2018–2019

Claim	Sources of evidence*				
	Test content	Response processes	Internal structure	Relations with other variables	Consequences of testing
1. Scores represent what students know and can do.	3.1, 3.2, 3.3, 3.4, 4.1, 4.2, 4.3, 7.1, 7.2, 9.1	4.1, 4.2, 4.3, 4.4, 9.2	3.3, 3.4, 5.1, 8.1, 9.3		7.1, 7.2, 9.4
2. Achievement level descriptors provide useful information about student achievement.	7.1, 7.2		8.1		7.1, 7.2, 9.4
3. Inferences regarding student achievement can be drawn at the conceptual area level.	7.2, 9.1		8.1		7.2, 9.4
4. Assessment scores provide useful information to guide instructional decisions.					9.4

Note. * See Table 11.3 for a list of evidence sources. Only direct sources of evidence are listed. Some propositions are also supported indirectly by evidence presented for other propositions.

Excerpt from Dynamic Learning Maps Consortium. (2019). 2018-2019 Technical Manual Update—Integrated Model. University of Kansas, Accessible Teaching, Learning, and Assessment Systems (ATLAS). https://dynamiclearningmaps.org/sites/default/files/documents/publication/2018-2019_IM_Technical_Manual_Update.pdf

Figure C.2. List of sources of evidence for the 2018-2019 DLM Assessment Systems



KU ACCESSIBLE TEACHING, LEARNING & ASSESSMENT SYSTEMS
The University of Kansas

2018–2019 Technical Manual Update
Dynamic Learning Maps
Alternate Assessment System – Integrated Model

Table 11.3. Evidence Sources Cited in Table 11.2

Evidence no.	Chapter	Section
3.1	3	Items and Testlets
3.2	3	External Reviews
3.3	3	Operational Assessment Items for 2018–2019
3.4	3	Field Testing
4.1	4	Administration Time
4.2	4	Instructionally Embedded Administration
4.3	4	User Experience With the DLM System
4.4	4	Accessibility
5.1	5	All
7.1	7	Student Performance
7.2	7	Score Reports
8.1	8	All
9.1	9	Evidence Based on Test Content
9.2	9	Evidence Based on Response Processes
9.3	9	Evidence Based on Internal Structure
9.4	9	Evidence Based on Consequences of Testing

Excerpt from Dynamic Learning Maps Consortium. (2019). 2018-2019 Technical Manual Update—Integrated Model. University of Kansas, Accessible Teaching, Learning, and Assessment Systems (ATLAS). https://dynamiclearningmaps.org/sites/default/files/documents/publication/2018-2019_IM_Technical_Manual_Update.pdf

Figure C.3. Evaluation of the Validity Evidence for the DLM Assessment System

Table 116. Evaluation of Evidence for Each Proposition.

Proposition	Overall Evaluation
1. Scores represent what students know and can do.	There is strong procedural evidence for content representation and response process. Alignment evidence for the operational assessment system is generally strong, although areas for improvement are noted. There is preliminary empirical response process evidence, although analysis will be ongoing. Evidence of internal structure is strong for this stage of the assessment program; future statistical modeling with additional data will provide stronger evidence.
2. Achievement level descriptors provide useful information about student achievement.	In 2014-15, the policy-level PLDs were reported. Grade and content-specific PLDs were developed for first use in 2015-16. Procedural evidence supports PLD relationship to the content and structure of the academic content standards. Additional evidence will be needed to evaluate the actual use of the descriptors.
3. Inferences regarding student achievement, progress, and growth can be drawn at the conceptual area level.	There is preliminary evidence to support the structure of the conceptual areas and the reporting of achievement in these areas. More substantial evidence, particularly for internal structure, will be gathered in future years. Evidence on inferences about measures of progress and growth will be collected once those are calculated and reported.
4. Assessment scores provide useful information that can guide instructional decisions.	Overall evidence is strong for the first year of the program. Stakeholders can interpret report contents and teachers can describe their use for instructional decision-making. Additional evidence is needed as the assessment program matures, including evidence of score use in school and program decision-making.

Excerpt from Dynamic Learning Maps (DLM) Consortium. (2016). 2014-2015 Technical Manual—Integrated Model. University of Kansas, Accessible Teaching, Learning, and Assessment Systems (ATLAS). https://dynamiclearningmaps.org/sites/default/files/documents/publication/Technical_Manual_IM_2014-15.pdf