frontiers in Education

HYPOTHESIS AND THEORY
published: 26 October 2021
doi: 10.3389/feduc.2021.724938

Instructionally Embedded Assessment: Theory of Action for an Innovative System

Amy K. Clark and Meagan Karvonen*

Accessible Teaching, Learning, and Assessment Systems, University of Kansas, Lawrence, KS, United States

Policy shifts in the United States are beginning to reduce the emphasis on using statewide assessment results primarily for accountability and teacher evaluation. Increasingly, there are calls for and interest in innovative and flexible assessments that shift the purposes of assessment and use of results toward instructional planning and student learning. Under the Innovative Assessment Demonstration Authority, some states are exploring options for replacing traditional large-scale summative assessments with innovative measures. However, many of these programs are still in early phases of planning and research and have not yet fully articulated how the innovative system achieves desired outcomes. This conceptual paper presents an argument in the form of a theory of action for a flexible and innovative assessment system already in operational use. The system replaces traditional summative assessments with large-scale through-year Instructionally Embedded assessments. We describe the components of the theory of action, detailing the theoretical model and supporting literature that illustrate how system design, delivery, and scoring contribute to the intended outcomes of teachers using assessment results to inform instruction and having higher expectations for student achievement, in addition to accountability uses. We share considerations for others developing innovative assessment systems to meet stakeholders' needs.

Keywords: innovative assessment, accountability, theory of action, instructionally embedded, large-scale assessment, alternate assessment, through-year assessment

OPEN ACCESS

Edited by:
Chad M. Gotch,
Washington State University,
United States

Reviewed by:
Carla Evans,
National Center for the Improvement of Educational Assessment,
United States
Divya Varier,
George Mason University,
United States

****Correspondence:***
Amy K. Clark
akclark@ku.edu

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *14 June 2021*
Accepted: *20 September 2021*
Published: *26 October 2021*

Citation:
Clark AK and Karvonen M (2021) Instructionally Embedded Assessment: Theory of Action for an Innovative System. Front. Educ. 6:724938. doi: 10.3389/feduc.2021.724938

INSTRUCTIONALLY EMBEDDED ASSESSMENT: EXPLICATING AN INNOVATIVE SYSTEM TO ELICIT CHANGE

While statewide summative assessments serve important purposes, they have been criticized for providing limited information. In response to that criticism, many states shifted to comprehensive or balanced assessment systems to provide different types of information about student learning and achievement throughout the year. Yet balanced systems have their own challenges, including increased time spent on testing. To take advantage of technology innovations and address the limitations of balanced assessment systems, some states are now considering innovative approaches. The purpose of this paper is to describe an innovative model in operational use in five states. We describe an argument for an Instructionally Embedded assessment model through the lens of the program's theory of action, illustrating how the system's design, delivery, and scoring contribute to intended outcomes.

CONTEXT FOR THE MODEL: SHIFTS IN STATEWIDE ASSESSMENT PROGRAMS

Changing education policy since 2000 contributed to substantial shifts in educational-assessment practices in the United States. Beginning with the No Child Left Behind Act of 2001 (2002) and later Race to the Top (2010), federal education policy emphasized using results from large-scale assessments in state accountability models and for teacher evaluation. These assessments were designed to differentiate student achievement between performance levels or around the cut point(s) and did not provide teachers with actionable information for instructional decision-making. Results were often at too coarse a grain size and delivered too late in the year to be instructionally useful (Marion, 2018; Wilson, 2018; Jacobson, 2019; Modan, 2020). Teachers and their unions pushed back against using assessment results for evaluating teachers' performance (Onosko, 2011; Olson and Jerald, 2020). Parents, teachers, students, and policymakers expressed concern that students spent too much academic time preparing for and taking assessments (Rentner et al., 2016; Olson and Jerald, 2020), and opt-out movements gained popularity (Mitra et al., 2016; Pizmony-Levy and Green Saraisky, 2016).

Researchers and practitioners have acknowledged the need for assessments that go beyond traditional, large-scale summative assessments and inform instruction (e.g., Council of Chief State School Officers, 2008; Pellegrino et al., 2016; Wilson, 2018). Comprehensive and balanced assessment systems—in which a variety of assessments administered throughout the year provide stakeholders with multiple sources of evidence for decision-making—emerged at the state and local levels to provide teachers with data throughout the year. These systems often include administration of formative and interim measures, in addition to summative assessments (Gong, 2010; Marion et al., 2019). While interim assessments can provide teachers with data throughout the year, they often provide a prediction of how the student will perform on the end-of-year assessment (Perie et al., 2009) rather than data to guide immediate instructional decisions. Formative assessments provide information that teachers can use to inform instruction (Heritage, 2008) but often require teacher design, which can detract from instructional planning time and may have implications for assessment quality (e.g., George and Sanders, 2017). The combination of formative, interim, and summative measures, which may be administered in addition to other district assessments such as benchmarks or progress monitoring, raises concerns about the amount of time students spend taking and preparing for assessments. Stakeholders emphasize the criticality of cohesion across the suite of assessments (Marion et al., 2019) and the complexity of teachers using data from multiple measures (Mandinach and Gummer, 2013; Farrell and Marsh, 2016).

Advances in technology have expanded capabilities of educational assessments (Veldkamp and Sluijter, 2019), positioning the field for innovative and flexible assessments that extend balanced assessment systems and provide teachers with actionable results. Diagnostic modeling (Bradshaw, 2017), principled assessment-design activities (Ferrara et al., 2017), and online data dashboards (Ahn et al., 2019) have advanced considerably in recent years. These and other advances allow for development of assessment practices that are ongoing, embedded in instruction, and informative to instructional practice (U.S. Department of Education, 2016; Bennett, 2018).

Beginning in 2018, state education agencies submitted applications under the Every Student Succeeds Act, 2015 (ESSA) Innovative Assessment Demonstration Authority (IADA) to produce innovative assessment systems, including competency-based, instructionally embedded, interim, cumulative year-end, or performance assessments, that would replace traditional summative assessments. New Hampshire, Louisiana, North Carolina, Georgia and Massachusetts submitted applications and were approved (U.S. Department of Education, 2020); Nebraska is also exploring an innovative approach. Louisiana, North Carolina, Georgia, and Nebraska are piloting assessments that are administered throughout the year, with some intending to replace summative assessments, while the New Hampshire model currently offers performance assessments in addition to traditional summative assessments (O'Keefe and Lewis, 2019; Hedger, 2020).

Another innovative assessment system, not submitted under IADA, is the Instructionally Embedded assessment system adopted by some states in the Dynamic Learning Maps (DLM) Alternate Assessment Consortium. This model replaces traditional summative assessments with a through-year model. The model prioritizes teacher choice within constraints, with teachers selecting content to meet requirements for breadth, depth, and complexity and when to administer assessments after instruction. Diagnostic modeling provides teachers with fine-grained, actionable skill-mastery profiles summarizing mastery of skills measured for each content standard (**Figure 1**). Results are combined throughout the year to produce summative results used for accountability purposes. Because the assessment is operationally administered in five states as their summative assessments, the system offers a unique perspective and opportunity to share lessons learned for programs seeking to adopt an innovative and flexible assessment model. Because the model is applied to an alternate assessment for students with the most significant cognitive disabilities, we also identify aspects of the system that are specific to this assessment population and which components are likely to be applicable to flexible assessments for all students.

INSTRUCTIONALLY EMBEDDED ASSESSMENT AS CHANGE AGENT

In addition to reporting assessment results, assessment programs sometimes intend that results elicit change in stakeholder behaviors (National Council on Measurement in Education (NCME), 2018). While literature often references assessment unintended outcomes (e.g., narrowing of the curriculum (Cawelti, 2006), reduced instructional time (Powell et al., 2009), among others), assessment programs can, and as

REPORT DATE: 04-23-2021
SUBJECT: English language arts
GRADE: 6

Individual Student End-of-Year Report
Learning Profile 2020-21

DYNAMIC® LEARNING MAPS

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

DISTRICT ID: 12345
STATE: DLM State

Student's performance in 6th grade English language arts Essential Elements is summarized below. This information is based on all of the DLM tests Student took during the 2020-21 school year. Grade 6 had 19 Essential Elements in 4 Areas available for instruction during the 2020-21 school year. The minimum required number of Essential Elements for testing in 6th grade was 9. Student was tested on 13 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.C1.1	ELA.EE.RI.6.5	Understand action verbs	Identify concrete details in an informational text	Understand the structural purpose of a text	Recognize that titles reflect text structure	Determine how a fact, step, or event fits the structure of a text
ELA.C1.2	ELA.EE.RL.6.1	Differentiate between text and pictures	Identify characters, setting, and major events	Recognize explicit information in a story	Determine explicit and implicit information	Identify where explicit information is stated and where inferences can be drawn
ELA.C1.2	ELA.EE.RL.6.2	Match a picture representation with a real object	Identify concrete details in a familiar story	Identify main idea	Identify details related to theme in a story	Identify events related to theme in a story

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

Page 1 of 4

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.

FIGURE 1 | Learning Profile portion of individual student score reports.

argued by Gordon (2020) and others, should, be designed to elicit positive change in learning and instruction. A theory of action can guide development, use, and evaluation of an assessment designed for that purpose (National Council on Measurement in Education (NCME), 2018).

The DLM Instructionally Embedded assessment program is designed for educators to use results to inform instruction and addresses many of the previously described challenges of traditional, large-scale standardized assessments. Instruction and assessment are linked in the system; teachers create instructional plans indicating the standard and level for instruction and system-assigned assessment content. The total time spent taking assessments is spread throughout the year, in short (approximately 5-minute) administrations of instructionally relevant assessments. Results are provided on demand as instruction and assessment occur and at a fine-grained level to support subsequent instruction. Diagnostic skill-mastery results are combined to produce summative reporting information describing overall achievement in the subject for use in state accountability models. This innovative approach to administration and reporting is designed to support educators in instructional planning using assessment data, in addition to traditional uses, such as state and district education agencies using results for educational program monitoring and resource allocation.

Another goal for the DLM Instructionally Embedded assessment program is to enact change by addressing persistent low expectations held for students with the most significant cognitive disabilities. These students have historically been taught a largely functional curriculum to prepare them for independent living (Ryndak et al., 2014) and only recently were mandated to be included in statewide academic assessments used for accountability (NCLB, 2001). Their achievement was often described as proficient despite their instruction on relatively low-level academic skills (Altman et al., 2010; Nitsch, 2013). Research shows that students with significant disabilities can and do learn academic content when provided with appropriate opportunity and instruction (Allor et al., 2010; Geist et al., 2014). They can interact with various forms of technology (Burnes and Clark, 2021) and contribute to their own goal setting and decision-making (Wehmeyer et al., 2012; Carter et al., 2013). The DLM program recognizes these findings and seeks to increase teachers' expectations for student learning and for students to make progress toward higher expectations.

Because the innovative Instructionally Embedded assessment model is intended to elicit change in stakeholder behaviors (e.g., teachers use assessment data), we adopt the National Council on Measurement in Education (NCME) (2018) recommendation to describe how changes will be achieved over time using a theory of action. We present the theory of action here as an example for innovative assessment systems designed to elicit change, specifically detailing the theoretical model for how the system is designed, delivered, and scored to achieve intended outcomes.

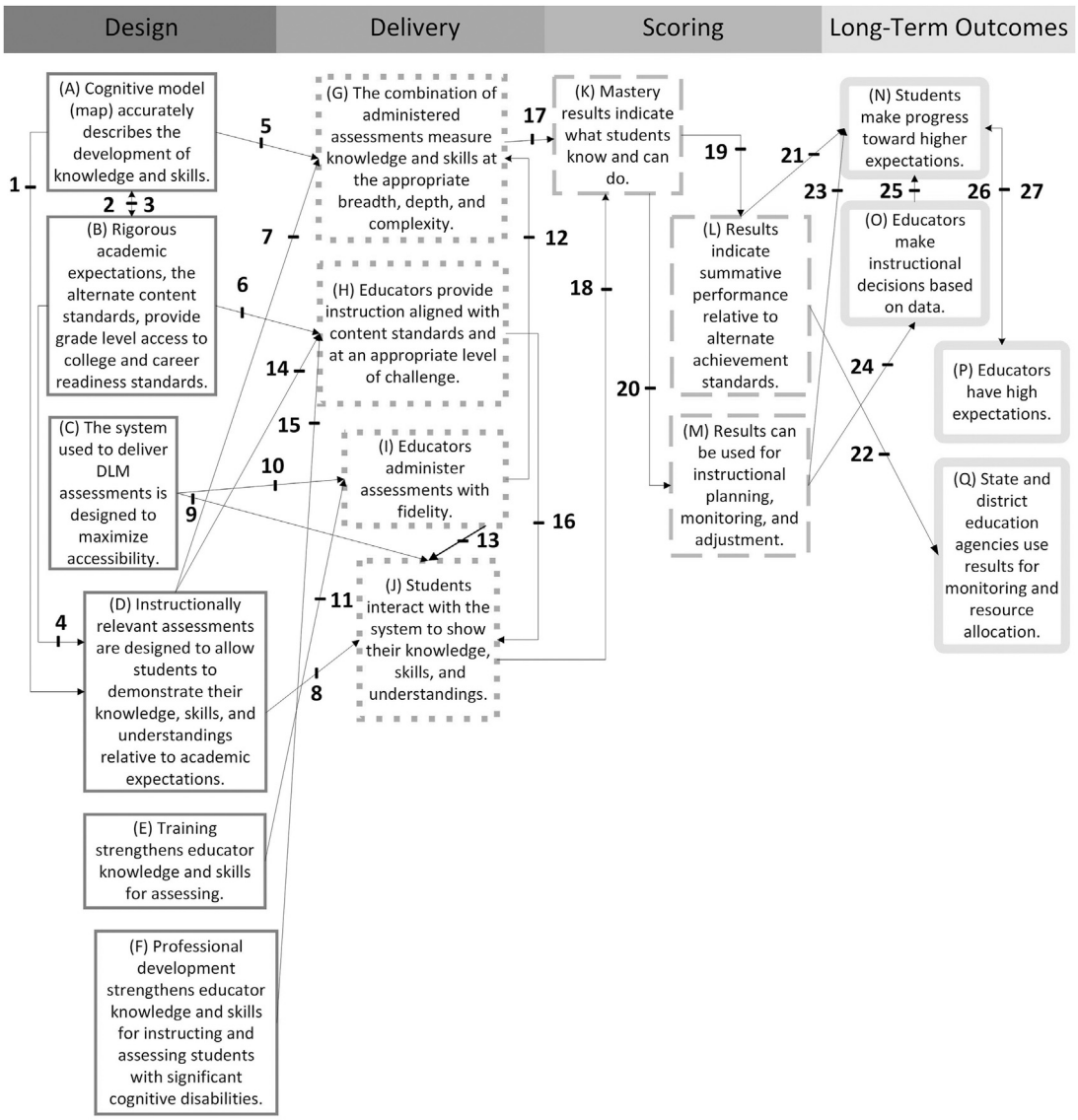


FIGURE 2 | Theory of action for the Dynamic Learning Maps Instructionally Embedded assessment system.

While this example is for an alternate assessment based on alternate academic achievement standards and some claims pertain to increasing expectations for students with significant cognitive disabilities, much of the theoretical foundation applies to other innovative assessment systems, including those for the general population of students (e.g., claims about test design, accessibility). We close by discussing considerations for developers of other innovative and flexible assessment programs.

THEORY OF ACTION FOR INSTRUCTIONALLY EMBEDDED SYSTEM

Required for Race to the Top (2010) grant applicants, theories of action depict causal models that include posited claims and relationships between claims (National Council on Measurement in Education (NCME), 2018). While their use originated in the program evaluation literature (e.g., Patton, 1989), programs have adopted theories of action for state accountability systems (Marion et al., 2016), educator development (Graziano et al., 2017), assessment systems (e.g., Bennett, 2010; Sireci, 2015), and assessment types (Council of Chief State School Officers, 2018; Gholson and Guzman-Orth, 2019). Theories of action are also used in assessment validation; claims in the theory of action are evaluated by collecting evidence to determine the extent to which they are supported and defensible (Chalhoub-Deville, 2016; Clark and Karvonen, 2020).

A theory of action's strength is derived from the logical argument upon which it is based. That is, there should be a sufficient theoretical rationale for the claims and relationships in

the model. We present an example for how a theory of action for an innovative assessment system documents the process for realizing programmatic change. In articulating this example, we drew from methods described by Jaakkola (2020) for conceptual papers, whereby authors develop the logical argument for an underlying model, providing the theoretical explanation for the model and the relationships depicted rather than summarizing empirical evidence. Specifically, we detail the logical argument for the claims and causal mechanisms in the theory of action and cite relevant literature to demonstrate the rationale for their inclusion so others can draw from it. Empirical evidence evaluating the validity argument is beyond the scope of this article (Clark and Karvonen, 2020).

Figure 2 shows the theory of action for the Instructionally Embedded assessment system. The model reads left to right. Rectangular boxes denote claims, organized within the four sections of the diagram: Design, Delivery, Scoring, and Long-Term Outcomes. The chain of reasoning, or theory of change, is demonstrated broadly by the ordered nature of the four sections and explicitly by the numbered arrows between claims. Design claims serve as inputs and inform delivery and implementation of the assessment system, which informs scoring and reporting, and the long-term outcomes for various stakeholders. Within and across the four sections, the chain of reasoning specifies a series of if–then statements hypothesizing relationships between individual claims. For instance, if the cognitive model is correctly specified, then students can be assessed at the appropriate depth, breadth, and complexity (shown as relationship 5 in **Figure 2** between claims A and G). Most relationships are unidirectional, but some are bidirectional (i.e., relationship 2/3 and 26/27). We describe the theoretical foundation for the claims and relationships between them in the following sections.

DESIGN

The Design section (boxes A–F in **Figure 2**) delineates the six inputs, or components, of the assessment system necessary to achieve the desired outcomes. **Figure 1A** further describes the relationships of the content structures (i.e., the cognitive model, content standards, and assessment content). Design of the system uses a principled approach. Both the student population and intended uses of results were considered throughout the design process.

Cognitive Model

Guidance for best practice in assessment dating back to *Knowing What Students Know* (National Research Council, 2001) emphasizes the importance of developing assessments from a strong cognitive model. Cognitive models are research-based representations of skills measured by the assessment. Models take a variety of forms, ranging from concept maps (Wilson, 2009) to organized learning models. For instance, learning progressions show the sequential ordering of skill development in a domain (e.g., Alonzo and Steedle (2009) describe the connection between skills associated with learning concepts of force and motion). Learning map models show a similar progression, but rather than constrain skill development to a linear path, they acknowledge the network of ways students acquire skills (Kingston et al., 2016). While they are often developed from available research literature describing the order of skill acquisition, they are validated using expert review and empirical evaluation (e.g., Osborne et al., 2016).

Without a strong connection between cognition and assessment contents, the validity of inferences made from results can be compromised (National Research Council, 2001). Although large-scale accountability measures have not historically been based on strong models of cognition (Pellegrino et al., 2016), the use of learning progressions to inform formative assessment practices is widely acknowledged (e.g., Wilson, 2009; Alonzo, 2018). A strong research-based model for how skills develop supports instructional practice (Shepard, 2018), which is critical for subsequent claims in the theory of action. In the present theory of action, both the cognitive model (i.e., maps) and standards in the Instructionally Embedded assessment system inform one another. The maps expand on the standards, specifying skill development that is grade agnostic and connects the content area. They include foundational and precursor skills and skills that extend beyond grade-level expectations defined in the standards. Together, the cognitive model and academic standards inform the development of assessment content (claim D), support teacher selection of instruction and assessment content at a level appropriate for students (claim G), and provide the framework for fine-grained reporting (claim K) that ultimately supports teachers in using results to inform subsequent instruction (claim M).

Rigorous Academic Expectations

The Individuals with Disabilities Education Act (2004) requires that all students have full and equal access to education. Students with disabilities are to be assessed on state academic standards. Students with the most significant cognitive disabilities, or the approximately 1% of students for whom general education assessments are not appropriate, may be instructed and assessed on academic content that is reduced in depth, breadth, and complexity from the grade-level college- and career-ready standards. Yet alternate academic achievement standards are expected to be challenging, be aligned to the content standards for their enrolled grade level, reflect the highest possible achievement expectations for the population, and ensure that students are on track to pursue postsecondary opportunities, including education and competitive integrated employment (U.S. Department of Education, 2018).

Various programs and legislation support students with significant disabilities to pursue postsecondary opportunities, including the Transition and Postsecondary Programs for Students with Intellectual Disabilities, (2008), and the Workforce Innovation and Opportunity Act (2014). To pursue postsecondary opportunities, students must demonstrate a range of academic skills. Historically, however, students with significant cognitive disabilities were not held to high academic expectations. Kearns et al. (2011) found only around one-third to one-half of these students could read sight words and simple sentences and

did not understand larger text sections, approximately one-quarter could not read sight words, and around one-fifth did not have print awareness. Similarly, in mathematics, only around one-third to one-half could perform calculations, around one-third demonstrated counting and one-to-one correspondence, and nearly one-fifth did not use or know numbers (Kearns et al., 2011). These skills do not represent the full breadth of academic knowledge, skills, and understandings students need to pursue postsecondary opportunities (Karvonen et al., 2020).

The theory of action for DLM Instructionally Embedded assessments reflects the need for higher academic expectations and to combat historic reduced opportunity for these students to learn the full breadth of academic content. Rigorous academic expectations are reflected in the content standards. Subject-specific standards span multiple conceptual areas to provide students broad access to academic content. Flexible assessment blueprints specify how many standards teachers should select for instruction and assessment per area from among available standards. Rigorous academic content standards intersect with the cognitive model to inform development of assessments and achievement standards that reflect high expectations for all students (claim D), as well as the academic instruction students receive and their opportunity to learn rigorous content (claim H).

Accessible System

The importance of designing accessible assessments is widely documented. Federal law requires that all students have equal access to instructional materials, including educational assessments (IDEA, 2004). The *Standards for Educational and Psychological Testing* (the *Standards*) emphasize that access to test content is a fairness issue (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014). Literature describes accessibility as an interaction between examinee and assessment characteristics that supports or detracts from test takers providing responses reflecting their knowledge (Roelofs, 2019) and provides test development recommendations. To the extent that students are disadvantaged by lack of access, they may not fully demonstrate their knowledge, affecting subsequent theory of action claims.

Assessment literature often describes accommodations when referring to adjustments made to support students in showing their knowledge and skills (e.g., Lazarus et al., 2014). However, accommodations implemented outside of test development can have adverse impacts on students because they are external to the assessment content, do not consider students' individual differences, and sometimes are unnecessarily packaged together (Ketterlin-Geller, 2005). We instead emphasize *accessibility by design* to reflect the principle that accessibility needs for all students are considered in system design, rather than treating some students and their needs as exceptions. We draw from principles of universal design (e.g., Dolan, 2000), which has roots in architecture and design features like sidewalk curb cuts that provide independent access for a range of individuals. Assessment systems can be similarly designed to maximize student access to content. For instance, research shows that providing breaks and read-aloud benefits all students (Sireci et al., 2005; Pariseau et al., 2010). Rather than limiting breaks to students with a documented accommodation, system design can allow breaks for all students (e.g., to promote engagement). These types of universal-design decisions can reduce the impact of construct-irrelevant variance on assessment results (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014).

Advances in technology support the delivery of online assessment systems that are increasingly customizable and have potential to reduce access barriers. However, the transition to computer-based systems does not inherently make assessment content accessible (Ketterlin-Geller and Tindal, 2007). Students with significant cognitive disabilities who take alternate assessments are a widely heterogeneous population. They have a range of primary disabilities and communication needs, with many co-occurring (Erickson and Geist, 2016). Around 50% can use a computer with human support, while around 40% can use a computer independently (Burnes and Clark, 2021). Therefore, these students show what they know and can do in varied ways. To support subsequent claims in the theory of action, the system should be designed so that all students can respond as intended.

The assessment system supports delivery of accessible content. Accessible Portable Item Protocol (APIP) standards provide some guidance for making content accessible to students with disabilities (IMS Global Learning Consortium, 2014). They advise implementing Personal Needs and Preferences profiles to customize interfaces according to student accessibility needs. Customizable features include functionality like system read-aloud; display features including color contrast, color overlay, reverse contrast, and magnification; and the ability to interface with communication devices like switches and screen readers. Design of the assessment platform also considers features like font readability and size; screen real estate allocations for spacing stems, answer options, and media; simplified navigation and tool buttons; and the location of clickable fields. Considering and reviewing for these features during system design better enables students to show their knowledge (claim J) and educators to administer assessments as intended (claim I); students can respond as independently as they are able, with supports consistent with those used during instruction.

Instructionally Relevant Assessments

Guidance provided by the *Standards* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014) and best practices in the literature (e.g., Mislevy et al., 1999; Kane and Bejar, 2014) indicate assessment design should consider intended uses of results. The Instructionally Embedded system intends for teachers to use results to inform subsequent instruction. For results to be meaningful and useful, they should be fine grained and based on assessment content that is instructionally relevant and reflective of the cognitive model and its alignment to the standards upon which teachers provide instruction.

As with good instruction, instructionally relevant assessments should allow students to demonstrate their knowledge without the content being substantially too easy or too challenging. Like the ways in which individualized and differentiated instruction account for differences in skills across learners (Landrum and McDuffie, 2010), assessments can measure grade-level academic

expectations at different levels so students can show their learning. This requires that the cognitive model and academic standards be connected and that assessment content be developed to measure them. Federal peer review requirements (U.S. Department of Education, 2018), the *Standards* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014), and best practices in the literature (Martone and Sireci, 2009; Forte, 2017) all provide guidance for developing assessment content that is aligned to the construct being measured (i.e., the cognitive model and academic standards).

Using methods adapted from principles of evidence-centered design and universal design for learning (UDL), the assessment design and development process translates the underlying cognitive structure and content standards into assessment items using design templates for item writers (Bechard et al., 2019). Because of its fine-grained nature, the cognitive model may be more informative to item writing than the academic content standards (Bennett et al., 2016). Item writers use the cognitive model to write content of appropriate complexity at levels that precede, meet, or extend the academic standard so all students can show their knowledge relative to grade-level expectations. These are represented as the five levels for each standard (i.e., row) in **Figure 1**. Assessment targets are written so students can demonstrate what they know regardless of their sensory or physical characteristics or their communication modes.

For content to be instructionally relevant, it should resemble instructional activities. Assessments are written as short learning checks on a specific instructional target (i.e., one standard at one level measured by three to nine items). Instructionally relevant assessments are developed by individuals familiar with instructional practice. Expert item writers typically have both content and teaching expertise (e.g., Nichols and Fulkerson, 2010) and tend to rely on their experience when writing items (Kim et al., 2010). Item writers consider the ways in which educators elicit responses during instruction. For instance, instruction often includes teachers asking students questions during a shared book-reading activity (e.g., Walsh and Hodge, 2018). Instructionally relevant items may similarly embed questions within a text. Consistent with UDL principles, these and other engagement activities orient test takers to content, activate prior knowledge, and link items to a shared stimulus, so that questions are asked with context as they are during instruction. Item writers also consider misconceptions, drawing from the cognitive structure, academic standards, and their professional expertise to determine where student learning may be challenged, which they use to inform distractors and content assessed at different levels. In the broader population, designing assessments to look like instruction might raise concerns about narrowing the curriculum or teaching to the test. Given special educators' limited training on academic instruction, historically limited opportunity for students with significant cognitive disabilities to learn academics, limited resources for teaching academics, and the fact that special educators are responsible for a range of curricular priorities (of which academics is just one), the potential unintended consequences are less likely. Instructionally embedded assessment is a well-made resource to inform instruction. The system is designed so teachers retain autonomy in when and what they assess.

Item writers also consider the examinee population when developing universally designed assessments. Content is developed to be accessible for all students, by, for instance, limiting the complexity of sentences, reducing jargon, and selecting accessible graphics. These item-writing principles combine to produce instructionally relevant assessments that support subsequent claims related to students' showing their knowledge (claim J) and being assessed at the appropriate depth, breadth, and complexity (claim G) on taught content (claim H).

Educator Training for Assessment

The theory of action includes a claim that training prepares educators for assessment administration. Research shows educator training is effective in other capacities, such as improving teachers' assessment literacy (Lukin et al., 2004) and providing reliable scores on constructed-response rating tasks (Shin et al., 2019). Because Instructionally Embedded assessment is innovative and unique and because teachers have an increased role in content selection for instruction and assessment, teachers complete annual required training before assessment administration. While test-administrator training is not unusual in statewide assessments, the Instructionally Embedded training was intentionally designed to go beyond procedural and compliance-related topics and support teachers' conceptual understanding and use of the system consistent with intended purposes and uses and prepare them to make choices in the system using both their professional knowledge and knowledge of the student.

Following best practice, training incorporates multiple means of representation (CAST, 2020) and is available in self-directed and facilitated trainings, both of which are shown to be effective (Mertler, 2009; Vu et al., 2014). Training orients educators to the cognitive model, standards, assessment content, and availability of assessments at different levels. Training also covers system accessibility, making appropriate content selections according to students' instructional levels, and the flexible blueprint and its requirements. Other resources are provided to support teachers' administration and use of results as intended, including manuals and short videos. Quality educator training contributes to subsequent claims that teachers administer assessments with fidelity (claim I) and select content of adequate depth, breadth, and complexity (claim G).

Instructional Professional Development

The theory of action includes an input claim for building educators' instructional professional development. While instruction is assumed in statewide assessment systems, building instructional practice through professional development modules is worth explicitly addressing in the DLM theory of action because of historic lack of academic opportunity to learn for this student population (Ryndak et al., 2014; Karvonen et al., 2011; see claim H discussion in Delivery section). Special educators have diverse backgrounds and training (Leko and Brownell, 2009; Brownell et al., 2010), but

special education teacher-preparation programs may not emphasize subject-matter pedagogy as much as general education programs do (Brownell et al., 2005; Copeland et al., 2011). Special educators report being more prepared to address behavioral and communication challenges (Ruppar et al., 2016) and benefiting from modules designed to increase their subject-matter expertise (Lee et al., 2016). These findings indicate that while special educators may be well prepared to instruct students with a range of disabilities, they may lack relevant pedagogical knowledge. Other programs should evaluate inclusion of professional development as an input in their own theory of action, for example, when adopting substantially different content standards or implementing a statewide initiative that is expected to require teachers to learn and use new strategies.

Research shows professional development to be effective in increasing special educators' content and pedagogical content knowledge (Brownell et al., 2017). Educators can learn to teach rigorous academic content to students with significant cognitive disabilities (Bock and Erickson, 2015). When focused on content and pedagogy and when developed by experts, professional development can be effective in improving student learning (Guskey and Yoon, 2009). To this aim, professional development modules and other resources prepare educators to teach rigorous academic content measured by Instructionally Embedded assessments and to support students' communication needs so they can demonstrate their knowledge. Module content (available at dlmpd.com) provides overview information (e.g., the content standards, instructing students with significant cognitive disabilities) and is organized by the conceptual areas measured by the assessment (Kingston et al., 2016), with topics ranging from instruction on shared reading, getting started with narrative writing, and composing, decomposing, and comparing numbers. Organizing module content by conceptual area rather than singular standards promotes comprehensive instruction in the subject (Erickson and Koppenhaver, 2020), which is a departure from typical academic instruction for this population that focuses on teaching selected standards in isolation. Self-directed and facilitated formats support a range of educator learning preferences.

Professional development prepares teachers to provide instruction on rigorous academic standards and to provide students opportunity to learn the full depth and breadth of instructional content (claim H). It also supports subsequent claims in the theory of action of teachers having higher expectations for students (claim P) and using results to inform instruction (claim M).

DELIVERY

Delivery of Instructionally Embedded assessments prioritizes flexibility within constraints. Teacher choice is incorporated into implementation and administration processes. Teachers choose the content for instructional plans within blueprint constraints and how and when to administer assessments. These choices affect the extent to which students can demonstrate their knowledge. Prioritizing teacher choice embraces teachers as decision makers and professionals who know their students' needs and gives teachers agency to make decisions about student learning (Priestley et al., 2013; Kelly et al., 2018) while providing parameters to guide the decision-making process and support their success (e.g., Cook et al., 2008). Assessment delivery claims (boxes G–J in **Figure 2**) are influenced by system design claims (as inputs) and influence subsequent scoring and reporting claims.

Breadth, Depth, and Complexity

The Race to the Top call for through-course assessments (U.S. Department of Education, 2010) spurred development of new assessment models, including the DLM Instructionally Embedded model. Instructionally Embedded assessments are administered during 15-week fall (September–January) and spring (February–June) windows at teacher-determined times following instruction. Teachers determine the depth (e.g., number, content) and complexity (i.e., level) of instructional plans and assessment administration, within blueprint requirements. They also can choose to administer assessments beyond the breadth of standards required by the blueprint.

The breadth of administered assessment content has implications for theory of action scoring claims. Construct representation is a critical consideration for the design, administration, and scoring of educational assessments (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014). Not fully sampling the underlying construct means important content may be left off the assessment. In other words, students may not have the opportunity to demonstrate the full breadth of their knowledge. However, sampling standards too broadly may pose fairness issues if it poses challenges for educators to sufficiently instruct on all content (Popham, 2004). Instructionally Embedded assessment balances these two areas with a flexible blueprint. The blueprint defines the standards available for assessment in broad conceptual areas (i.e., related standards). For each area, the blueprint specifies the number of standards to assess to meet blueprint coverage (e.g., Choose three standards from area 1.1). Teachers have flexibility to select standards according to students' instructional goals, with the expectation that blueprint requirements are met in each administration window. There is some precedent for flexibility in construct representation for assessment systems (Bennett, 2018). The flexible assessment is also consistent with the individualized curricular priorities common in special education and addresses historical challenges with assessing this population, where standardized achievement assessments were too rigorous for students to demonstrate their knowledge and portfolio-based measures were not rigorous enough.

During instructionally embedded assessment, teachers determine the depth of instruction and assessment to provide within and across administration windows. The blueprint specifies requirements for the number of standards to assess from each conceptual area in each window but does not constrain choices to the same standards or levels across windows. Based on student goals and instructional needs, a teacher may determine that a student needs greater depth of instruction for a particular

standard (i.e., working with more-complex applications) as opposed to prioritizing greater breadth (i.e., adding more standards), as long as overall coverage requirements are met. This flexibility also allows teachers to determine when to administer assessments after instruction is provided, including how frequently to assess students. For instance, if additional learning occurs after an assessment is administered in the spring, the teacher can choose to assess the student again so they can demonstrate the full the depth of their knowledge, unlike traditional summative assessments that offer students only one response opportunity (e.g., Gong, 2021).

Finally, teachers determine the content complexity from among levels aligned to the grade-level expectation for each standard. While the system provides a recommended level based on information about students' academic and expressive communication skills, teachers can adjust the level. Research shows teachers can appropriately determine students' level of understanding (Heritage et al., 2009) and can make accurate judgments about student performance, particularly when their judgments are about the content standard or area being measured (Südkamp et al., 2012). When teachers select content of appropriate depth, breadth, and complexity, students can show what they know relative to rigorous, grade-level standards (claim J).

Provide Instruction

Once teachers make instructional plans, the theory of action assumes they provide quality instruction. Best practice indicates that instruction, assessment, and curriculum should be interrelated and inform one another (Roach et al., 2008; Martone and Sireci, 2009). Like other statewide assessment models, DLM assessments and this theory of action do not include curriculum as part of the system due to variation at the state and local levels on provided curriculum and its intended scope and sequence (Porter et al., 2009), as well as the high-level of individualization of instruction for students in this population. To account for this need for individualization, instruction is also not intended to be prescriptive in nature. Professional development modules prepare educators to provide comprehensive instruction that integrates content of multiple standards (Erickson and Koppenhaver, 2020), and instructional resources support educators in their practice. Local enacted curriculum and instruction should each be aligned to rigorous academic standards (and the cognitive model).

A critical assumption of educators providing quality instruction is that teachers provide students with adequate opportunity to learn the content measured by the assessment, and at the correct breadth, depth, and complexity. Opportunity to learn is emphasized in the *Standards* as critical to fairness and validity (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014) and is especially important for students with significant cognitive disabilities, who have historically had less access to rigorous academic content. For instance, even years after IDEA and NCLB legislation required that students with significant cognitive disabilities be taught academics, instruction often emphasized early-childhood content like shapes and listening tasks (Karvonen et al., 2011). Adequate opportunity to learn is linked with student learning, engagement, and achievement (Floden, 2002; Roach and Elliott, 2006; Karvonen and Huynh, 2007; Mo et al., 2013). For students to show their knowledge (claim J) and progress toward increased academic expectations (claim N), they should have adequate opportunity to learn rigorous content during instruction.

Administer With Fidelity

Fidelity examines the extent to which actual practice matches intended practice (Century et al., 2008). While intervention research often examines implementation fidelity, there is limited evidence of fidelity concerning assessment administration (e.g., Reed and Sturges, 2012). To promote fairness, through-course assessments used to make comparisons at state or district levels should have shared implementation practices (Zwick and Mislevy, 2011). Shared implementation practices for Instructionally Embedded assessments are included in administrator training (i.e., Design claim in theory of action) to support teachers in administering assessments as intended.

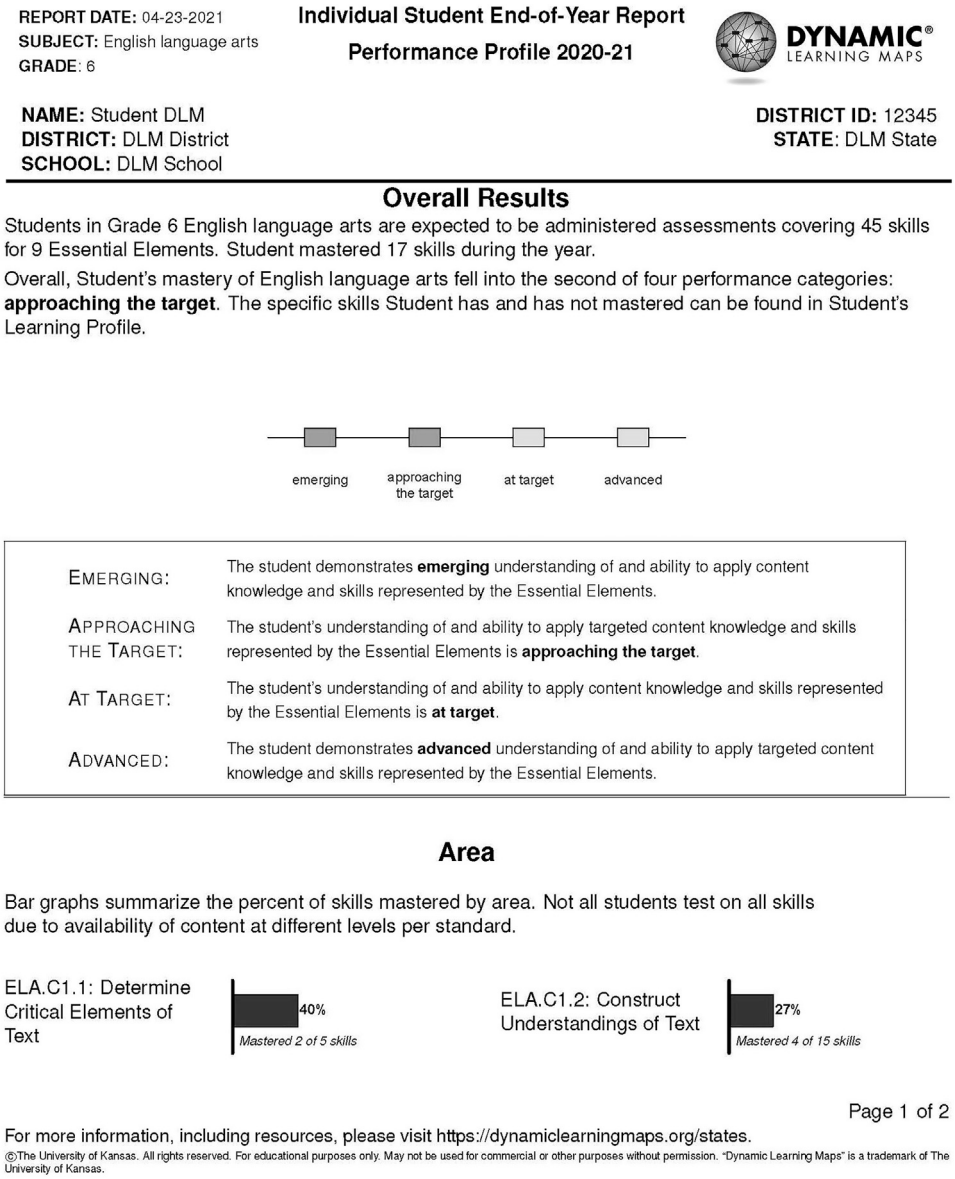
Flexibility in Instructionally Embedded administration draws from prior conceptions of intended variability (e.g., Gong and Marion, 2006). When administering assessments, teachers determine which accessibility supports to provide so that students receive supports they are familiar with from instruction. Teachers choose supports that enable students to respond as independently as they are able. In some cases, such as when students require mobility assistance or do not have access to communication devices, teachers may enter student answers or use partner-assisted scanning (Burnes and Clark, 2021; Sheldon and Erickson, 2020). This requires educators to accurately interpret student responses across modalities, such as gestures, eye gaze, or verbalization. When educators administer assessments with fidelity to intended practice, students can show their knowledge (claim J), and results more accurately reflect their knowledge claim K).

Show Knowledge and Skills

Students show what they know and can do when teachers make appropriate selections for the depth, breadth, and complexity of instruction and assessments; provide aligned instruction; and administer assessments with fidelity. These inputs together limit the impact of construct-irrelevant variance on students showing their knowledge and skills, the criticality of which is emphasized by the *Standards* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014). To the fullest extent possible, the delivery of assessments should limit external factors that could influence assessment results (such as difficulty responding to technology-enhanced items (Swinburne Romine et al., 2016) or limited engagement), while also allowing for flexibility in how students demonstrate their knowledge. To the extent that all educators administer assessments with fidelity (claim I) and receive instruction on the full breadth of content (claim H), students respond as intended (claim J) and mastery results accurately reflect students' knowledge and skills (claim K).

SCORING

Results are produced at two reporting levels (DLM Consortium, 2019). Diagnostic modeling uses all available



REPORT DATE: 04-23-2021
SUBJECT: English language arts
GRADE: 6

Individual Student End-of-Year Report
Performance Profile 2020-21

DYNAMIC® LEARNING MAPS

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

DISTRICT ID: 12345
STATE: DLM State

Overall Results

Students in Grade 6 English language arts are expected to be administered assessments covering 45 skills for 9 Essential Elements. Student mastered 17 skills during the year.

Overall, Student's mastery of English language arts fell into the second of four performance categories: **approaching the target**. The specific skills Student has and has not mastered can be found in Student's Learning Profile.

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.

Page 1 of 2

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

FIGURE 3 | Performance Profile portion of individual student score reports.

item responses to determine the probability of mastery for each assessed level per standard. The scoring model determines the highest level mastered for each content standard, and assumes all preceding levels were mastered as well. Fine-grained results summarize skill mastery (i.e., Learning Profile; **Figure 1**). Mastery is aggregated to describe overall achievement in the subject (i.e., Performance Profile; **Figure 3**). The delivery of reports at these two levels supports dual uses for instructional decision-making and inclusion in state accountability models, which supports intended outcomes in the theory of action. Scoring claims (boxes K–M in **Figure 2**) are influenced by assessment delivery claims (as inputs) and influence outcomes.

Fine-Grained Results

Research shows that to be useful, assessment data should be specific and at a fine-enough grain size to support instructional decision-making (Gill et al., 2014; Datnow and Hubbard, 2015). While there have been advances in subscore reporting for assessments with traditional scoring models (Haberman and Sinharay, 2010), these assessments require more items to support defensible delivery of subscores, and the level of reporting still may not be informative enough for teachers to plan subsequent instruction. In contrast, assessments designed for and scored by diagnostic models require fewer items to reliably determine mastery (Templin and Bradshaw, 2013), reducing the administration burden and time spent testing. Further, confidence in mastery determinations is inherent in scoring; because scoring uses the probability that students mastered the

assessed skills, values farther from 0.5 indicate greater certainty in the mastery classification. Setting a mastery threshold farther from 0.5 ensures greater confidence that the reported mastery results are accurate reflections of student knowledge.

Assessments scored with diagnostic models, such as attribute hierarchies and Bayesian inference networks (Leighton and Gierl, 2007; Rupp et al., 2010), are based on models of cognition, which supports fine-grained reporting on the mastered skills (Feldberg and Bradshaw, 2019; Karvonen et al., 2019). For the Instructionally Embedded system, reports summarize the highest level mastered for each standard assessed, connecting the cognitive model and academic standards directly to reporting. Mastered skills are shaded on score reports, emphasizing student strengths rather than deficits. This is particularly important when communicating results for students with significant cognitive disabilities, who have traditionally received reports indicating areas of weakness (Nitsch, 2013). During instructionally embedded assessment windows, mastery results (i.e., highest level mastered by standard to date) are summarized in on-demand progress reports (structured the same as the Learning Profile in **Figure 1**) to provide timely information to inform subsequent instruction (claim M). End of year reports also include the highest level students mastered per content standard (i.e., by the end of the year), which teachers can use to inform instructional planning in the subsequent year (Clark et al., 2018). This includes using information about mastery for related standards in the same conceptual area (first column in **Figure 1** report table).

Summative Performance

Researchers and practitioners have suggested that assessments can and should support both instructional and accountability purposes (e.g., Gordon, 2020). While instructional decision-making benefits from fine-grained results (Clark et al., 2018), accountability models rely on achievement levels for reporting performance. Researchers have developed methods for aggregating fine-grained skill-mastery information into descriptors of achievement in the subject (Clark et al., 2017; Skaggs et al., 2020). These methods modify existing standard-setting practices to account for reporting that is based on mastery profiles rather than a raw or scale score. This process provides consistent achievement expectations across all students, even with the added flexibility of instructionally embedded administration and teacher choice. Aggregate results support district and state users in describing achievement across students and comparability of score meaning (i.e., describing student achievement demonstrated throughout the year; Winter, 2010), without requiring teachers to administer additional assessments beyond those used to inform instruction. For the Instructionally Embedded system, we held early conversations with state partners about what summative results would describe and differentiated that results indicate what students know and can do "by" the end of the year rather than "at" the end of the year. States collectively arrived at four performance levels used to describe summative achievement. These results are summarized in the Performance Profile portion of student score reports (see Overall Results section of **Figure 3**). Aggregated class, school, district, and state reports also provide performance summaries across students to support accountability uses, resource allocation, and other state and district level needs (claim Q). Technical evidence (e.g., reliability at the performance level and in the subject overall; DLM Consortium, 2019) is collected to evaluate these score uses.

Used for Instructional Decision-Making

Assessment information can and should be used to inform instruction (Gordon, 2020). Research shows that, when supported, educators can use data to inform instructional decision-making (Means et al., 2011; Datnow et al., 2012). Research also indicates that when teachers use data, student learning improves (McMaster et al., 2020). However, data use can also be affected by building and district data use climate and other local factors (Levin and Datnow, 2012; Abrams et al., 2016).

When teachers use data to inform instruction, they draw on knowledge of content, curriculum, pedagogy, information about students they instruct, educational context, and purposes for instruction (Mandinach and Gummer, 2013; Poortman and Schildkamp, 2016). While teachers can often judge student achievement, they more often struggle with determining what instruction to provide next (Heritage et al., 2009), and special education teachers may lack the academic content knowledge to plan next instructional steps. When educators can situate current student performance within the broader learning trajectory, they are better prepared to make subsequent instructional decisions (Heritage, 2008). The cognitive model for Instructionally Embedded assessments as displayed in the Learning Profiles supports teachers in planning subsequent instruction, rather than teachers having to rely on their own content knowledge for what to instruct next (i.e., teachers can use the information in the Learning Profile (**Figure 1**) to know the order of skill acquisition and plan subsequent instruction as students work toward the grade-level target). Research shows that educators understand the contents of mastery profiles (Feldberg and Bradshaw, 2019) and can use information about current skill mastery to determine next steps for instruction (Karvonen et al., 2017), even when they do not receive local training or resources (Clark et al., 2018).

OUTCOMES

Outcomes in the theory of action (boxes N–Q in **Figure 2**) are intended to be long-term and over time. They are realized when previous claims in the model are fulfilled. The theoretical model described here, built on relevant literature in the field, provides support for the relationships depicted in the model. The Instructionally Embedded theory of action culminates in the intended outcomes of students progressing toward higher expectations (e.g., addressing historic challenges of reduced opportunity to learn and narrow scope of enacted curriculum, current AA-AAS reflecting increased expectations and alignment to postsecondary opportunities; Karvonen et al., 2020), teachers using data to make instructional decisions, and educators having increased expectations for student achievement (i.e., shifting perceptions of what students with significant cognitive disabilities are capable of achieving academically). There is a feedback loop between students making progress and teachers having high expectations, whereby teachers observe students achieving more over time and realize more

is possible (e.g., McGrew and Evans, 2004; Timberlake, 2014). Students, particularly those with significant cognitive disabilities, can progress toward higher expectations when they are instructed and assessed on rigorous academic standards, can access the assessment content, have adequate opportunity to learn, and can fully demonstrate their knowledge and skills.

Educators receive fine-grained mastery information that can inform instruction. Decision making may include using data to prioritize topics for instructional time, to determine individual strengths and weaknesses, and to provide individualized instruction (Hamilton et al., 2009). Educators' ability to use data to inform instruction is best facilitated through frequent opportunities to use data collected from their own students (Leko et al., 2015). It is recommended that data are used as part of a continuous cycle of instructional planning (Hamilton et al., 2009). The instructionally embedded nature of the assessment and the practice of providing fine-grained mastery information as assessments are completed support teachers in using assessment data to inform instruction, without requiring them to create or administer additional assessments beyond those required for state accountability purposes.

Research documents the relationship between student achievement and teacher expectations (e.g., Archambault et al., 2012; Rubie-Davies, 2007). Studies demonstrate it is possible to both raise teacher expectations and student achievement (de Boer et al., 2018). As students show their knowledge and progress toward increased expectations, and teachers engage in ongoing cycles of instructional planning using data that show students are achieving increased expectations, teachers' expectations for the academic content students with significant cognitive disabilities can achieve will increase. While the field of study on increased academic expectations for students with significant cognitive disabilities is slowly improving (e.g., Geist et al., 2014; Wakeman et al., in press), we believe the ongoing use of an assessment system based on rigorous academic standards and a cognitive model for skill development further supports teachers in having high expectations for their students.

DISCUSSION

As Gordon (2020) emphasized, it is not enough to merely report assessment results; assessments should be a catalyst for change. Although several innovative assessment programs are currently being piloted under the ESSA Innovative Assessment Demonstration Authority (U.S. Department of Education, 2020), they are still in early phases of development. This paper adds to the literature on flexible and innovative assessment systems by describing a conceptual model for realizing change with an Instructionally Embedded system that replaces traditional summative assessments and supports teachers in using results to inform instruction and increase their expectations for what students with the most significant cognitive disabilities can achieve.

When designing the assessment system and constructing the theory of action, consortium priorities and our collective beliefs about academic learning for the population drove many of the decisions. For instance, the claims and conceptual areas used to organize groups of standards and map segments emphasize conceptual understandings. In English language arts, the first claim is that students comprehend text in increasingly complex ways. This claim contrasts with historic priorities for the population, often limited to sight word recognition or oral reading fluency rather than comprehension. The DLM theory of action and assessment system reflect the extremely heterogeneous population and the individualized nature of their instruction and assessment. Assessment programs for other populations might require developers and other stakeholders to articulate characteristics of the examinee population and philosophies that influence design of the assessment system. For example, how do assumptions about teaching and learning for more homogeneous groups of students inform aggregated score reports that produce instructionally useful information for teachers who differentiate within a classroom?

We recognize that with innovation also come disruptions to status quo and identification of areas for improvement. Most states that collaborated to design the Instructionally Embedded model came from portfolio-based alternate assessment systems where teachers already had some degree of choice in what and how to assess. The Instructionally Embedded system still represented a shift in assessment practice. The consortium developed trainings and supports to help teachers make the transition. But the states also decided to move toward the Instructionally Embedded model in stages. Their first operational system used an instructionally embedded window open for more than half the year, followed by a computer-adaptive model that reassessed certain standards in late spring. This system was designed to ensure students met blueprint coverage and that their assessment results were based on evidence collected throughout the year. Only after evaluating implementation for several years and enhancing the online assessment management system to help teachers meet blueprint coverage did the states decide to change to a fully Instructionally Embedded model.

The Instructionally Embedded model promotes coherence between instruction and assessment. While we recognize that curriculum alignment is also essential, the consortium was not in a position to formally include curriculum in the model. This decision promoted independence across states, left local curricular decision-making intact within states, and supported the long history of individualized curricular priorities and methods for this population. Leaving curriculum out of the model has implications for claims in the theory of action and validity evaluation. This decision also means districts will need to supplement aggregated DLM assessment results with locally-available data on curriculum implementation when making decisions about programs and resource allocation.

We also recognize conceptual challenges arise when implementing innovative assessment systems like the Instructionally Embedded system. Challenges include departures from traditional conceptions of standardization and considerations for scoring and reporting student growth. Within these areas, we describe future considerations as Instructionally Embedded and other innovative assessment systems are increasingly adopted to meet stakeholder needs for assessments that inform instruction while also serving accountability requirements.

Standardization and Comparability

Despite recent attention on classroom assessment, the educational measurement field has largely prioritized measurement that is summative, classificatory, and used for accountability purposes (Gordon, 2020). Historically the field has emphasized standardization to produce score comparability (Lee et al., 2003; Sireci, 2020). However, as technology capabilities advance to support customization and personalized learning and stakeholder needs evolve, we recognize that "same" is not always better. Sometimes it is desirable to reduce standardization in the service of administering assessments that are useful to teaching and learning (Gallagher, 2003; Gordon, 2020; Sireci, 2020) and promoting equity across students (Winter, 2010). This includes a recognition that not all students need the same things when learning, and assessments do not have to adhere to a one-size-fits-all approach to be useful.

The Instructionally Embedded assessment adopts the practice of flexibility within constraints. That is, we intentionally build flexibility into assessment design and administration so students can more fully demonstrate their knowledge. This principle appears throughout the theory of action and affects how outcomes are realized. Because variability decreases standardization, it is important to consider its impact on practice and the extent that validity evidence supports intended uses. The *Standards* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014) indicate that when variability is permitted, the premise of standardization can be maintained by detailing the conditions in which choices can be made and the criteria for evaluation. Developers can put constraints in place to guide flexibility within intended practice and promote fairness and equity.

Political context and the extent to which state education agencies encourage teachers to use assessment results to inform instruction may affect the state's willingness to adopt a flexible assessment model or how much flexibility to offer. As demonstrated here, programs may want to consider the degree of flexibility to allow when defining test specifications, including the breadth of content coverage, the extent to which teachers make decisions about depth of instruction and assessment, and whether varying levels of complexity are offered, as well as the degree of flexibility to allow in the timing of assessment relative to instruction, or other areas.

While the field is increasingly receptive to assessments with less standardization (e.g., Sireci, 2020), developers should also consider the extent that flexibility in assessment design can introduce potential unintended consequences. For instance, developers could consider that when teachers determine assessment complexity level, there may be a risk that students are disadvantaged or subjected to teachers' low expectations. Developers might decide to constrain choices to prevent low expectations. Or, developers might incorporate conditions that facilitate high expectations, such as designing training to support educators in having high expectations, providing materials on appropriate administration to support students showing what they know at the level(s) assigned, prioritizing reporting that focuses on mastery rather than deficits that is useful for instruction, and messaging that reduces the emphasis on assessment for purely accountability purposes. When collecting validity evidence, developers should similarly be careful to not only collect confirmatory evidence but to also explore disconfirmatory evidence to rule out alternate explanations (such as teacher low expectations) for student performance.

Scoring and Growth

Large-scale assessments typically use item response theory to provide a scale score summarizing performance. In an Instructionally Embedded model, assessments are administered as instruction occurs, and there is no need to provide a scale score on demand. We instead provide fine-grained diagnostic mastery information as it is collected. While there is robust literature on diagnostic models in research applications, their operational use has been limited to date (Sessoms and Henson, 2018). Adopting innovative scoring models introduces the need for operational research into areas that have been less well-documented in the literature. For instance, methods for evaluating model fit for diagnostic models are still being defined (e.g., Chen et al., 2013; Hu et al., 2016), which has implications for demonstrating technical adequacy for these systems.

While there is also increased emphasis in the field on the use of results to inform teaching and learning taking precedence over accountability purposes, stakeholders may still desire (or require) information about student growth or progress over time. Growth metrics are often used for accountability and therefore rely on data at an aggregated level. Common approaches include transitional matrices and student growth percentiles (Domaleski and Hall, 2016); however, researchers note challenges for reporting growth in these ways, particularly for Instructionally Embedded assessments based on diagnostic scoring (Nehler et al., 2019). Diagnostic model-based growth measures have shown some promise for reporting within-year growth focused on student-specific change over time (e.g., Madison, 2019), but additional research and operational application are needed. However, the practice of assessing students as instruction occurs and making mastery results available throughout the year provides teachers with evidence of student progress over time as additional skills are mastered within and across standards, which shifts the focus of reporting student growth away from aggregate, year-to-year measures and allows for documenting academic progress within a year.

An additional concern often raised in high-stakes testing environments is that of test-score integrity and the potential for users to "game the system." As the paradigm shifts away from assessment for accountability toward assessment to inform learning, the impetus for gaming the system is reduced. The feasibility of gaming the system can also be addressed through the test design, implementation procedures, and scoring rules. Expanding data-forensics literature provides methods for evaluating (and ruling out) the presence of aberrant response patterns or anomalies (e.g., Kingston and Clark, 2014; Juhaňák et al., 2019), which can support the defensibility of results for a flexible assessment system.

Adoption of Flexible System

Perhaps one of the biggest hurdles to implementing flexible and innovative assessment systems, like the Instructionally Embedded

system, occurs before any of the design work. Stakeholders must be open to and ultimately buy into the concept of an innovative assessment model and agree to the intended changes the program seeks to bring about. This consensus spans a range of stakeholders, from local educators to state education agency staff and politicians. It requires a willingness to pursue areas of new research and a recognition that not all aspects will have already been completely studied and well-documented.

Once a program makes the decision to move forward with an innovative system that intends to elicit change in stakeholder behaviors, a theory of action should be developed, perhaps drawing from the concepts presented here. During this work, developers should consider intended uses of results and the population of students completing assessments. While some of the theoretical rationale for the present argument came from literature on students with significant cognitive disabilities, the general sentiments of having rigorous content expectations, developing an accessible system, and providing adequate opportunity to learn apply to all student populations.

CONCLUSION

Instructionally embedded assessments and other through-year assessment models offer opportunity for assessment results to inform instructional practice and also be used for state accountability purposes. These assessments demonstrate a departure from more traditional summative assessments administered at the end of an academic year. As more states consider adopting innovative and flexible assessment systems, they should consider how to balance aspects of comparability, standardization, and teacher autonomy in making decisions within flexible constraints. Future research should also continue to examine these issues related to comparability across students when assessments allow for flexible blueprints and other teacher choices and also the extent to which flexible assessments produce comparable results pertaining to student achievement as their traditional counterparts (e.g., in states offering local education agencies a choice between assessments).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

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.

REFERENCES

Abrams, L., Varier, D., and Jackson, L. (2017). Unpacking Instructional Alignment: The Influence of Teachers' Use of Assessment Data on Instruction. *PiE* 34 (4), 15–28. doi:10.18820/2519593X/pie.v34i4.2

Ahn, J., Campos, F., Hays, M., and DiGiacomo, D. (2019). Designing in Context: Reaching beyond Usability in Learning Analytics Dashboard Design. *J. Learn. Analytics* 6 (2), 70–85. doi:10.18608/jla.2019.62.5

Allor, J. H., Mathes, P. G., Roberts, J. K., Jones, F. G., and Champlin, T. M. (2010). Teaching Students with Moderate Intellectual Disabilities to Read: An Experimental Examination of a Comprehensive reading Intervention. *Education Train. Autism Dev. Disabilities* 45 (1), 3–22.

Alonzo, A. C. (2018). Exploring the Learning Progression-Formative Assessment Hypothesis. *Appl. Meas. Edu.* 31 (2), 101–103. doi:10.1080/08957347.2017.1408625

Alonzo, A. C., and Steedle, J. T. (2009). Developing and Assessing a Force and Motion Learning Progression. *Sci. Ed.* 93 (3), 389–421. doi:10.1002/sce.20303

Altman, J., Thurlow, M., and Vang, M. (2010). *Annual Performance Report: 2007-2008 State Assessment Data*. Minneapolis, Minnesota: University of Minnesota, National Center on Educational Outcomes.

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

Archambault, I., Janosz, M., and Chouinard, R. (2012). Teacher Beliefs as Predictors of Adolescents' Cognitive Engagement and Achievement in Mathematics. *J. Educ. Res.* 105 (5), 319–328. doi:10.1080/00220671.2011.629694

Bechard, S., Clark, A., Swinburne Romine, R., Karvonen, M., Kingston, N., and Erickson, K. (2019). Use of Evidence-Centered Design to Develop Learning Maps-Based Assessments. *Int. J. Test.* 19 (2), 188–205. doi:10.1080/15305058.2018.1543310

Bennett, R. E. (2010). Cognitively Based Assessment of, for, and as Learning (CBAL): A Preliminary Theory of Action for Summative and Formative Assessment. *Meas. Interdiscip. Res. Perspect.* 8 (2-3), 70–91. doi:10.1080/15366367.2010.508686

Bennett, R. E., Deane, P., and W. van Rijn, P. (2016). From Cognitive-Domain Theory to Assessment Practice. *Educ. Psychol.* 51 (1), 82–107. doi:10.1080/00461520.2016.1141683

Bennett, R. E. (2018). Educational Assessment: What to Watch in a Rapidly Changing World. *Educ. Meas. Issues Pract.* 37 (4), 7–15. doi:10.1111/emip.12231

Bock, A. K., and Erickson, K. A. (2015). The Influence of Teacher Epistemology and Practice on Student Engagement in Literacy Learning. *Res. Pract. Persons Severe Disabilities* 40 (2), 138–153. doi:10.1177/1540796915591987

B. P. Veldkamp and C. Sluijter (Editors) (2019). *Theoretical and Practical Advances in Computer-Based Educational Measurement* (Berlin, Germany: Springer Open).

Bradshaw, L. (2017). "Diagnostic Classification Models," in *The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications*. Editors A. A. Rupp and J. P. Leighton (Hoboken, NJ: John Wiley & Sons), 297–327.

Brownell, M., Kiely, M. T., Haager, D., Boardman, A., Corbett, N., Algina, J., et al. (2017). Literacy Learning Cohorts. *Exceptional Child.* 83 (2), 143–164. doi:10.1177/0014402916671517

Brownell, M. T., Ross, D. D., Colón, E. P., and McCallum, C. L. (2005). Critical Features of Special Education Teacher Preparation. *J. Spec. Educ.* 38 (4), 242–252. doi:10.1177/00224669050380040601

Brownell, M. T., Sindelar, P. T., Kiely, M. T., and Danielson, L. C. (2010). Special Education Teacher Quality and Preparation: Exposing Foundations, Constructing a New Model. *Exceptional Child.* 76 (3), 357–377. doi:10.1177/001440291007600307

Burnes, J. J., and Clark, A. (2021). *Characteristics Of Students Who Take Dynamic Learning Maps® Alternate Assessments (Technical Report No. 20-01)*. ATLAS: University of Kansas.

Carter, E. W., Lane, K. L., Cooney, M., Weir, K., Moss, C. K., and Machalicek, W. (2013). Parent Assessments of Self-Determination Importance and Performance for Students with Autism or Intellectual Disability. *Am. J. Intellect. Dev. Disabil.* 118 (1), 16–31. doi:10.1352/1944-7558-118.1.16

CAST (2020). *Principle: Provide Multiple Means of Representation*. Available at: https://udlguidelines.cast.org/representation.

Cawelti, G. (2006). The Side Effects of NCLB. *Educ. Leadersh.* 64 (3), 64–68.

Century, J., Freeman, C., and Rudnick, M. (2008). *A Framework for Measuring and Accumulating Knowledge about Fidelity of Implementation (FOI) of Science Instructional Materials [Paper Presentation]*. Baltimore, MD, United States: Annual meeting of the National Association for Research in Science Teaching.

Chalhoub-Deville, M. (2016). Validity Theory: Reform Policies, Accountability Testing, and Consequences. *Lang. Test.* 33 (4), 453–472. doi:10.1177/0265532215593312

Chen, J., de la Torre, J., and Zhang, Z. (2013). Relative and Absolute Fit Evaluation in Cognitive Diagnosis Modeling. *J. Educ. Meas.* 50 (2), 123–140. doi:10.1111/j.1745-3984.2012.00185.x

Clark, A. K., and Karvonen, M. (2020). Constructing and Evaluating a Validation Argument for a Next-Generation Alternate Assessment. *Educ. Assess.* 25 (1), 47–64. doi:10.1080/10627197.2019.1702463

Clark, A. K., Karvonen, M., Swinburne Romine, R., and Kingston, N. M. (2018). *Teacher Use of Score Reports for Instructional Decision-Making: Preliminary Findings [Presentation]*. New York: Annual meeting of the National Council on Measurement in Education.

Clark, A. K., Nash, B., Karvonen, M., and Kingston, N. (2017). Condensed Mastery Profile Method for Setting Standards for Diagnostic Assessment Systems. *Educ. Meas. Issues Pract.* 36 (4), 5–15. doi:10.1111/emip.12162

Cook, B. G., Tankersley, M., Cook, L., and Landrum, T. J. (2008). Evidence-based Practices in Special Education: Some Practical Considerations. *Intervention Sch. Clinic* 44, 69–75. doi:10.1177/1053451208321452

Copeland, S. R., Keefe, E. B., Calhoon, A. J., Tanner, W., and Park, S. (2011). Preparing Teachers to Provide Literacy Instruction to All Students: Faculty Experiences and Perceptions. *Res. Pract. Persons Severe Disabilities* 36 (3–4), 126–141. doi:10.2511/027494811800824499

Council of Chief State School Officers (2018). *An Integrated Approach to Defining a System-Level Theory of Action for Formative Assessment*. Washington, DC: CCSSO.

Council of Chief State School Officers (2008). Attributes of Effective Formative Assessment. Available at: https://ccsso.org/sites/default/files/2017-12/Attributes_of_Effective_2008.pdf.

Datnow, A., and Hubbard, L. (2015). Teachers' Use of Assessment Data to Inform Instruction: Lessons from the Past and Prospects for the Future. *Teach. Coll. Rec.* 117, 1–26.

Datnow, A., Park, V., and Kennedy-Lewis, B. (2012). High School Teachers' Use of Data to Inform Instruction. *J. Edu. Students Placed Risk (Jespar)* 17 (4), 247–265. doi:10.1080/10824669.2012.718944

de Boer, H., Timmermans, A. C., and van der Werf, M. P. C. (2018). The Effects of Teacher Expectation Interventions on Teachers' Expectations and Student Achievement: Narrative Review and Meta-Analysis. *Educ. Res. Eval.* 24 (3–5), 180–200. doi:10.1080/13803611.2018.1550834

Dolan, B. (2000). Universal Design for Learning [Guest Column]. *J. Spec. Edu. Tech.* 15 (4), 47–51.

Domaleski, C., and Hall, E. (2016). *Guidance for Estimating and Evaluating Academic Growth*. Minneapolis, MN: National Center and State Collaborative.

Erickson, K. A., and Geist, L. A. (2016). The Profiles of Students with Significant Cognitive Disabilities and Complex Communication Needs. *Augment Altern. Commun.* 32 (3), 187–197. doi:10.1080/07434618.2016.1213312

Erickson, K. A., and Koppenhaver, D. A. (2020). *Comprehensive Literacy for All: Teaching Students with Significant Disabilities to Read and Write*. Baltimore, MD: Paul H. Brookes Publishing Co., Inc.

Every Student Succeeds Act (2015). *P.L. 114-95, 20 U.S.C. §*, 6301.

Farrell, C. C., and Marsh, J. A. (2016). Contributing Conditions: A Qualitative Comparative Analysis of Teachers' Instructional Responses to Data. *Teach. Teach. Edu.* 60, 398–412. doi:10.1016/j.tate.2016.07.010

Feldberg, Z., and Bradshaw, L. (2019). *Reporting Results from Diagnostic Classification Models for Teachers. [Paper presentation]*. Toronto: National Council on Measurement in Education.

Ferrara, S., Lai, E., Reilly, A., and Nichols, P. D. (2017). "Principled Approaches to Assessment Design, Development, and Implementation," in *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications*. Editors A. A. Rupp and J. P. Leighton (Hoboken, NJ: John Wiley & Sons), 41–74.

Floden, R. E. (2002). "The Measurement of Opportunity to Learn," in *Methodological Advances in Cross-National Surveys of Educational Achievement*. Editors A. C. Porter and A. Gamoran (Washington, D.C.: The National Academies Press), 231–266.

Forte, E. (2017). *Evaluating Alignment in Large-Scale Standards-Based Assessment Systems*. Washington, DC: Council of Chief State School Officers.

Gallagher, C. J. (2003). Reconciling a Tradition of Testing with a New Learning Paradigm. *Educ. Psychol. Rev.* 15 (1), 83–99. doi:10.1023/a:1021323509290

Geist, L., Hatch, P., and Erickson, K. (2014). Promoting Academic Achievement for Early Communicators of All Ages. *Perspect. Augment Altern. Commun.* 23 (4), 173–181. doi:10.1044/aac23.4.173

George, A., and Sanders, M. (2017). Evaluating the Potential of Teacher-Designed Technology-Based Tasks for Meaningful Learning: Identifying Needs for Professional Development. *Educ. Inf. Technol.* 22, 2871–2895. doi:10.1007/s10639-017-9609-y

Gill, B., Coffee Borden, B., and Hallgren, K. (2014). Mathematica Policy Research Report: A Conceptual Framework for Data-Driven Decision Making. Available at: https://www.disabilitypolicyresearch.org/~/media/publications/pdfs/education/.

Gholson, M. L., and Guzman-Orth, D. (2019). Developing an Alternate English Language Proficiency Assessment System: A Theory of Action. *ETS Research Report Series* 2019, 1–19. doi:10.1002/ets2.12262

Gong, B., and Marion, S. (2006). *Dealing With Flexibility in Assessments for Students with Significant Cognitive Disabilities (Synthesis Report 60)*. Minneapolis, Minnesota: University of Minnesota, National Center on Educational Outcomes.

Gong, B. (2010). *Using Balanced Assessment Systems to Improve Student Learning and School Capacity: An Introduction*. Dover, NH: Council of Chief State School Officers.

Gong, B. (2021). Why Has it Been So Difficult to Develop a Viable through Year Assessment? Available at: https://www.nciea.org/blog/state-testing/why-has-it-been-so-difficult-develop-viable-through-year-assessment.

Gordon, E. W. (2020). Toward Assessment in the Service of Learning. *Educ. Meas. Issues Pract.* 39 (3), 72–78. doi:10.1111/emip.12370

Graziano, K. J., Herring, M. C., Carpenter, J. P., Smaldino, S., and Finsness, E. S. (2017). A TPACK Diagnostic Tool for Teacher Education Leaders. *TechTrends* 61, 372–379. doi:10.1007/s11528-017-0171-7

Guskey, T. R., and Yoon, K. S. (2009). What Works in Professional Development? *Phi Delta Kappan* 90, 495–500. doi:10.1177/003172170909000709

Haberman, S. J., and Sinharay, S. (2010). Reporting of Subscores Using Multidimensional Item Response Theory. *Psychometrika* 75, 209–227. doi:10.1007/s11336-010-9158-4

Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., Wayman, J., et al. (2009). *Using Student Achievement Data to Support Instructional Decision Making (IES Practice Guide; NCEE 2009-4067)*. Washington, D.C.: U.S. Department of Education.

Hedger, J. (2020). States experiment with Assessment through Innovative Pilots. *State. Edu. Stand.* 20 (3), 40–42.

Heritage, M., Kim, J., Vendlinski, T., and Herman, J. (2009). From Evidence to Action: A Seamless Process in Formative Assessment? *Educ. Meas. Issues Pract.* 28, 24–31. doi:10.1111/j.1745-3992.2009.00151.x

Heritage, M. (2008). *Learning Progressions: Supporting Instruction and Formative Assessment*. Washington, DC: Chief Council of State School Officers.

Hu, J., Miller, M. D., Huggins-Manley, A. C., and Chen, Y.-H. (2016). Evaluation of Model Fit in Cognitive Diagnosis Models. *Int. J. Test.* 16 (2), 119–141. doi:10.1080/15305058.2015.1133627

IMS Global Learning Consortium (2014). Accessible Portable Item Protocol (Version 1.0). Available at: http://www.imsglobal.org/apip/index.html.

Individuals with Disabilities Education Act (2004). *20 U.S.C. § 1400*.

Jaakkola, E. (2020). Designing Conceptual Articles: Four Approaches. *AMS Rev.* 10, 18–26. doi:10.1007/s13162-020-00161-0

Jacobson, L. (2019). Is This the End of End-Of-Year Testing? Education Dive. Available at: https://www.k12dive.com/news/is-this-the-end-of-end-of-year-testing/565850/.

Juhaňák, L., Zounek, J., and Rohlíková, L. (2019). Using Process Mining to Analyze Students' Quiz-Taking Behavior Patterns in a Learning Management System. *Comput. Hum. Behav.* 92, 496–506. doi:10.1016/j.chb.2017.12.015

Kane, M. T., and Bejar, I. I. (2014). Cognitive Frameworks for Assessment, Teaching, and Learning: A Validity Perspective. *Psicología Educativa* 20 (2), 117–123. doi:10.1016/j.pse.2014.11.006

Karvonen, M., Burnes, J. J., Clark, A. K., and Kavitsky, L. (2020). *Aligned Academic Achievement Standards to Support Pursuit of Postsecondary Opportunities: Instructionally Embedded Model (Technical Report No. 20-02)*. ATLAS: University of Kansas.

Karvonen, M., Clark, A. K., Swinburne Romine, R., and Kingston, N. (2019). *Development and Evaluation of Diagnostic Score Reports for an Alternate Assessment System. [Paper presentation]*. Toronto, Canada: American Educational Research Association.

Karvonen, M., and Huynh, H. (2007). Relationship between IEP Characteristics and Test Scores on an Alternate Assessment for Students with Significant Cognitive Disabilities. *Appl. Meas. Edu.* 20 (3), 273–300. doi:10.1080/08957340701431328

Karvonen, M., Swinburne Romine, R., Clark, A. K., Brussow, J., and Kingston, N. (2017). *Promoting Accurate Score Report Interpretation and Use for Instructional Planning [Paper Presentation]*. San Antonio, TX: National Council on Measurement in Education.

Karvonen, M., Wakeman, S. Y., Browder, D. M., Rogers, M. A., and Flowers, C. (2011). *Academic Curriculum for Students with Significant Cognitive Disabilities: Special Education Teacher Perspectives a Decade after IDEA 1997 (ED521407)*. ERIC. Available at: https://files.eric.ed.gov/fulltext/ED521407.pdf.

Kearns, J. F., Towles-Reeves, E., Kleinert, H. L., Kleinert, J. O. R., and Thomas, M. K.-K. (2011). Characteristics of and Implications for Students Participating in Alternate Assessments Based on Alternate Academic Achievement Standards. *J. Spec. Educ.* 45 (1), 3–14. doi:10.1177/0022466909344223

Kelly, K., Merry, J., and Gonzalez, M. (2018). Trust, Collaboration and Well-Being: Lessons Learned from Finland (EJ1186140). *ERIC. SRATE J.* 27 (2), 34–39.

Ketterlin-Geller, L. (2005). Knowing what All Students Know: Procedures for Developing Universal Design for Assessment. *J. Technol. Learn. Assess.* 4 (2), 4–22.

Ketterlin-Geller, L. R., and Tindal, G. (2007). Embedded Technology: Current and Future Practices for Increasing Accessibility for All Students. *J. Spec. Educ. Technol.* 22, 1–15. doi:10.1177/016264340702200401

Kim, J., Chi, Y., Huensch, A., Jun, H., Li, H., and Roullion, V. (2010). A Case Study on an Item Writing Process: Use of Test Specifications, Nature of Group Dynamics, and Individual Item Writers' Characteristics. *Lang. Assess. Q.* 7 (2), 160–174. doi:10.1080/15434300903473989

Kingston, N. M., Karvonen, M., Bechard, S., and Erickson, K. A. (2016). The Philosophical Underpinnings and Key Features of the Dynamic Learning Maps Alternate Assessment. *Teach. Coll. Rec.* 118 (14), 1–30.

Landrum, T. J., and McDuffie, K. A. (2010). Learning Styles in the Age of Differentiated Instruction. *Exceptionality* 18 (1), 6–17. doi:10.1080/09362830903462441

Lazarus, S. S., Kincaid, A., Thurlow, M. L., Rieke, R. L., and Dominguez, L. M. (2014). *2013 State Policies for Selected Response Accommodations on Statewide Assessments (Synthesis Report 93)*. Minneapolis, Minnesota: University of Minnesota, National Center on Educational Outcomes.

Lee, A., Browder, D. M., Flowers, C., and Wakeman, S. (2016). Teacher Evaluation of Resources Designed for Adapting Mathematics for Students with Significant Cognitive Disabilities. *Res. Pract. Persons Severe Disabilities* 41 (2), 132–137. doi:10.1177/1540796916634099

Lee, D., Reynolds, C. R., and Willson, V. L. (2003). Standardized Test Administration. *J. Forensic Neuropsychol.* 3 (3), 55–81. doi:10.1300/j151v03n03_04

Leighton, J., and Gierl, M. (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge: Cambridge University Press.

Leko, M. M., and Brownell, M. T. (2009). Crafting Quality Professional Development for Special Educators. *Teach. Exceptional Child.* 42, 64–70. doi:10.1177/004005990904200106

Leko, M. M., Brownell, M. T., Sindelar, P. T., and Kiely, M. T. (2015). Envisioning the Future of Special Education Personnel Preparation in a Standards-Based Era. *Exceptional Child.* 82, 25–43. doi:10.1177/0014402915598782

Levin, J. A., and Datnow, A. (2012). The Principal Role in Data-Driven Decision Making: Using Case-Study Data to Develop Multi-Mediator Models of Educational Reform. *Sch. Effectiveness Sch. Improvement* 23 (2), 179–201. doi:10.1080/09243453.2011.599394

Lukin, L. E., Bandalos, D. L., Eckhout, T., and Mickelson, K. (2004). Facilitating the Development of Assessment Literacy. *Educ. Meas. Issues Pract.* 23 (2), 26–32. doi:10.1111/j.1745-3992.2004.tb00156.x

Madison, M. J. (2019). Reliably Assessing Growth with Longitudinal Diagnostic Classification Models. *Educ. Meas. Issues Pract.* 38 (2), 68–78. doi:10.1111/emip.12243

Mandinach, E. B., and Gummer, E. S. (2013). A Systemic View of Implementing Data Literacy in Educator Preparation. *Educ. Res.* 42 (1), 30–37. doi:10.3102/0013189x12459803

Marion, S. F. (2018). The Opportunities and Challenges of a Systems Approach to Assessment. *Educ. Meas. Issues Pract.* 37 (1), 45–48. doi:10.1111/emip.12193

Marion, S., Lyons, S., and D'Brot, J. (2016). *Developing a Theory of Action to Support High-Quality Accountability System Design*. Dover, NH: Center for Assessment.

Marion, S., Thompson, J., Evans, C., Martineau, J., and Dadey, N. (2019). *The Challenges and Opportunities of Balanced Systems of Assessment: A Policy Brief*. Dover, NH: Center for Assessment.

Martone, A., and Sireci, S. G. (2009). Evaluating Alignment between Curriculum, Assessment, and Instruction. *Rev. Educ. Res.* 79 (4), 1332–1361. doi:10.3102/0034654309341375

McGrew, K. S., and Evans, J. (2004). Expectations for Students with Cognitive Disabilities: Is the Cup Half Empty or Half Full? Can the Cup Flow over? (Synthesis Report 55). National Center on Educational Outcomes. Available at: https://files.eric.ed.gov/fulltext/ED518644.pdf.

McMaster, K. L., Lembke, E. S., Shin, J., Poch, A. L., Smith, R. A., Jung, P.-G., et al. (2020). Supporting Teachers' Use of Data-Based Instruction to Improve Students' Early Writing Skills. *J. Educ. Psychol.* 112 (1), 1–21. doi:10.1037/edu0000358

Means, B., Chen, E., DeBarger, A., and Padilla, C. (2011). *Teachers' Ability to Use Data to Inform Instruction: Challenges and Supports*. Washington, D.C.: U. S. Department of Education, Office of Planning, Evaluation and Policy Development.

Mertler, C. A. (2009). Teachers' Assessment Knowledge and Their Perceptions of the Impact of Classroom Assessment Professional Development. *Improving Schools* 12 (2), 101–113. doi:10.1177/1365480209105575

Mislevy, R. J., Steinberg, L., and Almond, R. G. (1999). *Evidence-centered Assessment Design*. Princeton, NJ: Educational Testing Service.

Mitra, D., Mann, B., and Hlavacik, M. (2016). Opting Out: Parents Creating Contested Spaces to challenge Standardized Tests. *Edu. Pol. Anal. Arch.* 24 (31), 1–23. doi:10.14507/epaa.24.2142

Mo, Y., Singh, K., and Chang, M. (2013). Opportunity to Learn and Student Engagement: A HLM Study on Eighth Grade Science Achievement. *Educ. Res. Pol. Prac* 12 (1), 3–19. doi:10.1007/s10671-011-9126-5

Modan, N. (2020). Fast Forward: Will COVID-19 Trigger Shift from Standardized Assessments? Education Dive. Available at: https://www.educationdive.com/news.

National Council on Measurement in Education (2018). Position Statement on Theories of Action for Testing Programs. Available at: http://www.ncme.org/publications/statements.

National Research Council (2001). *Knowing what Students Know: The Science and Design of Educational Assessment*. Washington, D.C.: The National Academies Press.

Nehler, C., Clark, A., and Karvonen, M. (2019). *White Paper: Considerations for Measuring Academic Growth on Dynamic Learning Maps® (DLM®) Alternate Assessments*. ATLAS: University of Kansas.

Nichols, P., and Fulkerson, D. (2010). Informing Design Patterns Using Research on Item Writing Expertise. SRI International. Technical Report 9

Nitsch, C. (2013). Dynamic Learning Maps: The Arc Parent Focus Groups. *The Arc*.

N. M. Kingston and A. K. Clark (Editors) (2014). *Test Fraud: Statistical Detection and Methodology* (London, United Kingdom: Routledge).

No Child Left Behind Act of 2001 (2002). *20 U.S.C. §*, 6319.

O'Keefe, B., and Lewis, B. (2019). The State of Assessment: A Look Forward on Innovation in State Testing Systems. Bellwether Education Partners. Available at: https://bellwethereducation.org/.

Olson, L., and Jerald, C. (2020). *The Big Test: The Future of State Standardized Assessments*. Bengaluru, Karnataka: FutureEd.

Onosko, J. (2011). Race to the Top Leaves Children and Future Citizens behind: The Devastating Effects of Centralization, Standardization, and High Stakes Accountability. *Democracy Edu.* 19 (2), 1–11.

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., and Yao, S.-Y. (2016). The Development and Validation of a Learning Progression for Argumentation in Science. *J. Res. Sci. Teach.* 53 (6), 821–846. doi:10.1002/tea.21316

Pariseau, M. E., Fabiano, G. A., Massetti, G. M., Hart, K. C., and Pelham, W. E., Jr. (2010). Extended Time on Academic Assignments: Does Increased Time lead to Improved Performance for Children with Attention-Deficit/hyperactivity Disorder?. *Sch. Psychol. Q.* 25 (4), 236–248. doi:10.1037/a0022045

Patton, M. Q. (1989). A Context and Boundaries for a Theory-Driven Approach to Validity. *Eval. Program Plann.* 12 (4), 375–377. doi:10.1016/0149-7189(89)90054-2

Pellegrino, J. W., DiBello, L. V., and Goldman, S. R. (2016). A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments. *Educ. Psychol.* 51 (1), 59–81. doi:10.1080/00461520.2016.1145550

Perie, M., Marion, S., and Gong, B. (2009). Moving toward a Comprehensive Assessment System: A Framework for Considering Interim Assessments. *Educ. Meas. Issues Pract.* 28 (3), 5–13. doi:10.1111/j.1745-3992.2009.00149.x

Pizmony-Levy, O., and Green Saraisky, N. (2016). *Who Opts Out and Why? Results from a National Survey on Opting Out of Standardized Tests*. Teachers College: Columbia University.

Poortman, C. L., and Schildkamp, K. (2016). Solving Student Achievement Problems with a Data Use Intervention for Teachers. *Teach. Teach. Edu.* 60, 425–433. doi:10.1016/j.tate.2016.06.010

Popham, W. J. (2004). *America's Failing Schools: How Parents and Teachers Can Cope with No Child Left behind*. London, United Kingdom: Routledge.

Porter, A. C., Polikoff, M. S., and Smithson, J. (2009). Is There a De Facto National Intended Curriculum? Evidence from State Content Standards. *Educ. Eval. Pol. Anal.* 31 (3), 238–268. doi:10.3102/0162373709336465

Powell, D., Higgins, H. J., Aran, R., and Freed, A. (2009). Impact of No Child Left behind on Curriculum and Instruction in Rural Schools. *The Rural Educator* 31 (1), 19–28. doi:10.35608/ruraled.v31i1.439

Priestley, M., Biesta, G., and Robinson, S. (2013). "Teachers as Agents of Change: Teacher agency and Emerging Models of Curriculum," in *Reinventing the Curriculum: New Trends in Curriculum Policy and Practice*. Editors M. Priestley, and G. Biesta (London, United Kingdom: Bloomsbury), 187–206. doi:10.5040/9781472553195.ch-010

Race to the Top (2010). *Notice Inviting Applications for New Awards for Fiscal Year (FY) 2010, 75 Fed. Reg. 18171 (April 9, 2010)*.

Reed, D. K., and Sturges, K. M. (2012). An Examination of Assessment Fidelity in the Administration and Interpretation of reading Tests. *Remedial Spec. Edu.* 34 (5), 259–268. doi:10.1177/0741932512464580

Rentner, D. S., Kober, N., and Frizzell, M. (2016). *Listen to Us: Teacher Views and Voices*. Washington, D.C.: George Washington University, Center on Education Policy.

Roach, A. T., and Elliott, S. N. (2006). The Influence of Access to General Education Curriculum on Alternate Assessment Performance of Students with Significant Cognitive Disabilities. *Educ. Eval. Pol. Anal.* 28 (2), 181–194. doi:10.3102/01623737028002181

Roach, A. T., Niebling, B. C., and Kurz, A. (2008). Evaluating the Alignment Among Curriculum, Instruction, and Assessments: Implications and Applications for Research and Practice. *Psychol. Schs.* 45 (2), 158–176. doi:10.1002/pits.20282

Roelofs, E. (2019). "A Framework for Improving the Accessibility of Assessment Tasks," in *Theoretical and Practical Advances in Computer-Based Educational Measurement*. Editors B. P. Veldkamp, and C. Skuijter (Berlin, Germany: Springer Open), 21–45. doi:10.1007/978-3-030-18480-3_2

Rubie-Davies, C. M. (2007). Classroom Interactions: Exploring the Practices of High- and Low-Expectation Teachers. *Br. J. Educ. Psychol.* 77 (2), 289–306. doi:10.1348/000709906X101601

Rupp, A. A., Templin, J., and Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York: Guilford Press.

Ruppar, A. L., Neeper, L. S., and Dalsen, J. (2016). Special Education Teachers' Perceptions of Preparedness to Teach Students with Severe Disabilities. *Res. Pract. Persons Severe Disabilities* 41 (4), 273–286. doi:10.1177/1540796916672843

Ryndak, D. L., Taub, D., Jorgensen, C. M., Gonsier-Gerdin, J., Arndt, K., Sauer, J., et al. (2014). Policy and the Impact on Placement, Involvement, and Progress in General Education. *Res. Pract. Persons Severe Disabilities* 39 (1), 65–74. doi:10.1177/1540796914533942

Sessoms, J., and Henson, R. A. (2018). Applications of Diagnostic Classification Models: A Literature Review and Critical Commentary. *Meas. Interdiscip. Res. Perspect.* 16 (1), 1–17. doi:10.1080/15366367.2018.1435104

Sheldon, E., and Erickson, K. (2020). Emergent Literacy Instruction for Students with Significant Disabilities in the Regular Classroom. *Assistive Tech. Outcomes Benefits* 14 (1), 135–160.

Shepard, L. A. (2018). Learning Progressions as Tools for Assessment and Learning. *Appl. Meas. Edu.* 31 (2), 165–174. doi:10.1080/08957347.2017.1408628

Shin, H. J., von Davier, M., and Yamamoto, K. (2019). "Investigating Rater Effects in International Large-Scale Assessments," in *Theoretical and Practical Advances in Computer-Based Educational Measurement*. Editors B. Veldkamp, and C. Sluijter (Berlin, Germany: Springer Open), 249–268. doi:10.1007/978-3-030-18480-3_13

Sireci, S. G. (2015). "A Theory of Action for Test Validation," in *The Next Generation of Testing: Common Core Standards, Smarter-Balanced, PARCC, and the Nationwide Testing Movement*. Editors H. Jiao and R. Lissitz (Charlotte, North Carolin: Information Age Publishing), 251–269.

Sireci, S. G., Scarpati, S. E., and Li, S. (2005). Test Accommodations for Students with Disabilities: An Analysis of the Interaction Hypothesis. *Rev. Educ. Res.* 75 (4), 457–490. doi:10.3102/00346543075004457

Sireci, S. G. (2020). Standardization and UNDERSTAND Ardization in Educational Assessment. *Educ. Meas. Issues Pract.* 39 (3), 100–105. doi:10.1111/emip.12377

Skaggs, G., Hein, S. F., and Wilkins, J. L. M. (2020). Using Diagnostic Profiles to Describe Borderline Performance in Standard Setting. *Educ. Meas. Issues Pract.* 39, 45–51. doi:10.1111/emip.12228

Südkamp, A., Kaiser, J., and Möller, J. (2012). Accuracy of Teachers' Judgments of Students' Academic Achievement: A Meta-Analysis. *J. Educ. Psychol.* 104, 743–762. doi:10.1037/a0027627

Swinburne Romine, R., Karvonen, M., and Clark, A. K. (2016). *Validity Evidence to Support Alternate Assessment Score Uses: Fidelity And Response Processes [Paper Presentation]*. Washington, DC: Annual meeting of the National Council on Measurement in Education.

Templin, J., and Bradshaw, L. (2013). Measuring the Reliability of Diagnostic Classification Model Examinee Estimates. *J. Classif* 30, 251–275. doi:10.1007/s00357-013-9129-4

Timberlake, M. T. (2014). Weighing Costs and Benefits. *Res. Pract. Persons Severe Disabilities* 39 (2), 83–99. doi:10.1177/1540796914544547

Transition and Postsecondary Programs for Students with Intellectual Disabilities (2008). *20 U.S.C. §*, 1140f–1140i.

U.S. Department of Education (2018). A State's Guide to the U.S. Department of Education's Assessment Peer Review Process. Office of Elementary and Secondary Education. Available at: https://www2.ed.gov/admins/lead/account/saa/assessmentpeerreview.pdf.

U.S. Department of Education (2016). *Future Ready Learning: Reimagining the Role of Technology in Education*. Washington, DC: Office of Educational Technology.

U.S. Department of Education (2020). Innovative Assessment Demonstration Authority (IADA). Available at: https://www2.ed.gov/admins/lead/account/iada/index.html.

U.S. Department of Education (2010). *Race to the Top Fund Assessment Program. 75 Fed. Reg.*, 18171–18185.

Vu, P., Cao, V., Vu, L., and Cepero, J. (2014). Factors Driving Learner success in Online Professional Development. *Int. Rev. Res. Open Distributed Learn.* 15 (3), 120–139. doi:10.19173/irrodl.v15i3.1714

Wakeman, S. Y., Karvonen, M., Flowers, C., and Ruther, L. (in press). "Alternate Assessments and Monitoring Student Progress in Inclusive Classrooms," in *Handbook of Research and Practice for Inclusive Schools*. Editors J. McLeskey, F. Spooner, B. Algozzine, and N. L. Waldron (New York: Routledge).

Walsh, R. L., and Hodge, K. A. (2018). Are We Asking the Right Questions? an Analysis of Research on the Effect of Teachers' Questioning on Children's Language during Shared Book reading with Young Children. *J. Early Child. Literacy* 18 (2), 264–294. doi:10.1177/1468798416659124

Wehmeyer, M. L., Shogren, K. A., Palmer, S. B., Williams-Diehm, K. L., Little, T., and Boulton, A. (2012). Impact of the Self-Determined Learning Model of Instruction on Self-Determination: A Randomized-Trial Control Group Study. *Except Child.* 78 (2), 135–153. doi:10.1177/001440291207800201

Wilson, M. (2018). Making Measurement Important for Education: The Crucial Role of Classroom Assessment. *Educ. Meas. Issues Pract.* 37 (1), 5–20. doi:10.1111/emip.12188

Wilson, M. (2009). Measuring Progressions: Assessment Structures Underlying a Learning Progression. *J. Res. Sci. Teach.* 46 (6), 716–730. doi:10.1002/tea.20318

Winter, P. C. (2010). "Comparability and Test Variations," in *Evaluating the Comparability of Scores from Achievement Test Variations*. Editor P. C. Winter (Washington, DC: CCSSO), 1–12.

Workforce Innovation and Opportunity Act (2014). *29 U.S.C. §*, 3101.

Zwick, R., and Mislevy, R. (2011). *Scaling Aand Llinking Tthrough-Ccourse Ssummative Aassessments [Paper Ppresentation]*. Atlanta, GA, United States: Invitational Research Symposium on Through-Course Summative Assessments.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX A

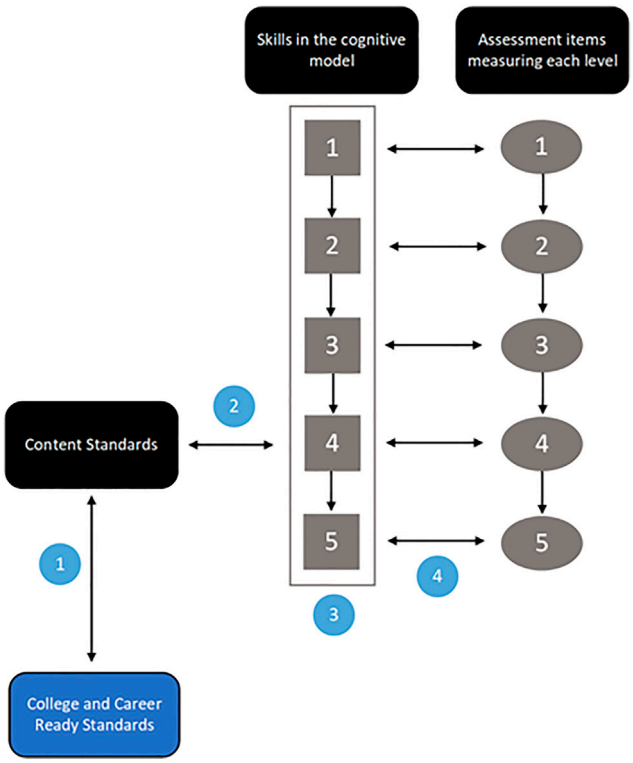


FIGURE A1 | Relationships among content structures of the DLM assessment system. 1) Alternate content standards specify academic targets and are aligned to college and career standards. 2) For each alternate content standard, skills in the cognitive model (learning map) align with the standard as the "target" level for that standard. 3) In English language arts and mathematics, each standard has three additional levels before the target level and one that extends beyond the target level to give all students opportunity to work on grade-level academic content. 4) The skills measured at each level have associated items that are grouped together and administered in a short assessment. Each student takes a series of assessments to cover blueprint requirements. Teachers choose which content standards and at which levels to administer assessments to meet all blueprint requirements.