2021–2022 Technical Manual

Year-End Model

December 2022

Copyright © 2022 Accessible Teaching, Learning, and Assessment Systems (ATLAS)

All rights reserved. Any or all portions of this document may be reproduced and distributed without prior permission provided the source is cited as:

Dynamic Learning Maps Consortium. (2022, December). *2021–2022 Technical Manual—Year-End Model*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems.

Acknowledgements

The publication of this technical manual update represents the culmination of a body of work in the service of creating a meaningful assessment system designed to serve students with the most significant cognitive disabilities. Hundreds of people have contributed to this undertaking. We acknowledge them all for their contributions.

Many contributors made the writing of this technical manual possible. Dynamic Learning Maps® (DLM®) staff who made significant writing contributions to this technical manual are listed below with gratitude.

W. Jake Thompson, Ph.D., *Assistant Director for Psychometrics*
Amy K. Clark, Ph.D., *Associate Director for Operational Research*

The authors also wish to acknowledge Ashley Hirt, Jeffrey Hoover, Elizabeth Kavitsky, Jennifer Kobrin, Brooke Nash, and Noelle Pablo for their role in developing, organizing, and compiling this manual. The authors also wish to acknowledge Brianna Beitling, Lisa Braun, Steve Breedlove, Amber Cavasos, Alson Cole, Karen Erickson, Zachary Hopper, Eun Mi Kim, Sarah Koebley, Jessica Lancaster, Mari Langas, Susan Martin, Jonathan Schuster, Vanessa Simmering, Mark Thompson, Delaney Wilson, and Nicole Zdeb for their contributions to this manual. Finally, the authors wish to thank Kristy Bledsoe, Lucas Cooper, Justin Dean, Aaron Gates, Whitney Lohrenz, and Sara Lundberg for their editing and project management work. The present technical manual is a refresh of the 2014–2015 technical manual. For a list of project staff who contributed to the original technical documentation for the DLM System, please see the *2014–2015 Technical Manual—Year-End Model*, and the subsequent annual technical manual updates.

We are also grateful for the contributions of the members of the DLM Technical Advisory Committee who graciously provided their expertise and feedback on the DLM System, including drafts of the technical manual chapters. Members of the Technical Advisory Committee during the 2021–2022 operational year include:

Russell Almond, Ph.D., *Florida State University*
Karla Egan, Ph.D., *EdMetric*
Claudia Flowers, Ph.D., *University of North Carolina at Charlotte*
Robert Henson, Ph.D., *University of North Carolina at Greensboro*
Joan Herman, Ed.D., *University of California, Los Angeles*
James Pellegrino, Ph.D., *University of Illinois Chicago*
Edward Roeber, Ph.D., *Michigan Assessment Consortium*
David Williamson, Ph.D., *The College Board*
Phoebe Winter, Ph.D., *Independent Consultant*

Contents

List of Tables

List of Figures

1. Overview

The Dynamic Learning Maps® (DLM®) Alternate Assessment System assesses student achievement in English language arts (ELA), mathematics, and science for students with the most significant cognitive disabilities in grades 3–8 and high school. The purpose of the system is to improve academic experiences and outcomes for students with the most significant cognitive disabilities by setting high and actionable academic expectations and providing appropriate and effective supports to educators. Results from the DLM alternate assessment are intended to support interpretations about what students know and are able to do and to support inferences about student achievement in the given subject. Results provide information that can guide instructional decisions as well as information for use with state accountability programs.

The DLM System is developed and administered by Accessible Teaching, Learning, and Assessment Systems (ATLAS), a research center within the University of Kansas's Achievement and Assessment Institute. The DLM System is based on the core belief that all students should have access to challenging, grade-level content. Online DLM assessments give students with the most significant cognitive disabilities opportunities to demonstrate what they know in ways that traditional paper-and-pencil assessments cannot.

A complete technical manual was created after the first operational administration of ELA and mathematics in 2014–2015. Because the DLM System adopts a continuous improvement model, each year incremental changes are implemented and additional technical evidence is collected, which are described in annual technical manual updates. The 2015–2016 to 2020–2021 technical manual updates summarize only the new information and evidence for that year and refer the reader to the 2014–2015 version for complete descriptions of the system. Example DLM System changes across years include updates to the online test delivery engine, item banks, and administration procedures.

The 2021–2022 manual is a complete technical manual that provides comprehensive information and evidence for the ELA and mathematics assessment system. The manual provides a full description of the current DLM System, but also describes the more substantive changes made to the system over time. To help orient readers to this manual, a glossary of common terms is provided in Appendix A.1. Due to differences in the development timeline for science, the science technical manual is prepared separately (see Dynamic Learning Maps Consortium [DLM Consortium], 2022a).

1.1. Current DLM Collaborators for Development and Implementation

The DLM System was initially developed by a consortium of state education agencies (SEAs) beginning in 2010 and expanding over the years, with a focus on ELA and mathematics. The development of a DLM science assessment began with a subset of the participating SEAs in 2014. Due to the differences in the development timelines, separate technical manuals are prepared for ELA and mathematics and science. During the 2021–2022 academic year, DLM assessments were available to students in 21 states: Alaska, Arkansas, Colorado, Delaware, District of Columbia, Illinois, Iowa, Kansas, Maryland, Missouri, New Hampshire, New Jersey, New Mexico, New York, North Dakota, Oklahoma, Pennsylvania, Rhode Island, Utah, West Virginia, and Wisconsin. One SEA partner, Colorado, only administered assessments in ELA and mathematics; one SEA partner, District of Columbia, only administered assessments in science. The DLM Governance Board is comprised of two representatives from the SEAs of each member state. Representatives have expertise in special education and state assessment administration. The DLM

Governance Board advises on the administration, maintenance, and enhancement of the DLM System.

In addition to ATLAS and governance board states, other key partners include the Center for Literacy and Disability Studies at the University of North Carolina at Chapel Hill and Agile Technology Solutions at the University of Kansas.

The DLM System is also supported by a Technical Advisory Committee (TAC). DLM TAC members possess decades of expertise, including in large-scale assessments, accessibility for alternate assessments, diagnostic classification modeling, and assessment validation. The DLM TAC provides advice and guidance on technical adequacy of the DLM assessments.

1.2. Student Population

The DLM System serves students with the most significant cognitive disabilities, sometimes referred to as students with extensive support needs, who are eligible to take their state's alternate assessment based on alternate academic achievement standards. This population is, by nature, diverse in learning style, communication mode, support needs, and demographics.

Students with the most significant cognitive disabilities have a disability or multiple disabilities that significantly impact intellectual functioning and adaptive behavior. When adaptive behaviors are significantly impacted, the individual is unlikely to develop the skills to live independently and function safely in daily life. In other words, significant cognitive disabilities impact students in and out of the classroom and across life domains, not just in academic settings. The DLM System is designed for students with these significant instruction and support needs.

The DLM System provides the opportunity for students with the most significant cognitive disabilities to show what they know, rather than focusing on deficits (Nitsch, 2013). These are students for whom general education assessments, even with accommodations, are not appropriate. These students learn academic content aligned to grade-level content standards, but at reduced depth, breadth, and complexity. The content standards are derived from the Common Core State Standards (CCSS, National Governors Association Center for Best Practices and Council of Chief State School Officers, 2010), often referred to in this manual as college and career readiness standards, and are called Essential Elements (EEs). The EEs are the learning targets for DLM assessments for grades 3–12 in ELA and mathematics. Chapter 2 of this manual provides a complete description of the content structures for the DLM assessment, including the EEs.

While all states provide additional interpretation and guidance to their districts, three general participation guidelines are considered for a student to be eligible for the DLM alternate assessment.

1. The student has a significant cognitive disability, as evident from a review of the student records that indicates a disability or multiple disabilities that significantly impact intellectual functioning and adaptive behavior.
2. The student is primarily being instructed (or taught) using the DLM EEs as content standards, as evident by the goals and instruction listed in the IEP for this student that are linked to the enrolled grade level DLM EEs and address knowledge and skills that are appropriate and challenging for this student.
3. The student requires extensive direct individualized instruction and substantial supports to achieve

measurable gains in the grade-and age-appropriate curriculum. The student (a) requires extensive, repeated, individualized instruction and support that is not of a temporary or transient nature and (b) uses substantially adapted materials and individualized methods of accessing information in alternative ways to acquire, maintain, generalize, demonstrate and transfer skills across multiple settings.

The DLM System eligibility criteria also provide guidance on specific considerations that are not acceptable for determining student participation in the alternate assessment:

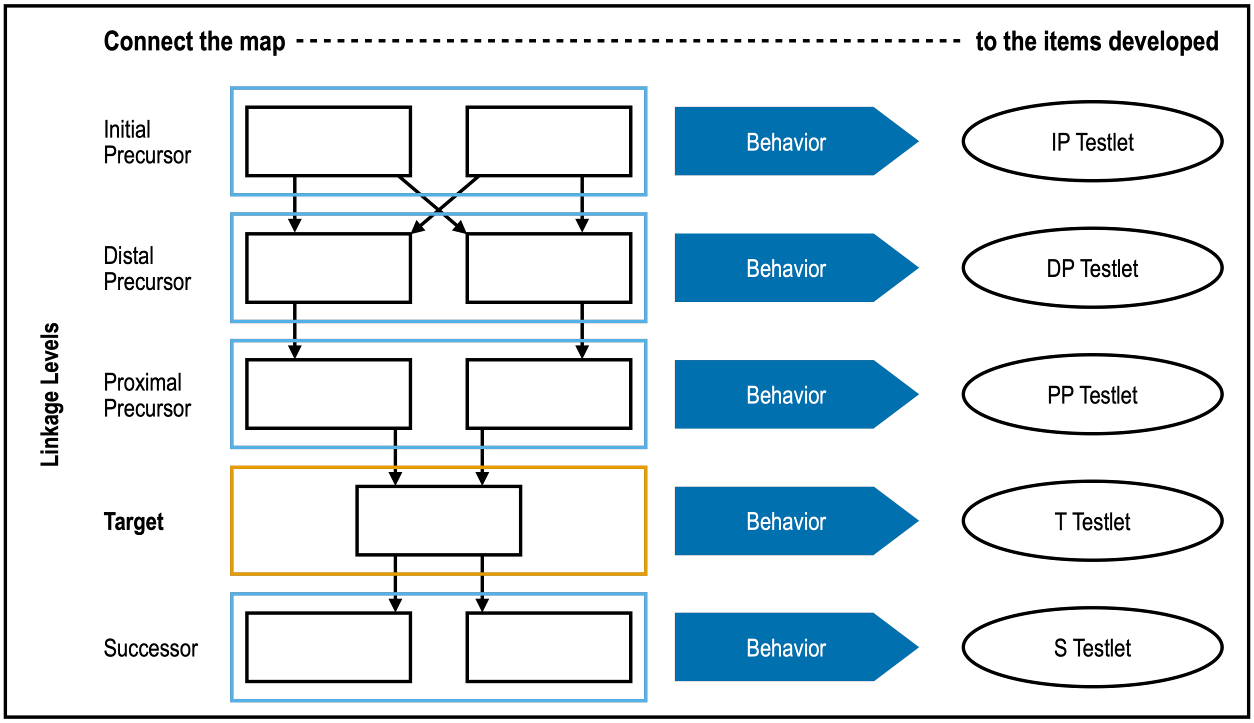
- a disability category or label
- poor attendance or extended absences
- native language, social, cultural, or economic differences
- expected poor performance on the general education assessment
- receipt of academic or other services
- educational environment or instructional setting
- percent of time receiving special education
- English Language Learner status
- low reading or achievement level
- anticipated disruptive behavior
- impact of student scores on accountability system
- administrator decision
- anticipated emotional duress
- need for accessibility supports (e.g., assistive technology) to participate in assessment

1.3. Assessment

The DLM assessments are delivered as a series of testlets, each of which contains an unscored engagement activity and three to nine items. The assessment items are aligned to specific nodes in the learning map neighborhoods corresponding to each EE. These neighborhoods are illustrations of the connections between the knowledge and skills necessary to meet extended grade-level academic content standards. Individual concepts and skills are represented as points on the map, known as nodes. Small collections of nodes are called linkage levels, which provide access to the EEs at different levels of complexity to meet the needs of all students taking the DLM assessments. Assessment items are written to align to map nodes at one of the five linkage levels and are clustered into testlets (see Figure 1.1; for a complete description of the content structures, including learning map neighborhoods and nodes, and linkage levels, see Chapter 2 of this manual). Except for writing testlets, each testlet measures a single linkage level.

Figure 1.1

Relationship Between DLM Map Nodes and Items in Testlets



Note. Small black boxes represent nodes in the DLM learning map. Blue and orange boxes represent collections of nodes in linkage levels. The orange box denotes the Target linkage level for the EE. There may be more than one node at any linkage level.

Assessment blueprints consist of the EEs, or targeted skills that align most closely with grade-level expectations, prioritized for assessment by the DLM Governance Board. The blueprint specifies the number of EEs that must be assessed in each subject. To achieve blueprint coverage, each student is administered a series of testlets. Testlet delivery and test management are achieved through an online platform, Kite® Suite. Kite Suite contains applications for educators to manage student data and assign testlets to students, and also contains a portal through which students can access testlets. Student results are based on evidence of mastery of the linkage levels for every assessed EE. The student interface used to administer the DLM assessments was designed specifically for students with the most significant cognitive disabilities. It maximizes space available to display content, decreases space devoted to tool-activation buttons within a testing session, and minimizes the cognitive load related to test navigation and response entry. More information about the Kite Suite can be found in Chapter 4 of this manual.

For all aspects of the DLM System, our overarching goal is to align with the latest research from a full range of accessibility lenses (e.g., universal design of assessment, physical and sensory disabilities, special education) to ensure the assessments are accessible for the widest range of students who will be interacting with the content. In order to exhibit the assessed skills, students must be able to interact with the assessment in the means most appropriate for them. Thus, the DLM assessments provide different pathways of student interaction and ensure those means can be used while maintaining the validity of the inferences from and intended uses of the DLM System. These pathways begin in the earliest stages of assessment and content development, from the creation of the learning maps to item writing and assessment administration. We seek both content adherence and accessibility by maximizing the quality of the assessment process while preserving evidence of the targeted cognition. This balance of ensuring accessibility for all students while protecting the validity of the intended uses is discussed throughout this manual where appropriate. Additionally, the overarching goal of accessible content is reflected in the Theory of Action for the DLM System, which is described in the following section.

1.4. Assessment Models

There are two assessment models for the DLM alternate assessment. Each state chooses its own model.

- **Instructionally Embedded model.** There are two instructionally embedded testing windows: fall and spring. The assessment windows are structured so that testlets can be administered at appropriate points in instruction. Each window is approximately 15 weeks long. Educators have some choice of which EEs to assess, within constraints defined by the assessment blueprint. For each EE, the system recommends a linkage level for assessment, and the educator may accept the recommendation or choose another linkage level. Recommendations are based on information collected about the students' expressive communication and academic skills. At the end of the year, summative results are based on mastery estimates for the assessed linkage levels for each EE (including performance on all testlets from both the fall and spring windows) and are used for accountability purposes. There are different pools of operational testlets for the fall and spring windows. In 2021–2022, the states adopting the Instructionally Embedded model were Arkansas, Delaware, Iowa, Kansas, Missouri, and North Dakota.
- **Year-End model.** During a single operational testing window in the spring, all students take testlets that cover the whole blueprint. The window is approximately 13 weeks long, and test administrators may administer the required testlets throughout the window. Students are assigned their first testlet based on information collected about their expressive communication and academic skills. Each testlet assesses one linkage level and EE. The linkage level for each subsequent testlet varies according to student performance on the previous testlet. Summative assessment results reflect the student's performance and are used for accountability purposes each school year. In Year-End states, instructionally embedded assessments are available during the school year but are optional and do not count toward summative results. In 2021–2022, the states adopting the Year-End model were Alaska, Colorado, Illinois, Maryland, New Hampshire, New Jersey, New Mexico, New York, Oklahoma, Pennsylvania, Rhode Island, Utah, West Virginia, and Wisconsin.

Information in this manual is common to both models wherever possible and is specific to the Year-End model where appropriate. The Instructionally Embedded model has a separate technical manual.

1.5. Theory of Action and Interpretive Argument

The Theory of Action that guided the design of the DLM System was formulated in 2011, revised in December 2013, and revised again in 2019. It expresses the belief that high expectations for students with the most significant cognitive disabilities, combined with appropriate educational supports and diagnostic tools for educators, results in improved academic experiences and outcomes for students and educators.

The process of articulating the Theory of Action started with identifying critical problems that characterize large-scale assessment of students with the most significant cognitive disabilities so that the DLM System design could alleviate these problems. For example, traditional assessment models treat knowledge as unidimensional and are independent of teaching and learning, yet teaching and learning are multidimensional activities and are central to strong educational systems. Also, traditional assessments focus on standardized methods and do not allow various, non-linear approaches for demonstrating learning even though students learn in various and non-linear ways. In addition, using assessments for accountability pressures educators to use assessments as models for instruction with assessment preparation replacing best-practice instruction. Furthermore, traditional assessment systems often emphasize objectivity and reliability over fairness and validity. Finally, negative, unintended consequences for students must be addressed and eradicated.

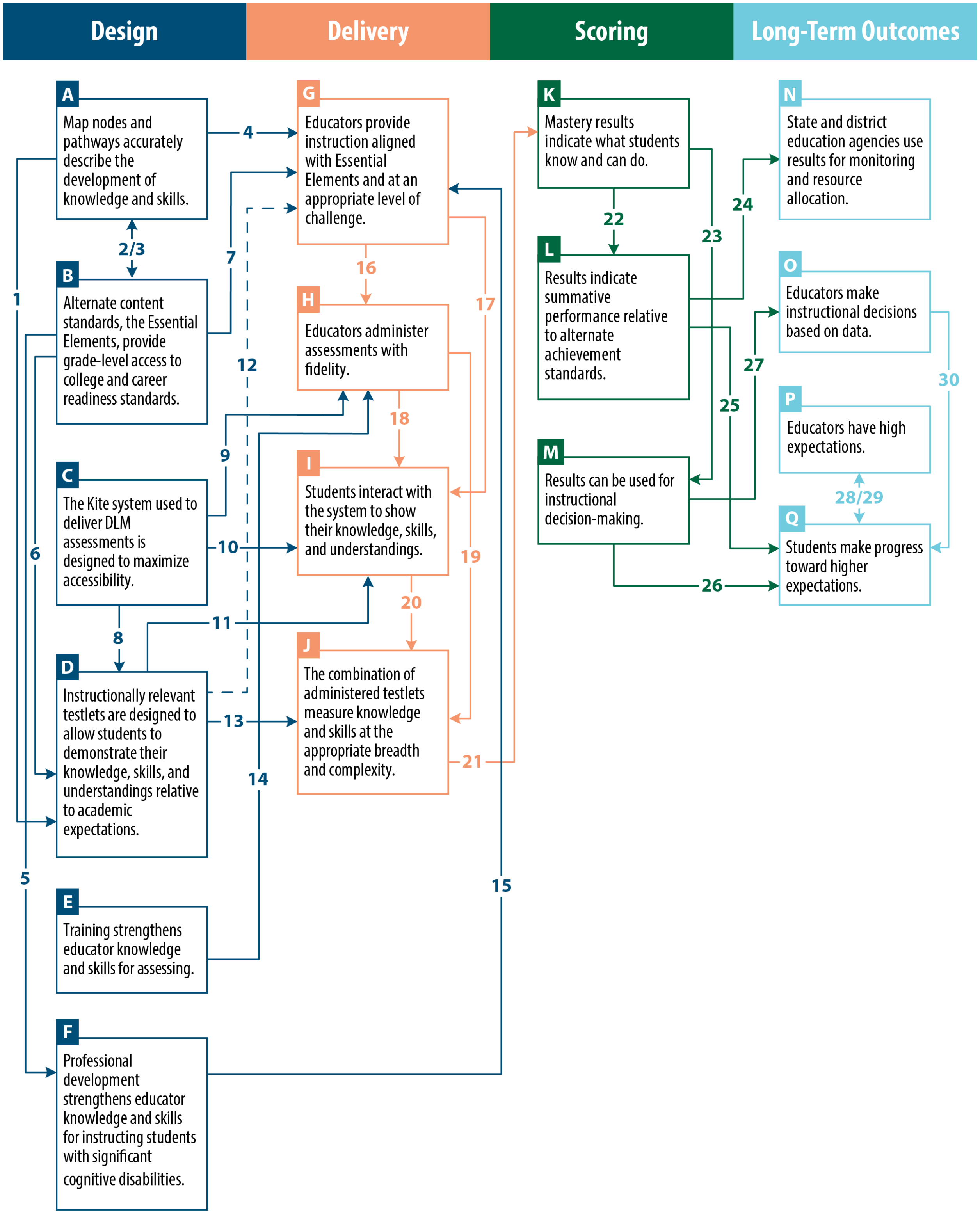
The DLM Theory of Action expresses a commitment to provide students with the most significant cognitive disabilities access to highly flexible cognitive and learning pathways and an assessment system that is capable of validly and reliably evaluating their achievement. Ultimately, students will make progress toward higher expectations, educators will make instructional decisions based on data, educators will hold higher expectations of students, and state and district education agencies will use results for monitoring and resource allocation.

The DLM Governance Board adopted an argument-based approach to assessment validation. The validation process began in 2013 by defining with governance board members the policy uses of DLM assessment results. We followed this with a three-tiered approach, which included specification of 1) a Theory of Action defining statements in the validity argument that must be in place to achieve the goals of the system; 2) an interpretive argument defining propositions that must be evaluated to support each statement in the Theory of Action; and 3) validity studies to evaluate each proposition in the interpretive argument.

After identifying these overall guiding principles and anticipated outcomes, specific elements of the DLM Theory of Action were articulated to inform assessment design and to highlight the associated validity arguments. The Theory of Action includes the assessment's intended effects (long-term outcomes), statements related to design, delivery and scoring, and action mechanisms (i.e., connections between the statements; see Figure 1.2). The chain of reasoning in the Theory of Action is demonstrated broadly by the order of the four sections from left to right. Dashed lines represent connections that are present when the optional instructionally embedded assessments are utilized. Design statements serve as inputs to delivery, which informs scoring and reporting, which collectively lead to the long-term outcomes for various stakeholders. The chain of reasoning is made explicit by the numbered arrows between the statements.

Figure 1.2

Dynamic Learning Maps Theory of Action



1.6. Key Features

Consistent with the Theory of Action and interpretive argument, key elements were identified to guide the design and delivery of the DLM System. The following section briefly describes the key features of the DLM System, followed by an overview of chapters indicating where to find evidence for each statement in the Theory of Action. Terms are defined in the glossary (Appendix A.1).

1. **Fine-grained learning maps that describe how students acquire knowledge and skills.**
   Learning maps are a unique key feature of the DLM System and drive the development of all other components. While the DLM learning maps specify targeted assessment content, they also reflect a synthesis of research on the relationships and learning pathways among different concepts, knowledge, and cognitive processes. Therefore, DLM maps demonstrate multiple ways that students can acquire the knowledge and skills necessary to reach targeted expectations, and they provide a framework that supports inferences about student learning needs (Bechard et al., 2012). The use of DLM maps helps to realize a vision of a cohesive, comprehensive system of assessment. DLM learning map development is described in Chapter 2 of this manual.
2. **A set of learning targets for instruction and assessment, as defined by the EEs and linkage levels, that provide grade-level access to college and career readiness standards.**
   Crucial to the use of fine-grained learning maps for instruction and test development is the selection of critical nodes that serve as learning targets aligned to the grade level expectations defined in the EEs. These are accompanied by the selection of nodes that build up to and extend the knowledge, skills, and abilities required to achieve the EEs for each grade and subject (i.e., linkage levels). This neighborhood of nodes forms a local EE-specific learning progression that is informative to both instruction and assessment. The development of EEs and the selection of nodes for assessment at each linkage level are described in Chapter 2 of this manual.
3. **Instructionally relevant assessments that reinforce the primacy of instruction and are designed to allow students to demonstrate their knowledge, skills, and understandings relative to academic expectations.**
   The development of an instructionally relevant assessment begins by creating items using principles of evidence-centered design and Universal Design for Learning and linking related items together into meaningful groups, called testlets. These assessments necessarily take different forms depending on the learning characteristics of students and the concepts being taught. Item and testlet design are described in Chapter 3 of this manual.
4. **Accessibility by design.**
   Accessibility is a prerequisite to validity, the degree to which interpretation of test results is justifiable for a particular purpose and supported by evidence and theory (American Educational Research Association et al. [AERA et al.], 2014). Therefore, throughout all phases of development, the DLM System was designed with accessibility in mind to support both learning and assessment. Students must understand what is being asked in an item or task and have the tools to respond in order to demonstrate what they know and can do (Karvonen, Bechard, & Wells-Moreaux, 2015). The DLM alternate assessment provides accessible content, accessible delivery via technology and student-specific linkage level assignment. Since all students taking an alternate assessment based on alternate academic achievement standards are students with significant cognitive disabilities, accessibility supports are universally available. The emphasis is on helping educators select the appropriate accessibility features and tools for each individual student. Accessibility considerations

are described in Chapter 3 (testlet development) and Chapter 4 (accessibility during test administration) of this manual.

5. **Training and professional development that strengthens educator knowledge and skills for instructing and assessing students with significant cognitive disabilities.**
   The DLM System provides comprehensive support and training to state education agency staff and local educators in order to administer the assessment with fidelity. The DLM System also supports educators with professional development modules which address instruction in ELA and mathematics and support educators in creating Individual Education Programs that are aligned with the DLM EEs. Modules are designed to meet the needs of all educators, especially those in rural and remote areas, offering educators just-in-time, on-demand training. DLM professional development modules support educator continuing education through learning objectives designed to deepen knowledge and skills in instruction and assessment for students with the most significant cognitive disabilities. Training and professional development are described in Chapter 9 of this manual.
6. **Assessment results that are readily actionable.**
   Due to the unique characteristics of a map-based system, DLM assessments require new approaches to psychometric analysis and modeling, with the goal of assuring accurate inferences about student performance relative to the content as it is organized in the DLM learning maps. Each EE has related nodes at five associated levels of complexity (i.e., linkage levels). Diagnostic classification modeling is used to estimate a student's likelihood of mastery for each assessed EE and linkage level. A student's overall performance level in the subject is determined by aggregating linkage level mastery information across EEs. The DLM modeling approach is described in Chapter 5, information about standard setting is described in Chapter 6, and the technical quality of the assessment and student performance is summarized in Chapter 7 of this manual.
   The DLM scoring model supports reports that can be immediately used to guide instruction and describe levels of mastery. Student reports include two parts: a performance profile and a learning profile. The performance profile aggregates linkage level mastery information for reporting on each conceptual area and for the subject overall. The learning profile provides fine-grained skill mastery for each assessed EE and linkage level and is designed to support educators in making individualized instructional decisions. Score report design is described in Chapter 7 of this manual.

1.7. Technical Manual Overview

This manual provides evidence to support the assertion of technical quality and the validity of intended uses of the assessment based on statements in the Theory of Action.

Chapter 1 provides the theoretical underpinnings of the DLM System, including a description of the DLM collaborators, the target student population, an overview of the assessment, and an introduction to the Theory of Action and interpretive argument.

Chapter 2 describes the content structures of the DLM System and addresses the design statements in the Theory of Action that the map nodes and pathways accurately describe the development of knowledge and skills, and that the alternate content standards (i.e., the EEs) provide grade level access to college and career readiness standards. The chapter describes how EEs were used to build bridges from grade-level college and career readiness content standards to academic expectations for students with the most significant cognitive disabilities. Chapter 2 also describes how the EEs were developed, the process by which the DLM learning maps were developed, and how the learning maps were linked to the EEs. Extensive, detailed work was necessary to create the DLM learning maps in light of the CCSS and the needs of the student population. Based on in-depth literature reviews and research as well as extensive input from experts and practitioners, the DLM learning maps are the conceptual and content basis for the DLM System. The chapter describes how the learning maps are organized by claims and conceptual areas, and the relationship between elements in the DLM System. Chapter 2 then describes the expert evaluation of the learning map structure, the development of the assessment blueprints and subsequent blueprint revision, and the external alignment study for the DLM System. The chapter concludes with a description of the learning maps that are used for the operational assessment.

Chapter 3 outlines procedural evidence related to test content and item quality, addressing the Theory of Action's statement that instructionally relevant testlets are designed to allow students to demonstrate their knowledge, skills, and understandings relative to academic expectations. The chapter relates how evidence-centered design was used to develop testlets—the basic unit of test delivery for the DLM alternate assessment. Further, the chapter describes how nodes in the DLM learning maps and EEs were used to develop concept maps to specify item and testlet development. Using principles of Universal Design, the entire development process accounted for the student population's characteristics, including accessibility and bias considerations. Chapter 3 includes summaries of external reviews for content, bias, and accessibility, and field tests. The final portion of the chapter describes results of differential item functioning (DIF) analyses ensuring that testlets are fair for all student sub-groups.

Chapter 4 provides an overview of the fundamental design elements that characterize test administration and how each element supports the DLM Theory of Action, specifically the statement that the system used to deliver DLM assessments is designed to maximize accessibility. The chapter describes the assessment delivery system (Kite Suite) and assessment modes (computer delivery and educator delivery). The chapter also relates how students are assigned testlets and describes the test administration process, resources that support test administrators, and test security. Chapter 4 also describes evidence related to each of the delivery statements in the Theory of Action: the combination of administered testlets measure knowledge and skills at the appropriate breadth, depth, and complexity; educators provide instruction aligned with EEs and at an appropriate level of challenge; educators administer assessments with fidelity; and students interact with the system to show their knowledge, skills, and understandings.

Chapter 5 addresses the Theory of Action's scoring statement that mastery results indicate what students know and can do. The chapter demonstrates how the DLM project draws upon a well-established research base in cognition and learning theory and uses operational psychometric methods that are relatively uncommon in large-scale assessments to provide feedback about student progress and learning acquisition. This chapter describes the psychometric model that underlies the DLM project and describes the process used to estimate item and student parameters from student test data and evaluate model fit.

Chapter 6 addresses the scoring statement that results indicate summative performance relative to alternate achievement standards. This chapter describes the methods, preparations, procedures, and results of the standard setting meeting and the follow-up evaluation of the impact data and cut points based on the 2014–2015 standard setting study, as well as the standards adjustment process that occurred in spring 2022. This chapter also describes the process of developing grade- and subject-specific performance level descriptors in ELA and mathematics.

Chapter 7 reports the 2021–2022 operational results. The chapter first reports student participation data, and then details the percent of students achieving at each performance level, as well as subgroup performance by gender, race, ethnicity, and English learner status. The chapter describes scoring rules used to determine linkage level mastery and the percent of students who showed mastery at each linkage level. The chapter next describes the inter-rater reliability of writing sample ratings. The chapter also presents evidence related to postsecondary opportunities for students taking DLM assessments. Finally, the chapter addresses the statement that results can be used for instructional decision-making with a description of data files, score reports, and interpretive guidance.

Chapter 8 focuses on reliability evidence, including a description of the methods used to evaluate assessment reliability and a summary of results by the linkage level, EE, conceptual area, and subject (overall performance). This evidence is used to support all three scoring statements in the Theory of Action.

Chapter 9 addresses the Theory of Action's statements that training strengthens educator knowledge and skills for assessing; and professional development strengthens educator knowledge and skills for instructing and assessing students with significant cognitive disabilities. The chapter describes the training and professional development that was offered across states administering DLM assessments, including the 2021–2022 training for state and local education agency staff, the required test administrator training, and the professional development available to support instruction. Participation rates and evaluation results from 2021–2022 instructional professional development are included.

Chapter 10 synthesizes the evidence provided in the previous chapters. It evaluates how the evidence supports the statements in the Theory of Action as well as the long-term outcomes: students make progress toward higher expectations, educators make instructional decisions based on data, educators have high expectations, and state and district education agencies use results for monitoring and resource allocation. The chapter ends with a description of our future research and ongoing initiatives for continuous improvement.

2. Content Structures

This chapter describes the key assessment structures that support the Dynamic Learning Maps® (DLM®) Alternate Assessment System's purpose and program goals.

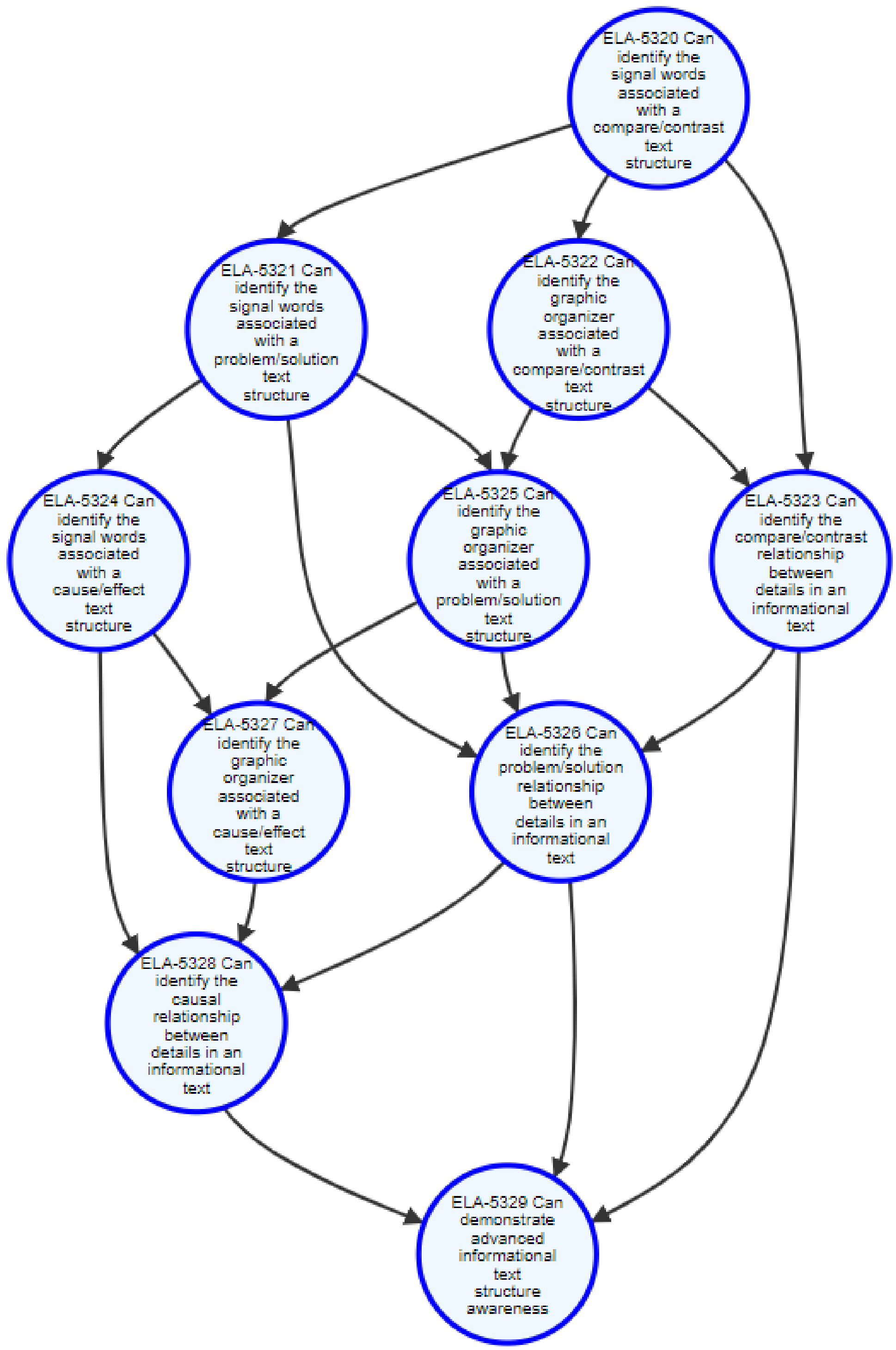
The DLM System is based on large, fine-grained learning maps. These learning maps are highly connected representations of the pathways for how academic content is acquired, as reflected in research literature. The DLM maps consist of nodes that represent discrete knowledge, skills, and understandings in either English language arts (ELA) or mathematics, as well as important cognitive skills that support the acquisition of the learning targets associated with grade-level content standards. Connections between nodes represent the development of the knowledge, skills, and understandings. With approximately 1,900 nodes in the ELA map, 2,400 nodes in the mathematics map, and 150 foundational nodes[1] that are associated with both subjects, the DLM learning maps go beyond traditional learning progressions to include multiple pathways by which students with the most significant cognitive disabilities may acquire the academic content.

Seen in its entirety, the DLM learning maps are highly complex. A closer look at smaller sections of the DLM learning maps reveals how the discrete nodes are described and connected. Figure 2.1 provides an illustration of a small segment of the DLM ELA learning map. The learning maps are read from the top down, moving from the least to most complex knowledge, skills, and understandings.

[1] Foundational nodes represent basic skills that are required across subjects and are important precursors to developing competency in learning targets associated with grade-level academic standards.

Figure 2.1

Sample Excerpt From the DLM English Language Arts Learning Map



Extensive, detailed work was necessary to establish and refine the DLM learning maps in light of the

Common Core State Standards (CCSS, National Governors Association Center for Best Practices and Council of Chief State School Officers, 2010) and the needs of the student population. Guided by in-depth reviews of literature and research, as well as extensive input from experts and practitioners, the DLM learning maps are the conceptual and content basis for the DLM System.

This chapter begins a description of the Essential Elements (EEs), which are the learning targets for the DLM assessments. We also describe the development process for the EEs. Then we describe the development process of the DLM learning maps and how they were linked to the EEs. The EEs and the learning maps were developed concurrently, but separately, and then integrated. We then explore how the learning map is organized into claims and conceptual areas and describe the evaluation of the learning map structure. The chapter then describes the development of the DLM assessment blueprints defining EEs for assessment by grade and subject. Finally, the chapter concludes with a description of an external alignment study.

2.1. Essential Elements

EEs are specific statements of knowledge, skills, and understandings in a subject. The purpose of the EEs is to build a bridge from grade-level college and career readiness content standards to academic expectations for students with the most significant cognitive disabilities for both instruction and assessment. They are based on the grade-level general education content standards but are at reduced complexity, linking the grade-level general education content standards to academic expectations that are at an appropriate level of challenge for students with the most significant cognitive disabilities. In other words, EEs are the alternate content standards of the grade-level college and career readiness content standards used in general education assessments.

The progression of grade-level EEs across years of instruction reflects the changing priorities for instruction and learning as students move from grade to grade. The differences between EEs at different grade levels are subtler than what is typically seen in content standards for general education; the grade-to-grade differences in the EEs may consist of added knowledge, skills, and understandings that are not of obvious increasing rigor compared to the grade-to-grade differences found in the general education content standards. However, to the degree possible, the knowledge, skills, and understandings represented by the EEs increase in complexity across the grades, with clear links to the shifting emphases at each grade level in the general education content standards.

An example of three related EEs from the CCSS ELA “Key Ideas and Details” strand is shown in Table 2.1. The content shown is from elementary (Grade 3), middle (Grade 7), and high school (Grades 9–10). There is an increase in what students are asked to do as grade levels increase. Tested EEs for each grade are available on the DLM website for ELA[2] and mathematics[3].

[2] https://dynamiclearningmaps.org/essential-elements/ela

[3] https://dynamiclearningmaps.org/essential-elements/math

Table 2.1

Example of Increasing Complexity in Related Essential Elements Across Grades

Essential Element	Grade	Essential Element Description
ELA.EE.RI.3.2	Grade 3	Identify details in a text.
ELA.EE.RI.7.2	Grade 7	Determine two or more central ideas in a text.
ELA.EE.RI.9-10.2	Grades 9–10	Determine the central idea of the text and select details to support it.

The EEs specify grade-level learning targets, and the DLM learning maps clarify how students can reach those learning targets. For each EE, critical junctures on the path toward the EE's learning target(s) are identified by one or more nodes in the DLM learning maps. Nodes are also identified past the EE's learning target(s) to give students an opportunity to grow toward the learning targets of grade-level general education content standards.

These critical junctures of one or more related nodes are called linkage levels. The Target linkage level aligns to the EE's grade-level learning target(s). There are three linkage levels below the Target (i.e., Initial Precursor, Distal Precursor, and Proximal Precursor) and one linkage level beyond the Target (i.e., Successor). Table 2.2 shows the increasing complexity from linkage level to linkage level for the same EEs shown in Table 2.1. This example provides an illustration of how complexity increases both across linkage levels and across grade levels. The linkage levels are the unit of assessment for the DLM System.

Table 2.2

Example of Increasing Complexity of Skills in Related Linkage Levels for Three Essential Elements Across Grades

Linkage level	ELA.EE.RI.3.2	ELA.EE.RI.7.2	ELA.EE.RI.9-10.2
Initial Precursor	The student can demonstrate an understanding that absent objects still exist despite not being visible by searching for objects that are hidden or not visible.	When provided with a picture of an object, or other symbolic representation of that object, the student can correctly match the picture with the real object.	The student can identify concrete details, such as individuals, events, or ideas, in a familiar informational text.
Distal Precursor	When provided with language cues, the student can pay attention to the entire object, a characteristic of the object, or an action the object can perform.	After hearing or reading a beginner-level informational text, the student can identify a concrete detail in the text.	After reading or hearing an informational text, the student can identify the topic of the text and textual details that are related to the topic.

Table 2.2

Example of Increasing Complexity of Skills in Related Linkage Levels for Three Essential Elements Across Grades (continued)

Linkage level	ELA.EE.RI.3.2	ELA.EE.RI.7.2	ELA.EE.RI.9-10.2
Proximal Precursor	When provided with illustrations that are related and unrelated to a familiar text, the student can identify the illustrations that relate to aspects of the familiar text, such as people, places, things, and ideas.	After hearing or reading an informational text, the student can identify the implicit main idea of the text and identify the relationships between concrete details.	After reading or hearing an informational text, the student is able to summarize the information from the text.
Target	After hearing or reading a beginner-level informational text, the student can identify a concrete detail in the text.	After reading or hearing an informational text, the student can identify more than one main idea in the text.	After reading or hearing an informational text, the student can identify the central idea of the text and the details that contribute to the understanding of the central idea.
Successor	After hearing or reading an informational text, the student can identify explicit details that are key to the information in the text.	After reading or hearing an informational text, the student can demonstrate an understanding of the text by summary of the text by identifying an accurate summary or expressing the main ideas of the text.	The student can identify both the implicit and explicit meaning of an informational text by identifying specific details and citations within the text that support the meaning.

2.2. Development of the Essential Elements

As previously mentioned, the EEs and the learning maps were developed concurrently, but separately. The development of the EEs involved DLM project staff; Edvantia, Inc., a DLM subcontractor; the DLM Governance Board; and content experts and educators of students with significant cognitive disabilities who were recruited by the DLM Governance Board. Initial planning meetings were held February 2011 to ensure that governance board members were in agreement with the process designed by Edvantia and the goals of the EEs. At the initial meeting and throughout the development process, stakeholders and DLM staff prioritized EEs that increased in complexity across grades and reflected high academic expectations aligned to college and career readiness standards for students with significant cognitive disabilities.

Stakeholder meetings were held via webinar in March 2011 to prepare materials for development meetings.

A series of subject-specific webinars were conducted in April 2011 to train panelists before meeting face-to-face to draft the EEs in ELA and mathematics in April–May 2011.

Led by Edvantia, representatives from each of the then DLM partner state education agencies (SEAs) and the selected educators and content specialists developed the original draft of the DLM EEs. The first meeting was held in Kansas City, Missouri, in April 2011, to draft the ELA EEs from kindergarten through twelfth grade. More than 70 participants participated, representing 12 of the 13 states that were working to develop the DLM assessments at that time. A similar meeting was held to draft the mathematics EEs in May 2011, with more than 70 participants representing all 13 member states.

Drafts of the EEs developed at the meetings were compiled and released to participants for review and feedback. Panelists and other stakeholders took part in webinars from July through October 2011 to review drafts. The last drafts were reviewed by SEA and content experts in November 2011. The finalized version was released for state approval in February 2012 and, when approved, was released online in March 2012.

Additional revisions to the EEs were made during the concurrent development of the learning maps, as described in section 2.3.5. Final documents are available publicly on the DLM website[4] for ELA and mathematics.

2.3. Development of the Learning Maps

Learning maps are a type of cognitive model composed of multiple interconnected learning targets and other critical knowledge, skills, and understandings supporting student learning. The development of the DLM assessment system's learning maps began with a review of the existing literature on learning progressions, a widely accepted and similar approach to assessing student progress over time (Confrey et al., 2014, 2017; Shepard, 2018). Learning progressions depict student learning of the critical concepts related to a topic and depicted in grade-level content standards (Bechard et al., 2012) through multiple, sequenced building blocks that precede and support their mastery (Hess, 2012; Popham, 2011). These building blocks increase in complexity over time and provide different intermediary goals that students may target toward acquiring the learning targets (Bechard et al., 2012). Formative assessments often draw on learning progressions to assist educators in understanding the gap between current student performance and a learning target. Despite these benefits, learning progressions often represent only one individual, straightforward pathway depicting how the average general education student commonly learns the content.

Due to this feature, learning progressions may have difficulties in representing the learning of students with the most significant cognitive disabilities (Hess, 2012; Kearns et al., 2011). The singular and linear pathway of these learning progressions toward a grade-level content standard cannot accurately and sufficiently represent the complexity and diversity in the learning among all of these students (e.g., acquiring writing skills with limited mobility or learning to read with hearing impairments). Students with the most significant cognitive disabilities have a range of sensory differences that may require demonstrating their knowledge, skills, and understandings in different ways than those represented in learning progressions for students who do not use assistive technology.

DLM maps expanded on existing notions of learning progressions by accounting for the multiple pathways

4 https://dynamiclearningmaps.org/model

students may use to acquire the same learning targets. They also depict the hypothesized connections and interactions described in the research literature between conceptually related content standards within and across topics and grade levels. These changes formed the learning maps guiding the development of the DLM assessment. The resulting maps represent a web-like network of connected learning targets and the critical knowledge, skills, and understandings supporting them (Bechard et al., 2012). The supporting knowledge, skills, and understandings can depict intermediary learning targets advancing students toward acquiring individual content standards. To complement the progression of grade-level learning targets, the DLM learning maps also include the early cognitive skills developing between birth and school entry that form the basis for all subsequent content learning. Finally, the multiple pathways through the DLM learning maps provide accessibility for students with the most significant cognitive disabilities to the learning targets and the critical supporting knowledge, skills, and understandings (Erickson & Karvonen, 2014).

Project staff developed the DLM learning maps in ELA and mathematics, both of which begin with a common set of basic cognitive skills that provide a basis for academic skill development. The project staff consisted of individuals with expertise in cognitive psychology, literacy, and mathematics, as well as individuals with experience with students with significant cognitive disabilities, among other areas. To create these interconnected DLM learning maps, map developers followed a four-step process.

1. Identification and representation of learning targets
2. Identification and representation of critical supporting knowledge, skills, and understandings
3. Development of connections between nodes and building accessible pathways
4. Linking the learning maps to the EEs

In each step, the synthesized literature review informed development activities, which were followed by rounds of internal review to ensure the learning maps were consistent with the guidelines outlined within each step. These four steps are described in the subsequent sections.

2.3.1. Identification and Representation of Learning Targets

The first step in the development of the DLM learning maps was to identify learning targets. The DLM learning maps consist of two basic elements: nodes and connections. The nodes are essential, unique, observable, and testable knowledge, skills, and understandings. Nodes can either directly represent learning targets or may represent the critical knowledge, skills, and understandings supporting the acquisition of the learning targets. The second element, connections, forms the relationship between nodes.

Because the DLM assessments measure student achievement of the alternate grade-level expectations—the EEs, aligned to college and career-readiness content standards—the CCSS served as a starting place for node development. Specifically, grade-level CCSS became individual nodes within the DLM learning maps. The CCSS were initially used in early map development. The EEs were later integrated into the DLM learning maps as an additional set of learning targets that largely preceded the CCSS learning targets.

During initial map development, when a content standard contained multiple knowledge, skills, and understandings unsuitable to be combined into a single node, these different knowledge, skills, and understandings were represented as distinct nodes in the DLM learning maps. These nodes representing the grade-level learning targets are called learning target nodes. Once the learning target nodes had been created, they were arranged in the DLM learning maps according to grade-level(s).

2.3.2. Identification and Representation of Critical Supporting Knowledge, Skills, and Understandings

After identifying the learning targets, the next step in the development of the DLM learning maps focused on identifying and representing the critical knowledge, skills, and understandings supporting the acquisition of the learning targets and filling the gaps between the learning target nodes in the DLM learning maps.

The nodes representing these critical supporting knowledge, skills, and understandings are called supporting nodes. The results from a systematic literature review provided the primary input for creating these supporting nodes. Given that the CCSS for kindergarten begins at a relatively complex cognitive and language level, the map developers employed a bottom-up approach in the literature search, looking initially for research concerning early cognitive development (e.g., attending to object characteristics due to language cues) and then building toward the more advanced grade-level learning targets (e.g., answering *wh*-questions about details in a narrative). Wherever possible, the map developers used empirical research to drive the development and sequence of the supporting nodes. Table 2.3 provides an example of the alignment between a research-based learning progression and the supporting nodes based on them.

Table 2.3

Example Alignment Between Learning Progression in Research Literature and Learning Map Nodes

Stages of spelling development†	Nodes in the learning map
Precommunicative spelling	Spells words by including random letters
Semiphonetic spelling	Spells words by partially using letters to represents individual sounds
Phonetic spelling	Spells words by using letters to represent individual sounds
Transitional spelling	Spells words by using letters based on common letter patterns found in print
Conventional spelling	Spells words correctly using knowledge of letter-sound relationships and common spelling patterns

† See Gentry (1982) for details.

This step also involved making hypotheses and logical analyses about potential supporting nodes to fill in sections of the learning maps where the literature review provided no input. The map developers used common instructional practices, other curricular information, and expert judgment to provide ideas for the supporting knowledge, skills, and understandings that complete the gaps between learning target nodes. Regardless of their origin, each supporting node had a succinct name that summarizes the knowledge, skill, or understanding; an extended description that provides additional detail; and an observation that describes a context in which students could demonstrate their learning of the node's content.

Despite the DLM System's focus on students with the most significant cognitive disabilities, the empirical literature on the acquisition of academic skills by typical students primarily provided the basis for developing the learning maps in ELA and mathematics.[5] As a result, the map developers focused on first building a “super highway” to represent typical development with multiple pathways to learning targets. The map developers then adapted the learning maps by adding any additional pathways specific to students with significant cognitive disabilities as needed.

2.3.2.1. Critical Sources

Node development was based on a systemic literature review of articles, books, and book chapters summarizing the developmental research in a domain area. Book chapters and research syntheses broadly surveying the literature in a given domain were most useful to the map developers in developing the learning maps in ELA and mathematics. The grade-level content standards provided the parameters to guide the literature search. Node development began with the map developers identifying key terms within the content standards and locating relevant research handbooks or edited chapter books. These broad literature reviews were of the greatest utility because they often synthesized research findings into a developmental learning trajectory of the academic skills pertinent to the domain (e.g., Nippold, 2007; Sarama & Clements, 2009). Additionally, the map developers identified individual studies that were considered seminal to a particular domain, which could be used when building supporting nodes for a specific section of the DLM learning maps. If a particular researcher’s empirical work was sought out, the map developers looked for articles summarizing a series of findings into a developmental sequence (often using “acquisition” as a search term). The map developers also identified articles reporting the findings of longitudinal and cross-sectional samples that provide insight into the developmental acquisition of academic skills. When these sources were unavailable or did not cover the entire learning map area, the map developers synthesized the findings from multiple empirical studies to generate appropriate supporting nodes.

2.3.2.2. Nodes Reflect the Products of Learning and Cognitive Advancement

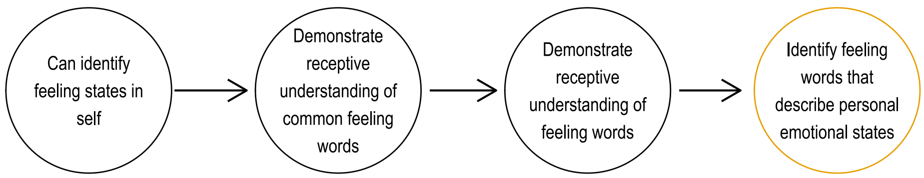
The learning maps depict learning between birth and school entry. The supporting nodes developing between birth and school entry reflect the learning and cognitive growth that occurs during this period by becoming increasingly more complex. For example, early developing cognitive skills, such as seeking the attention of others, provide the basis for more complex and later-developing ones, such as using words to request, comment, and command.

The supporting knowledge, skills, and understandings that develop after school entry provide steps that help students meet grade-level learning targets. Nodes throughout the learning maps reflect increased cognitive skills (e.g., improving logical and analytical thought and increasing declarative, procedural, and metacognitive knowledge), resources (e.g., increasing working memory span and attention), and instruction (e.g., learning basic abstract symbols following exposure and explicit instruction). Figure 2.2 represents an example of the gradually increasing complexity of supporting nodes in the DLM learning maps. These shifts in the knowledge, skills, and understandings depicted in the supporting nodes complete the framework established by the grade-level learning targets.

[5] Systematic literature reviews revealed a dearth of research related to academic skill development among students with the most significant cognitive disabilities.

Figure 2.2

Example Progressions of Supporting Nodes Toward the Target Node



Note. The target node for the Essential Element is the orange circle.

2.3.2.3. Foundational Nodes

Even in the early grades, learning targets associated with grade-level content standards require the application of basic cognitive skills. These basic cognitive skills are required across subjects and include such things as attention, self-regulation, language, and categorization. Foundational nodes represent these basic cognitive skills in ELA and mathematics. Some students with the most significant cognitive disabilities must be taught learning targets associated with foundational nodes to work toward learning targets associated with grade-level content standards (Kleinert et al., 2009).

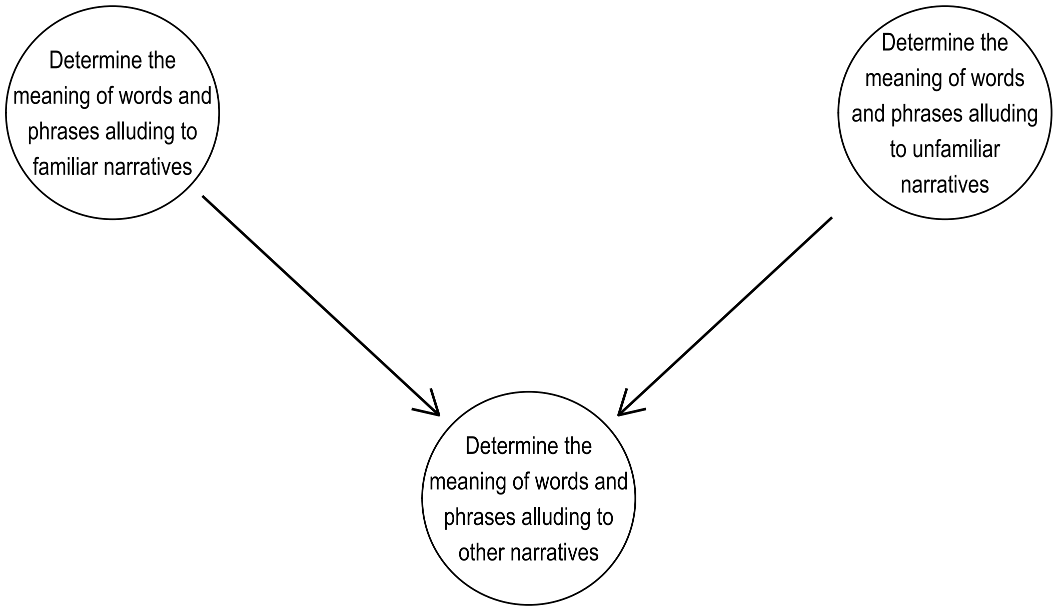
2.3.2.4. Node Development Criteria

The nodes representing the critical knowledge, skills, and understandings supporting the acquisition of grade-level learning targets met specific criteria before inclusion. These criteria focus on the content and accessibility of the nodes and ensure that the learning maps consisted of only the nodes that were critical to represent student content learning. The criterion were that the node should be (1) essential, (2) unique, (3) appropriately sized, (4) accessible, and (5) observable and measurable. The first criterion for node development was whether the nodes were essential for students to advance toward one or more grade-level learning targets in the DLM learning maps. The results of the systematic literature review and map developer judgment guided the process of determining whether to include the node.

The second criterion of node development required nodes to not duplicate the knowledge, skills, and understandings depicted in the surrounding nodes. In other words, the map developers created nodes distinct from other nodes by extending the knowledge, skills, and understandings covered in the preceding node(s) and contributing to those associated with the succeeding node(s). If a new node was not unique, the map developers combined its content with a current node in the DLM learning maps, as shown in Figure 2.3.

Figure 2.3

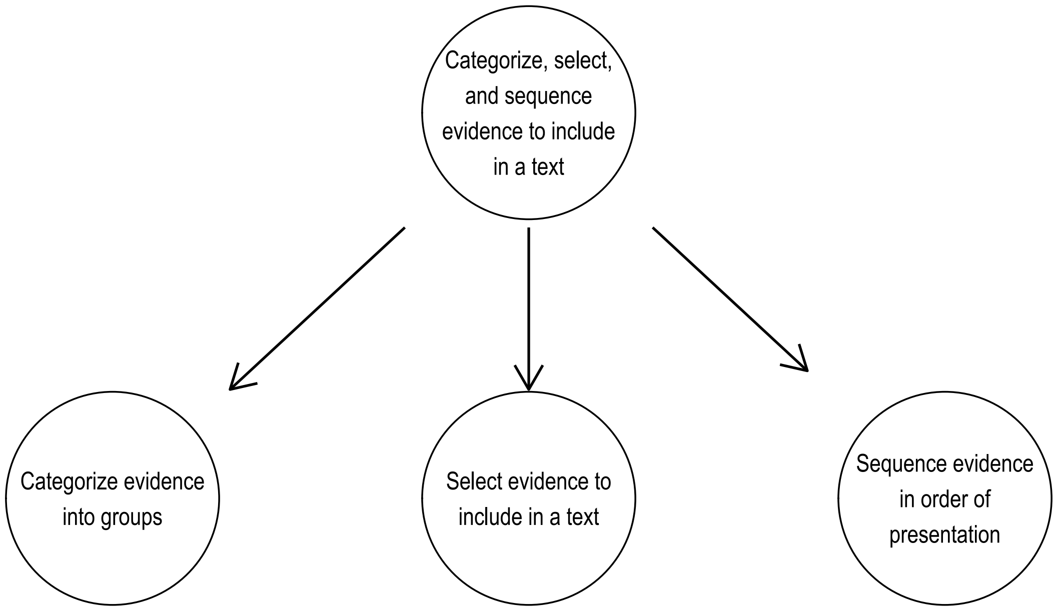
Combining Two Indistinct Nodes Into One Combined Node



The third criterion was that a node should represent only a single concept or skill in node development. This ensured that a new node was of a similar grain size to the surrounding nodes in the maps. The map developers divided nodes that were too large into multiple, more reasonably sized nodes (see Figure 2.4) or combined nodes that were too small into a single node. The characteristics of students with the most significant cognitive disabilities guided the identification of the appropriate node size.

Figure 2.4

Dividing One Node Into Multiple, Reasonably Sized Nodes



The fourth criterion of node development required the map developers to generate nodes accessible to students with the most significant cognitive disabilities. Accessible nodes reflect the principles of Universal Design for Learning (Center for Applied Special Technology [CAST], 2018) and are free of barriers for students with specific sensory, mobility, or communication disabilities. More specifically, they should account for variability among students, increasing the range of access to the node's content. When provided with the necessary support, all students, regardless of their disability, should have the opportunity to demonstrate their learning. Inaccessible nodes require students to demonstrate their understanding of the node's content through only one modality or format. The map developers made nodes more accessible by allowing multiple modalities or formats (see Table 2.4).

Table 2.4

Examples of the Accessible Node Criterion for Students With the Most Significant Cognitive Disabilities

Accessible nodes	Nonaccessible nodes
Introduce the topic or book being written about and then state an opinion on it when writing an opinion piece	Introduce the topic or book being written about and then state an opinion on it when writing an opinion piece using pencil and paper
Sort information related to a topic from print or digital sources into given categories	Sort visual information related to a topic from print or digital sources into given categories

The final criterion for development was that the nodes were observable and measurable. If nodes are observable and measurable, students should have the opportunity to demonstrate their learning of the node's content. These nodes require some form of expression that reflects the knowledge, skill, or understanding depicted in the node's content, allowing educators and test developers to gauge student mastery. Nodes lack observability and measurability when they only occur within the student's mind or require inferences from other behaviors.

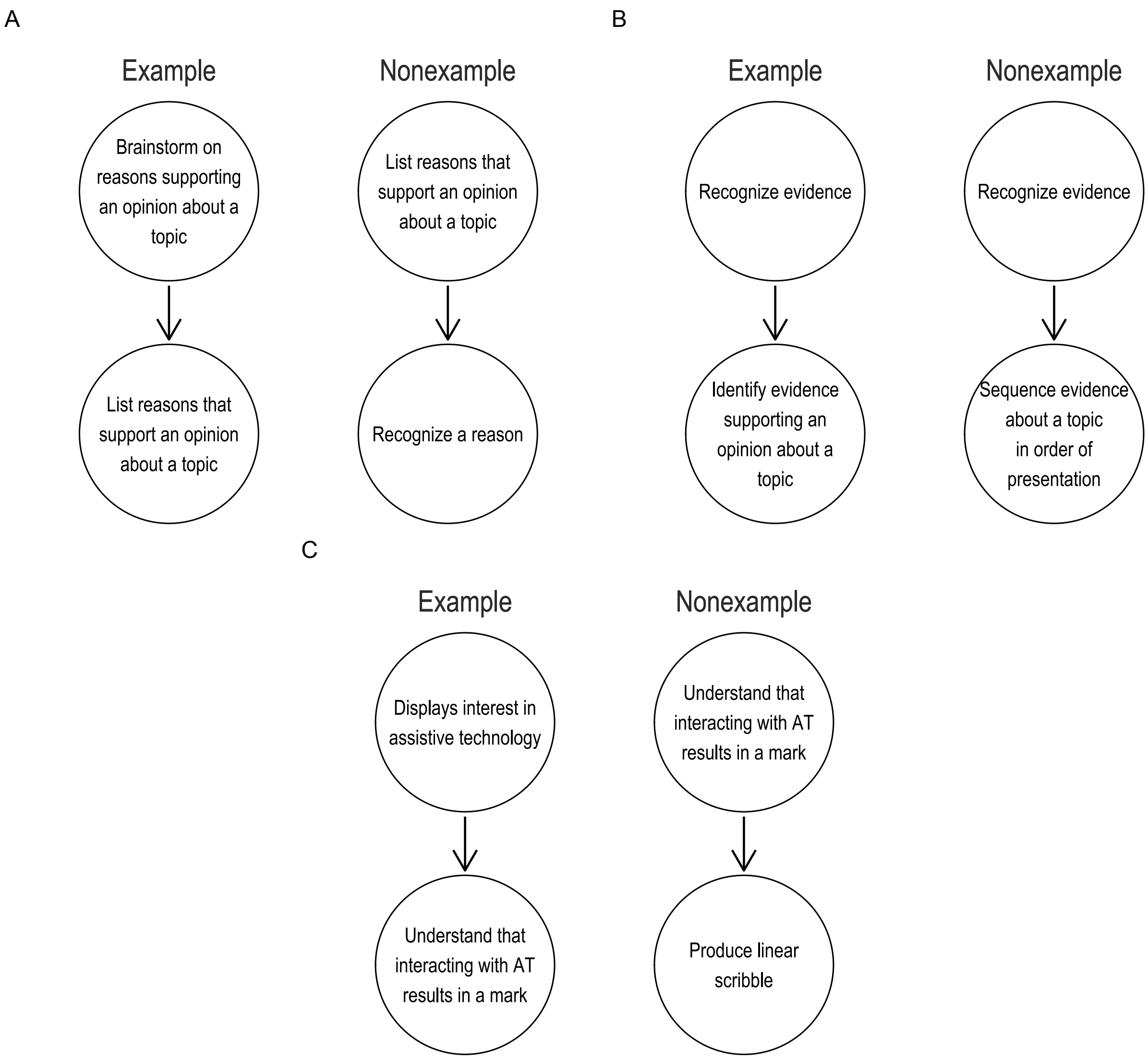
In summary, the learning maps contain only nodes meeting the above requirements, and each node contains critical supporting knowledge, skills, and understandings for the acquisition of the learning targets.

2.3.3. Development of Connections Between Nodes

After the learning target and supporting nodes were identified, they were arranged and connected according to their developmental acquisition, based on the empirical literature or in order of common instructional or curricular practices. An individual connection represents the directionality of the relationship between two nodes—the origin node and the destination node. For inclusion in the DLM learning maps, a connection should be logical, appropriate, and accessible. The map developers built logical origin-to-destination connections that represent increased cognitive progressions, where origin nodes are cognitively less complex than the destination nodes (see Figure 2.5A). Origin nodes precede and are hypothesized to develop before the destination nodes. Connections should not represent an overly large leap in complexity between the origin and destination nodes (see Figure 2.5B). The map developers built accessible origin-to-destination connections that do not contain any barriers limiting the access to the origin and destination nodes' content for students with the most significant cognitive disabilities (see Figure 2.5C).

Figure 2.5

Examples of Criteria for Making Connections Between Nodes



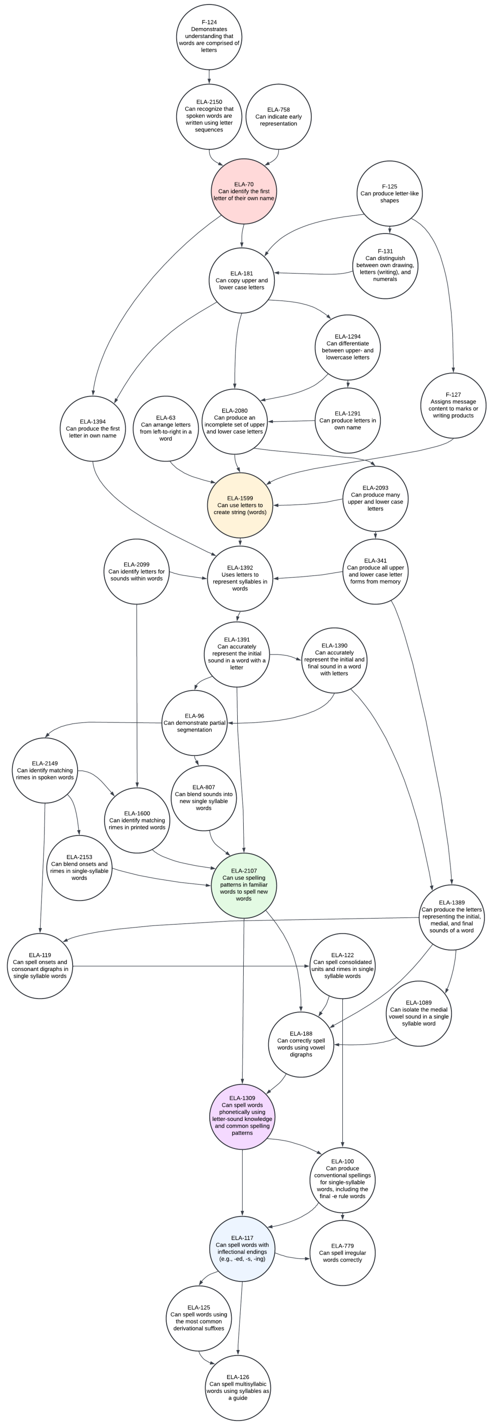
Note. Panel A: Example and Nonexample of a logical connection between two nodes. Panel B: Example and Nonexample of an appropriate connection between two nodes. Panel C: Example and Nonexample of an accessible connection between nodes.

As an example, small sections of the DLM learning map are provided in Figure 2.6 and Figure 2.7. This map section covers the nodes leading up to the spelling of untaught words phonetically. Figure 2.6 illustrates the structure of the DLM learning maps, including multiple pathways, while Figure 2.7 displays the nodes used for assessment. Both figures highlight a specific pathway to demonstrate the interconnected nature of the DLM learning maps. The pathway depicts how a student would progress from one node (i.e., can identify the first letter of their own name) to a later node (i.e., can spell words with inflectional endings). Items used in the DLM assessment system have been created to measure some of the nodes in this pathway, and these nodes have been color-coded to identify their location within the map section.

The colored nodes in both figures represent tested nodes, which are nodes that have been identified as making a significant contribution to the acquisition of the learning target by the map developers and content experts. These nodes typically precede or directly follow the learning target node (e.g., can spell words phonetically using letter-sound knowledge and common spelling patterns), but they are not the only nodes contributing to the acquisition of the learning target, nor do they prescribe the only route that can be taken toward acquiring it.

Figure 2.6

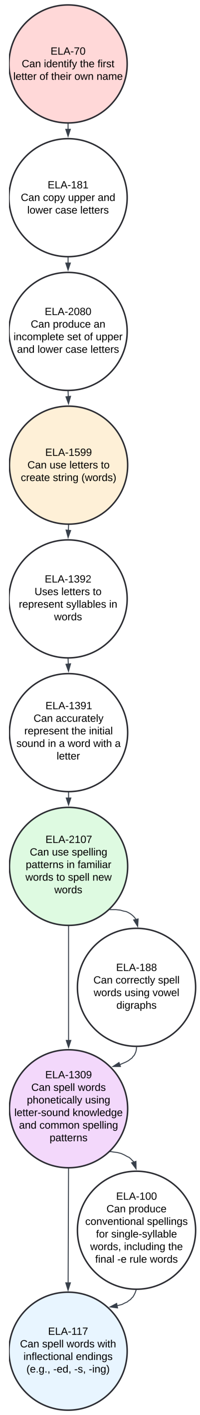
Section of the Learning Map Representing Basic Cognitive and Academic Skills



Note. The color-coded nodes in the map section represent a pathway of tested nodes, with the learning target for this Essential Element highlighted in purple.

Figure 2.7

Pathway of Nodes Covering Essential Element ELA.EE.L.6.2.b



Note. The color-coded nodes in the map section represent a pathway of tested nodes, with the learning target for this Essential Element highlighted in purple.

2.3.4. Accessibility of Nodes and Pathways

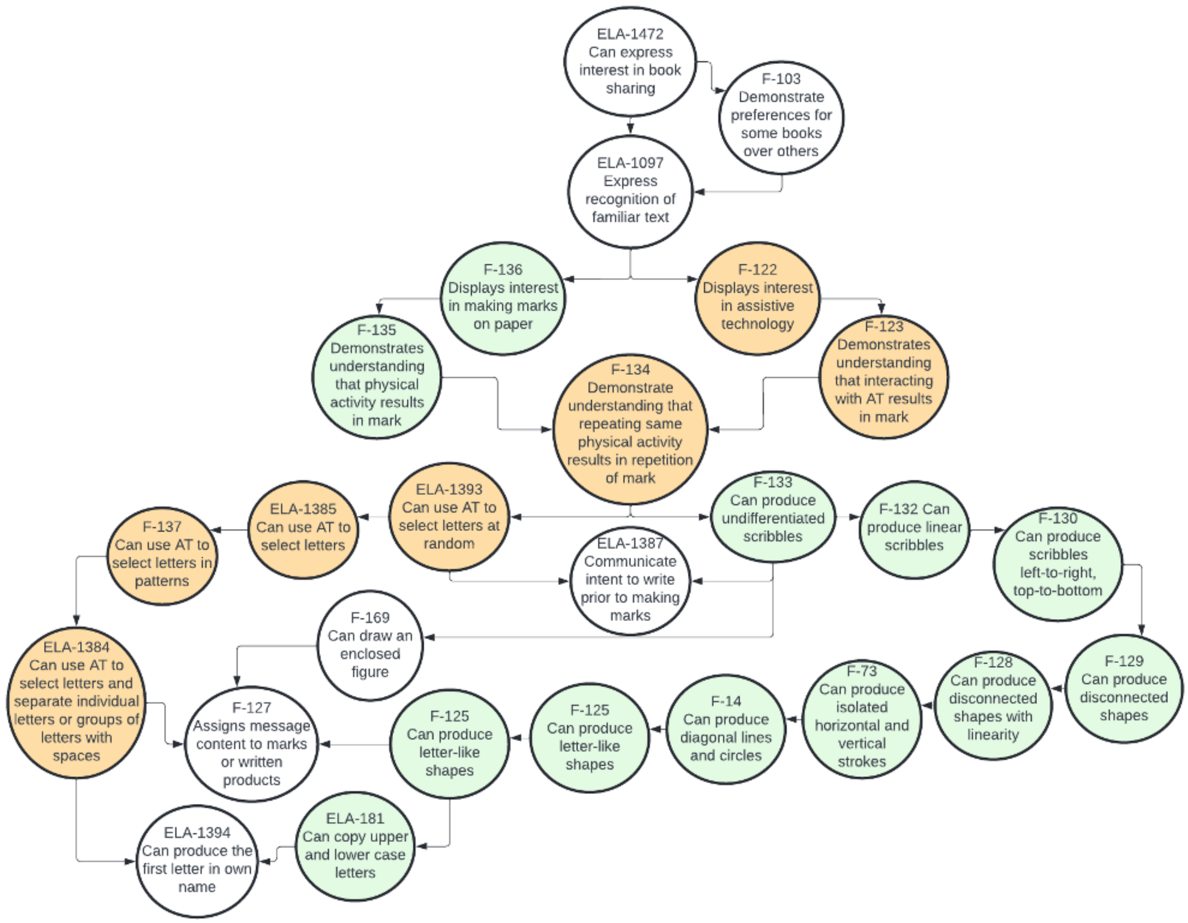
Creating appropriate learning targets alone does not sufficiently provide all students with the most significant cognitive disabilities access to the content in the DLM learning maps. Some students exhibit sensory, mobility, or communication disabilities that require different means for representation and action or expression to provide evidence of mastery for some nodes in the DLM learning maps. A critical step in making the DLM learning maps accessible to all students involves verifying that nodes and pathways are appropriate for all students and their sensory, mobility, and communication needs. These pathways allow all students to achieve grade-level learning targets when provided with appropriate support and access to instruction and assessments based on the principles of Universal Design for Learning (CAST, 2018).

The map developers, in partnership with the Center for Literacy and Disability Studies (CLDS) at the University of North Carolina, enhanced accessibility of the DLM learning maps for students with specific disabilities. The CLDS team reviewed each node and considered whether the node was accessible to individuals with learning differences across four primary areas: vision, hearing, mobility, and communication. Nodes flagged during this process were most likely inaccessible to students with specific sensory, mobility, or communication disabilities even when considering potential accommodations. For example, many early writing nodes involve skills like scribbling before students can produce letters and numerals. For students with mobility differences, the writing acquisition process will include learning to use assistive technology to select letters and numbers. In this example, an accommodation allowing the students to choose scribbles would be inappropriate. As a result, the CLDS team flagged as inaccessible the early writing nodes related to scribbling because the cognitive process of learning to write involves some fundamental differences for students using assistive technology to communicate. Clusters of these flagged nodes represented a section within the DLM learning maps that posed challenges for students with specific sensory, mobility, or communication disabilities.

For example, in Figure 2.8, students with mobility impairments would demonstrate their writing knowledge and skills in different ways than students who do not use assistive technology. In this example, the learning target may be ELA-1394, “Can produce the first letter in their own name.” The green pathway describes how students who do not use assistive technology might acquire the knowledge and skills for this target (e.g., drawing scribbles, diagonal lines, circles). Conversely, students with mobility impairments may instead acquire the skills to use assistive technology, such as selecting using the technology to select letters first at random and then in a pattern, as shown in the orange pathway. These nodes represent the cognitive steps involved with learning to use writing methods different from those used by students without mobility impairments.

Figure 2.8

Hypothesized Pathways in Writing for Students With and Without Mobility Disabilities



Note. Green shading indicates the writing development of mobility-typical students. Orange shading suggests an accessible path for students with mobility impairments using assistive technology.

2.3.5. Linking the Learning Maps to the Essential Elements

Because the primary goal of DLM assessments is to measure what students with the most significant cognitive disabilities know and can do, EEs were created to reflect more accurately the knowledge, skills, and understandings that are appropriately challenging grade-level learning targets for the students. Within each subject, the EEs were derived from CCSS to represent similar academic skill development sequences as the CCSS.

EEs were first written based on the CCSS, independent of the learning maps development process in 2012 (see section 2.2 for a description of the EE development). At the same time that the EEs were being developed, map developers were actively engaged in building the maps in ELA and mathematics. Because the development of the EEs and the learning maps happened simultaneously, alignment between the EEs and the learning maps was not possible until the fall of 2012. The process of evaluating the alignment between the EEs and the DLM learning maps involved reconciling the content of the EEs to the content represented in the nodes and connections of the DLM learning maps in ELA and mathematics.

Teams of content experts worked together to revise the initial 2012 version of the EEs and the DLM maps to ensure appropriate alignment of these two elements of the assessment system. Alignment involved horizontal alignment of the EEs with the CCSS and vertical alignment of the EEs with meaningful progressions of skills represented by nodes in the DLM maps. The process of aligning the maps and the EEs began by identifying nodes in the maps that represented the EEs in mathematics and ELA. This process revealed areas in the maps where additional nodes were needed to account for incremental increases in expectations across related EEs from one grade to the next. Areas were also identified in which an EE was out of place developmentally with other EEs in the same or adjacent grades according to research that was incorporated into the maps. For example, adjustments were made when an EE related to a higher-grade map node appeared earlier on the map than an EE related to a lower-grade map node (e.g., a fifth-grade skill preceded a third-grade skill). Finally, the alignment process revealed EEs that were actually written as instructional tasks rather than learning outcomes. These EEs were revised to represent knowledge and skills rather than instructional tasks. These revisions were compiled and reviewed by the governance board in early 2013, with an approved final version of the EEs published in May 2013. Final documents are available publicly on the DLM website[6] for ELA and mathematics.

This process of aligning the EEs and learning maps also resulted in significant revisions to the DLM learning maps to ensure that the nodes and connections represented a solid framework from which assessments could be developed. Depending on the complexity of the EE, one or more nodes in the DLM learning maps were aligned to the EE as learning targets. If no existing node(s) corresponded to the content of the EE, new nodes were created and placed in the DLM learning maps at appropriate locations according to their content. New nodes were placed in the DLM learning maps by analyzing the existing map structure to identify precursor and successor nodes to the new node. Once identified, the map developers proposed placements of new nodes and connections based on literature reviews and expert judgment.

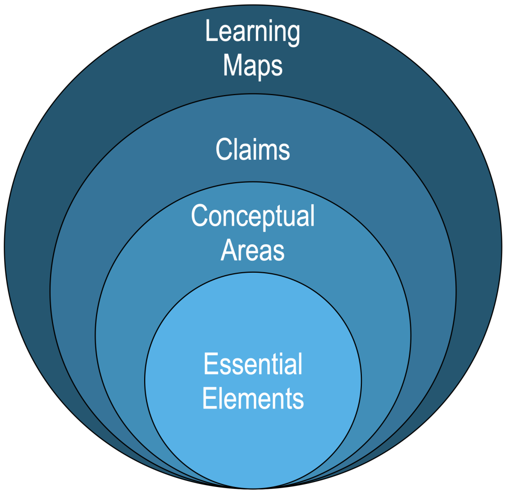
2.4. Organizing the Learning Maps: Claims and Conceptual Areas

Large sections of the DLM learning maps are too complex to depict in a manageable map view or describe on a node-by-node basis. Instead, the larger sections are described by the claims and conceptual areas they represent. This organizational structure was designed to articulate where the content standards are located and their relationships to important cognitive concepts. Organization of the academic content in the DLM assessment system is illustrated conceptually in three layers (claims, conceptual areas, and EEs), as shown in Figure 2.9.

[6] https://dynamiclearningmaps.org/model

Figure 2.9

Layers of Content in the DLM Alternate Assessment System



Modern test development approaches, such as evidence-centered design (Mislevy & Riconscente, 2005; Mislevy et al., 2003), are founded on the idea that test design should start with specific claims about what students know and can do and the evidence needed to support such claims. While evidence-centered design is multifaceted, it starts with a set of claims regarding the major knowledge, skills, and understandings in the domains of interest (i.e., ELA and mathematics), as well as how it is acquired.

The DLM System divides both ELA and mathematics content into four broad claims, which are subdivided into nine conceptual areas for each subject. Claims are overt statements of what students with the most significant cognitive disabilities are expected to learn and be able to demonstrate when mastering the knowledge, skills, and understandings within a broad section of the DLM learning maps. As broad statements about expected student content learning, claims focus the scope of the assessment. Each claim includes two or three conceptual areas.

Conceptual areas are clusters of related concepts made up of multiple conceptually associated EEs and nodes that support, represent, and extend beyond the EE's learning target(s). Conceptual areas further define the knowledge, skills, and understandings required to meet the broader claims and serve as models of how students may acquire and organize these knowledge, skills, and understandings in a subject. The DLM claims and conceptual areas apply to all grades in the system.

The four claims and nine conceptual areas for each subject are shown in Table 2.5. Following Table 2.5, Figure 2.10 provides an example of a conceptual area in the learning map.

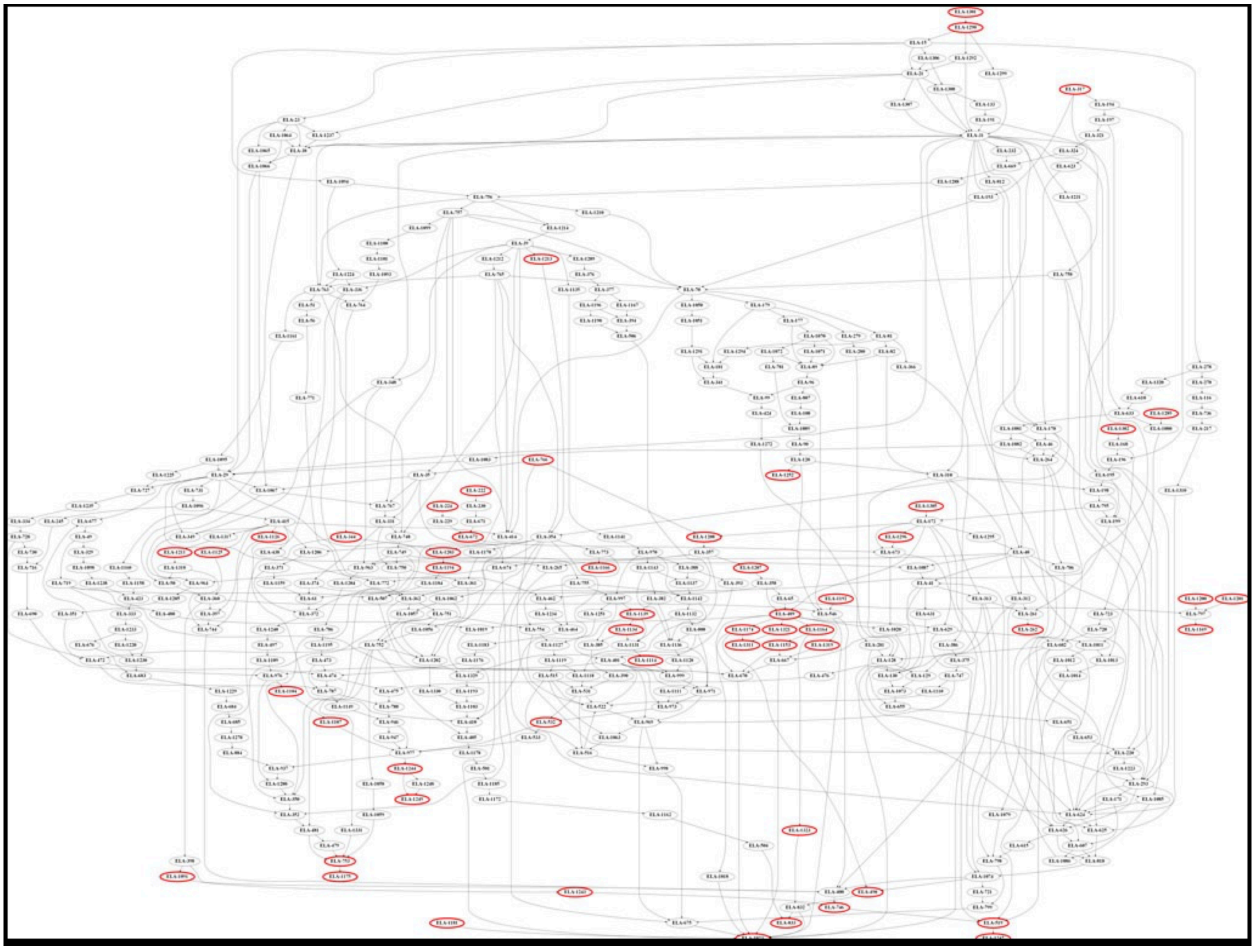
Table 2.5

DLM Claims and Conceptual Areas

Claim	Conceptual area
ELA.C1: Students can comprehend text in increasingly complex ways.	ELA.C1.1: Determine Critical Elements of Text ELA.C1.2: Construct Understandings of Text ELA.C1.3: Integrate Ideas and Information From Text
ELA.C2: Students can produce writing for a range of purposes and audiences.	ELA.C2.1: Use Writing to Communicate ELA.C2.2: Integrate Ideas and Information in Writing
ELA.C3: Students can communicate for a range of purposes and audiences.	ELA.C3.1: Use Language to Communicate With Others ELA.C3.2: Clarify and Contribute in Discussion
ELA.C4: Students can engage in research/inquiry to investigate topics and present information.	ELA.C4.1: Use Sources and Information ELA.C4.2: Collaborate and Present Ideas
M.C1: Number Sense: Students demonstrate increasingly complex understandings of number sense.	M.C1.1: Understand Number Structures (Counting, Place Value, Fractions) M.C1.2: Compare, Compose, and Decompose Numbers and Sets M.C1.3: Calculate Accurately and Efficiently Using Simple Arithmetic Operations
M.C2: Geometry: Students demonstrate increasingly complex spatial reasoning and understanding of geometric principles.	M.C2.1: Understand and Use Geometric Properties of Two- and Three-Dimensional Shapes M.C2.2: Solve Problems Involving Area, Perimeter, and Volume
M.C3: Measurement Data and Analysis: Students demonstrate increasingly complex understanding of measurement, data, and analytic procedures.	M.C3.1: Understand and Use Measurement Principles and Units of Measure M.C3.2: Represent and Interpret Data Displays
M.C4: Algebraic and Functional Reasoning: Students solve increasingly complex mathematical problems, making productive use of algebra and functions.	M.C4.1: Use Operations and Models to Solve Problems M.C4.2: Understand Patterns and Functional Thinking

Figure 2.10

Section of the English Language Arts Learning Map for Conceptual Area ELA.C1.2

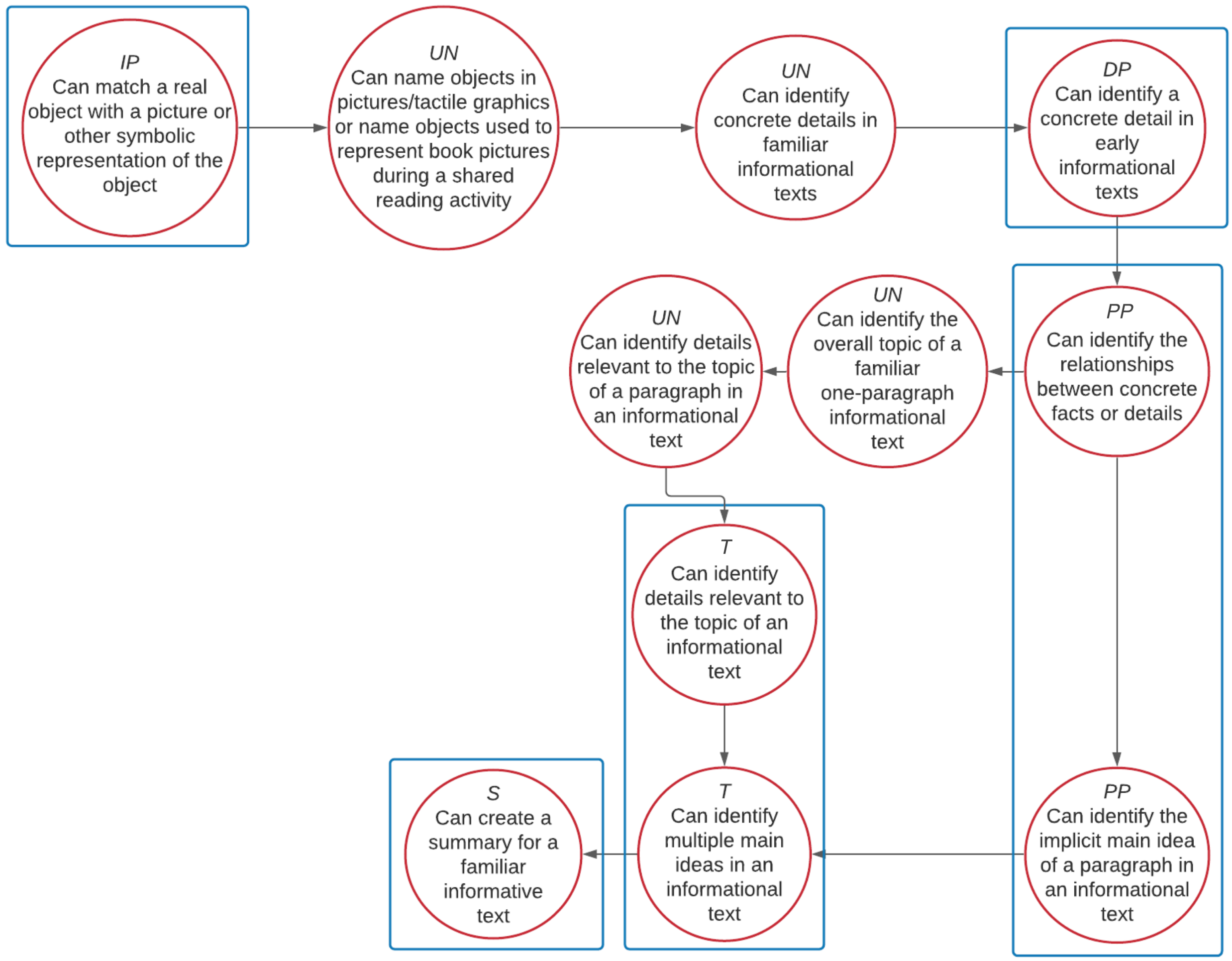


Note. Red circles indicate nodes aligned to Essential Elements.

Within each claim and conceptual area, smaller regions of the DLM learning maps are displayed in mini-maps for single EEs, as shown in Figure 2.11. Mini-maps identify which nodes are assessed at each linkage level for an EE, as well as untested nodes between the assessed nodes to support educators' instructional decisions.

Figure 2.11

Example Mini-Map for an English Language Arts Essential Element



Note. Blue boxes indicate nodes organized into linkage levels. IP = Initial Precursor, DP = Distal Precursor, PP = Proximal Precursor, T = Target, S = Successor, UN = Untested node.

In summary, the DLM learning maps represent the paths students with the most significant cognitive disabilities may take to acquire the knowledge, skills, and understandings within a subject, claim, or conceptual area. EEs (the DLM learning targets) within a particular claim or conceptual area link to one another. Linkage levels are small collections of nodes which represent critical junctures on the path toward and beyond the learning target. The Target linkage level reflects the grade-level expectation aligned directly to the EE. The DLM claims and conceptual areas provide a framework for organizing nodes on the DLM learning maps and, accordingly, the EEs.

2.5. System Structure

Table 2.6 shows examples of the relationship between DLM system elements in ELA and mathematics. Assessment system elements are listed from broadest to most specific (i.e., from claim to learning map node).

Table 2.6

Assessment System Elements With Examples for English Language Arts and Mathematics

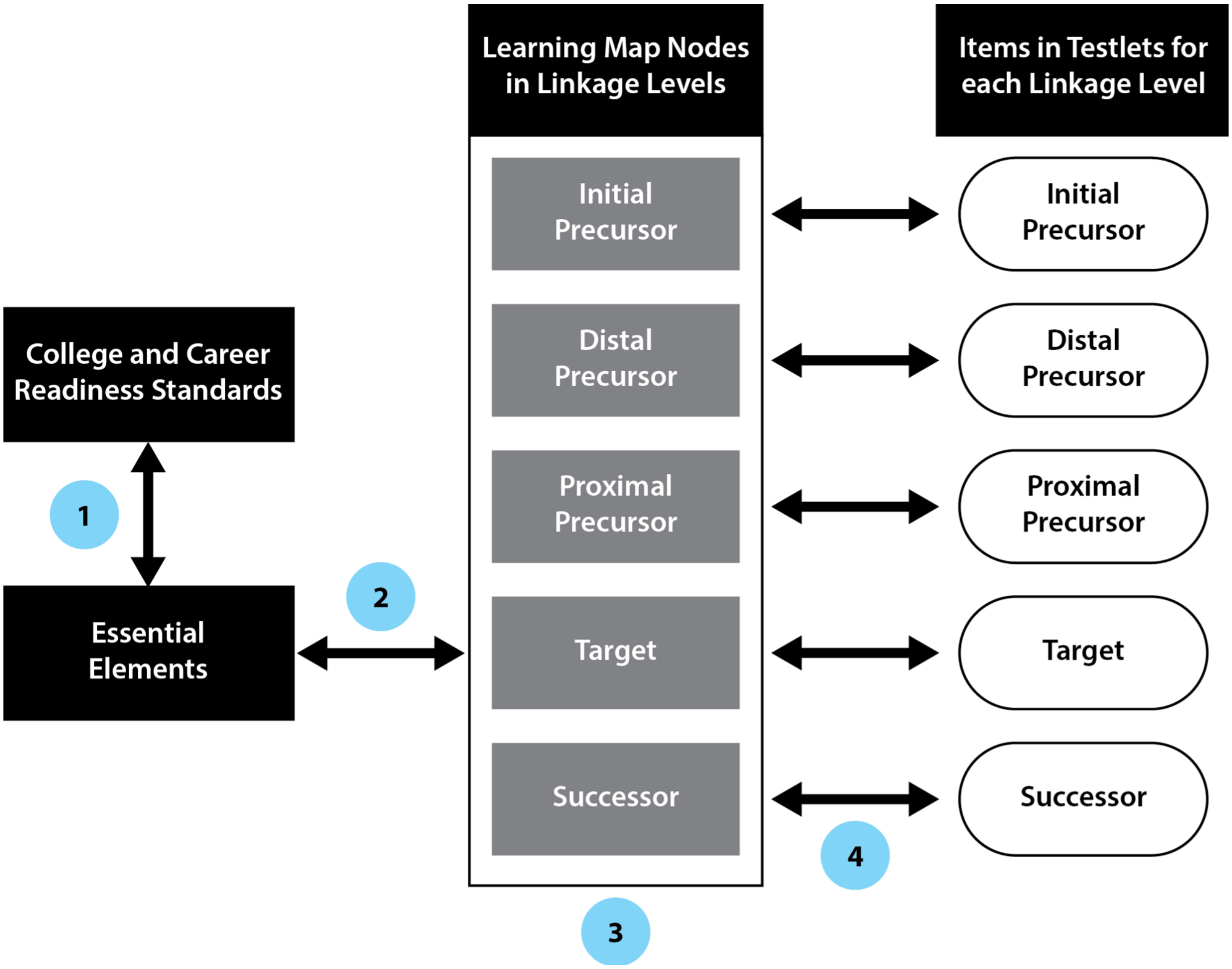
Element	Description
English language arts	
Claim ELA.C1	Student can comprehend text in increasingly complex ways.
Conceptual area ELA.C1.1	Determine Critical Elements of Text
Essential Element ELA.EE.RL.3.1	Answer who and what questions to demonstrate understanding of details in a text.
Target linkage level	The student can answer who and what questions about details in a story.
DLM learning map node	Can answer who and what questions about details in a narrative.
Mathematics	
Claim M.C1	Number Sense: Students demonstrate increasingly complex understanding of number sense.
Conceptual area M.C1.3	Calculate Accurately and Efficiently Using Simple Arithmetic Operations
Essential Element M.EE.6.NS.2	Apply the concept of fair share and equal shares to divide.
Target linkage level	Demonstrate understanding of division by splitting a set into an equal number of subsets and communicating the quotient as the number of equal subsets (e.g., a set consisting of 15 objects has three subsets, each containing 5 objects).
DLM learning map node	Demonstrate the concept of division.

The overall structure of the DLM System has four key relationships between system elements (see Figure 2.12; the numbers below are indicated in Figure 2.12):

1. College and career readiness content standards and EEs for each grade level
2. An EE and its target-level node(s)
3. An EE and its associated linkage levels
4. DLM learning map nodes within a linkage level and assessment items

Figure 2.12

Relationships in the DLM Alternate Assessment System



2.6. Evaluation of the Learning Map Structure

Once developed, the first evaluation of the DLM learning maps consisted of educator and expert review. Subsequent empirical analyses of the structure of the DLM learning maps have also been conducted. Each of these evaluations are described in turn.

2.6.1. Educator and Expert Review

By 2014, the DLM learning maps underwent three major external reviews by educators and experts. The first two review panels (K–5 and 6–12) leveraged the content expertise of general educators in ELA and mathematics, identified by state education agency (SEA) personnel from members of the DLM Governance Board, to examine the nodes and connections in the DLM learning maps by grade level. For each node, the external content reviewers considered

(a) the appropriateness of cognitive complexity,

(b) the relationship to the CCSS, and
(c) the properties of the node (e.g., grain size and redundancy).

The external content reviewers then reviewed individual origin-to-destination connections for appropriateness (e.g., is the connection from skill A to skill B logical?). If the external content reviewers identified a node or connection they disagreed with, found illogical, or contained a gap, they stated the reason for their decision and attempted to provide evidence supporting it. The external content reviewers then offered potential solutions for improving the problematic node or connection, such as adding a new or current node between the two nodes in a connection. During these reviews of the DLM learning maps, the external content reviewers focused on only the typical progression of the average student in acquiring the grade-level learning targets. Following the K–5 and 6–12 external content reviews, the map developers revised the DLM learning maps to incorporate the external content reviewers' feedback.

Additionally, as described in section 2.3.4 of this chapter, collaborators from CLDS at the University of North Carolina identified multiple sections in the DLM learning maps in which students with specific sensory, mobility, or communication disabilities might have difficulty demonstrating the node's content. SEA personnel from the participating states identified accessibility experts across various disabilities to participate in an external accessibility review. Special educators and related service providers reviewed specific sections of the DLM learning maps to make their content and structure accessible to all students with the most significant cognitive disabilities, regardless of any sensory, mobility, or communication disabilities. The external accessibility reviewers also evaluated areas flagged by CLDS and recommended pathways to make them more accessible.

In some cases, the implementation of the principles of Universal Design for Learning (e.g., flexibility and equitability of use, CAST, 2018) could make the node's content accessible by altering the manner of their assessment (i.e., allowing for multiple ways to demonstrate skills). When possible, the resulting nodes did not depend on information exclusively available through only one sense. In other cases, some students need to acquire unique cognitive skills to achieve a learning target (see the writing example provided in section 2.3.4). The external accessibility reviewers then identified additional pathways to support multiple means of instruction and assessment that were accessible to all students. In summary, the external accessibility reviewers proposed node and connection revisions to increase the accessibility of the DLM learning maps' content and structure when considering assistive technology for students with specific sensory, mobility, or communication disabilities.

2.6.2. Empirical Analyses of the Learning Maps

Although items on the DLM assessments are written to measure specific nodes,[7] the linkage level is the unit of analysis for the diagnostic classification model (DCM) used to score the assessment and the unit of reporting for student score reports.[8] Because of this test design, students typically complete 3–5 items for each assessed linkage level, but may complete as few as 0 or 1 item for each node (i.e., not all nodes that comprise a linkage level are measured on every testlet). Therefore, a direct evaluation of the learning map structure is not possible with the available data, as the DCM cannot be identified at the node level (Fang

[7] For a description of item writing practices, see Chapter 3 of this manual.

[8] See Chapter 5 and Chapter 7 of this manual for descriptions of the psychometric model used to score DLM assessments and score reporting, respectively.

et al., 2019; Xu, 2019; Xu & Zhang, 2016).

However, it is possible to evaluate the learning map structures by using the linkage levels as a proxy for the underlying maps. As described above, the linkage levels represent clusters of nodes, which are the learning targets for each EE. The five linkage levels are hypothesized to follow a linear hierarchy (i.e., more advanced linkage levels cannot be mastered if the student has not also mastered the prerequisite skills). Thus, we can begin to evaluate the learning map structure by evaluating the linear ordering of the linkage levels.

Consistent with the scoring model for DLM assessments, the linkage level ordering is evaluated using DCMs. The framework for evaluating a learning map structure using DCMs is detailed by W. J. Thompson and Nash (2022). The framework consists of three methods that each evaluate a proposed structuring of attributes (i.e., linkage levels):

1. **Patterns of Mastery Profiles.** The first method is the most direct test of an attribute structure. In this method, two DCMs are estimated: the log-linear cognitive diagnostic model (LCDM, Henson et al., 2009) and a hierarchical DCM (HDCM, Templin & Bradshaw, 2014). The LCDM allows students to have any profile of attribute (i.e., linkage level) mastery. In the HDCM, the model is constrained to only allow profiles of mastery that conform with the hierarchical structure (i.e., the five ordered DLM linkage levels). The two models can then be evaluated for model fit and compared directly using relative fit indices such as the widely applicable information criterion (WAIC, Watanabe, 2010) or Pareto-smoothed importance sampling leave-one-out cross validation (PSIS-LOO, Vehtari et al., 2017, 2022). EEs are flagged if the HDCM fits significantly worse than the LCDM, which indicates that the subset of mastery profiles included in the HDCM are insufficient for representing the range of mastery profiles demonstrated by students.
2. **Patterns of Attribute Mastery.** In this method, each attribute is estimated and scored separately as a 1-attribute LCDM (i.e., master or nonmaster of each individual linkage level). The patterns of mastery are then compared to the patterns that are expected, given the hypothesized structure (i.e., a student should not be able to master the Target level without also mastering the Proximal Precursor level). Any EE where more than 25% of students exhibit an unexpected pattern of attribute mastery is flagged for possible violations of the hierarchical linkage level ordering.
3. **Patterns of Attribute Difficulty.** In the final method, students are placed into cohorts based on their complexity band.[9] We then calculate the percentage of items answered correctly (*p*-value) by each cohort for each linkage level. If the hierarchical structure is correct, the *p*-values for any one cohort should decrease as the linkage level increases (i.e., the items should get harder as the linkage level and underlying nodes become more complex). A standard error is calculated for each *p*-value, and 95% confidence interval is created based on the standard error. An EE is flagged if the *p*-value for a higher linkage level is higher (i.e., items are easier) than for the lower linkage level, and the confidence intervals do not overlap. That is, the *p*-value is in the opposite direction of our expectation, beyond what might be expected due to noise in the data.

As with any analysis evaluating the statistical relationship between variables, the methods outlined in the W. J. Thompson and Nash (2022) framework require data on each variable from the same participant. For the purposes of evaluating linkage level ordering, this means students must be assessed on multiple

[9] See Chapter 4 for a description of how complexity bands are determined.

linkage levels for the same EE. Under the current assessment administration design, there are limited opportunities for collecting these cross-linkage-level data. These data can only be collected from students in Instructionally Embedded model states for whom educators choose to administer the same EE at multiple linkage levels, students in Year-End model states who participate in the optional instructionally embedded assessment window, or through targeted field testing. For example, in 2017–2018 and 2018–2019, rather than administering new content during the spring field-test window, operational forms were administered at a linkage level adjacent to the level that was tested during the operational assessment. This was done for the expressed purpose of collecting additional cross-linkage-level data to support an evaluation of the linkage level ordering. Although the targeted field test was limited to a single EE in each grade, the combined data from all sources have allowed for preliminary empirical evaluations of linkage level ordering.

Initial findings supporting the linkage level order are described in full by W. J. Thompson and Nash (2019). The patterns of mastery profiles method had limited results due to the limited availability of cross-linkage-level data. In the patterns of attribute mastery method, over 80% of EEs received no flag, indicating that the patterns of attribute mastery were consistent with the hierarchical ordering of linkage levels. Similarly, the patterns of attribute difficulty method also showed over 80% of EEs supporting the hierarchical structure of linkage levels. For a complete description of results, see W. J. Thompson and Nash (2019). Notably, most flags from the second and third methods were due to potential reversals among the higher linkage levels. This is expected, as the skills measured by the higher linkage levels are closer conceptually than the lower linkage levels. For example, in Figure 2.7, there are gaps of untested nodes between the earlier linkage levels, whereas the later tested nodes are consecutive. Therefore, it is more likely that we might observe a reversal in this situation. Although the initial findings show positive support for the linkage level ordering, additional research is planned to refine the methods and flagging criteria for identifying potential violations of the hierarchical structure. As the method are further refined, flagged EEs will undergo qualitative reviews by the test development and learning maps teams to inform potential revisions to the map structure. Updated analyses will be included in future technical manual updates as additional cross-linkage-level analyses and reviews are conducted.

2.7. Development of Assessment Blueprints

The DLM assessment blueprints specify the EEs, organized by conceptual area, on which students are assessed in each grade and subject.

Blueprint development began with a proposed plan in October 2013. The test development teams in each subject area developed blueprint options following several guiding principles. State representatives and subject matter experts then reviewed multiple iterations of blueprints, as did the senior DLM staff and psychometricians, through September 2014. The governance board adopted finalized blueprints for the spring 2015 administration.

The assessment blueprints were subsequently revised in 2019–2020, with additional feedback from the governance board and the DLM Technical Advisory Committee (TAC). In the following sections, we describe the initial development of the assessment blueprints, followed by the revision process.

2.7.1. Original Blueprint Development Process

The DLM Governance Board identified three overarching needs when developing the original blueprints: the blueprints should have broad coverage of academic content, emphasize connections across grades, and limit administrative burden. The learning maps described above were used to prioritize EEs for inclusion in the blueprint in each subject. EEs were evaluated by determining the position within the maps of EE-aligned nodes. EEs selected for inclusion in the blueprint had the potential to maximize student progress in academic skills across grades. The general principles that guided the use of the DLM maps to develop the blueprints were to:

- prioritize interrelated content to allow for opportunities to learn ELA and mathematics skills and conceptual understandings within and across grades,
- use knowledge of academic content and instructional methods to prioritize content considered important by stakeholders,
- maximize the breadth of content coverage of EEs within each grade and subject,
- balance a need for representativeness across grades with the need to prioritize a narrower range of interconnected content to allow students the opportunity to demonstrate growth within and across grade levels, and
- select an appropriate number of EEs in a grade to prevent excessive time for administration of an assessment to students with significant cognitive disabilities.

In both subjects, some EEs were not included on the blueprint. Reasons for excluding EEs from the blueprint included:

- the EE would be very difficult to assess in a standardized, computer-based assessment,
- the EE content relied on specific sensory information (e.g., an EE that was excluded because it would likely provide a barrier to access for students with visual impairments is ELA.EE.RL.3.7, "Use information gained from visual elements and words in the text to answer explicit *who* and *what* questions."),[10] and
- the EE content was more aligned to instructional goals (e.g., demonstrating understanding of text while engaged in group reading of stories) than to an assessment.

These principles were applied when making decisions about the EEs that were included in the blueprint. For instance, in ELA, the decision was made, in consultation with the governance board, to only assess Claim 1 (reading) and Claim 2 (writing). It is important to recognize that these principles were not implemented as rigid rules, but as guidelines for prioritization of the content of the EEs within and across the grades.

Test development teams for ELA and mathematics produced initial blueprint drafts by conducting a substantive review of each EE in conjunction with the location of the EE within the learning maps. The processes for ELA and mathematics differed slightly given the structural differences in the way the EEs were grouped thematically,[11] but adhered to these basic steps:

[10] In this case, a different EE in the same grade, describing a similar construct, ELA.EE.RL.3.1, "Answer *who* and *what* questions to demonstrate understanding of details in a text," was included on the blueprint, as it did not require specific attention to visual elements.

[11] These structural differences in groupings refer to the use of strands in ELA and clusters in mathematics. These elements were used in the CCSS and maintained in the EEs.

1. Review the content of the EE and its relationship to the associated grade-level content standard.
2. Review the location of the node(s) associated with the Target content of the EE in the maps.
3. Review the location of the node(s) associated with the Proximal, Distal, and Initial Precursors for each EE.
4. Review the location of the node(s) associated with the Successor for each EE.
5. Examine the relative location in the maps of all linkage levels associated with the EE to the location of related EEs in the preceding grade.
6. Examine the relative location in the maps of the contents of the EE to the location of related EEs in the following grade.
7. Using the map locations, prioritize EEs that were most interconnected with EEs in the same grade level.
8. Using the map locations, prioritize EEs that were most interconnected with EEs at the preceding and following grade levels.

Initial drafts of test blueprints were reviewed by the DLM Governance Board and TAC members in early 2014.

2.7.1.1. Original English Language Arts Blueprint

In the original ELA blueprint, after seeking input and consent from the governance board, content in the areas of Claim 1 (reading) and Claim 2 (writing) was prioritized for inclusion. In addition to a variety of reading testlets at each grade level, all students complete structured writing assessments in which test administrators engage students in a writing activity that addresses between one and six EEs in Claim 2. The EEs selected for the blueprint have:

- a broad range of potential application in novel contexts,
- the most connections to content at subsequent grade levels, and
- content that is relevant to a conceptual pathway in ELA that has applications in multiple domains or contexts.

2.7.1.2. Original Mathematics Blueprint

Like the original ELA blueprint, the breadth of mathematics EEs available on the original blueprint for assessment was deliberately broad. In each grade, the original blueprint addressed all four claims and each conceptual area relevant to the grade. All but a few EEs were included in the blueprint, excluding only those EEs that are very difficult to represent in a computer-based assessment environment. In addition to implementing these general guidelines, the mathematics blueprint reflected additional attempts to streamline the assessment across the grades to:

- avoid unnecessary redundancy in what is tested from year to year,
- highlight concepts and skills that provide students knowledge and skills for future mathematical learning during and beyond school, and
- acknowledge mathematical learning trajectories that connect the EEs over the course of several grades.

2.7.2. Blueprint Revision

Discussions began in the summer of 2016 with the DLM Governance Board about the need to review and revise the ELA and mathematics blueprints. At the December 2016 governance meeting, the governance board agreed that a reduction in the number of EEs on the ELA and mathematics Year-End blueprints was necessary to support more fine-grained reporting in student score reports (i.e., provide the Learning Profile in individual student score reports[12]). As part of this change, the governance board also agreed to change test specifications to increase the number of items that assess a single EE, as existing testlets typically measured each EE with one or two items. Additionally, because the testing times were lower than originally expected, the governance board agreed to increase the number of items administered per grade and subject. Thus, the governance board agreed to reduce the total number of EEs on the blueprint so that the number of items measuring each EE retained on the blueprint could be increased without over burdening educators and test administrators. These changes would allow reporting of fine-grained information in the Learning Profile.[13] The revised blueprints were adopted for use beginning with the 2019–2020 academic year.

2.7.2.1. Blueprint Revision Process

The revision process began with a content review of the original blueprints and a comparison to the DLM Instructionally Embedded model blueprints, which required fewer EEs per conceptual area and claim.

Three principles guided the selection of EEs for the revised blueprints, based on the criteria used for developing the original blueprints in 2014. Below, “goal(s)” included in each principle describe outcomes related to each principle that will meet the guidelines stated above.

Principle 1: The blueprint should have broad coverage of academic content as described by the EEs. The goals related to this principle are to provide appropriate breadth of content coverage of EEs within each grade and subject, select EEs for the blueprint that represent useful and valuable content for students, keep proportional coverage of claims and/or conceptual areas close to or identical to the original blueprint, and select a number of EEs in each grade that are more consistent with the number of EEs required in the Instructionally Embedded model blueprint.

Principle 2: The blueprint should emphasize connections in skills and understandings from grade to grade. The goals related to this principle are to select EEs in each grade that are conceptually related to EEs in later grades and use the learning map structure to inform grade-to-grade decisions to provide a connected, continuous delivery of content across all grades.

Principle 3: The revised blueprint should allow for a testlet design where each EE is assessed by 3-5 items for a total of 35-45 items in each subject in each grade. The goal related to this principle is to reduce the number of EEs and increase the number of items per EE, supporting the delivery of fine-grained mastery information without exceeding the maximum allowable assessment length, as specified by the DLM Governance Board.

[12] See Chapter 7 of this manual for a complete description of score reports.

[13] For a description of score reports, see Chapter 7 of this manual.

2.7.2.2. Overview of Blueprint Revision

This section provides a content overview of the revisions to the blueprints in ELA and mathematics for grades 3–8 and high school. The revised blueprints prioritize a set of EEs using a set of rationale categories to provide appropriate breadth and depth of content coverage in each discipline. The rationale categories used in the development of this version of the blueprints are:

Category 1: Include EEs that introduce or extend critical academic skills to form particular learning pathways in a topic/subject across grades. The EEs under Category 1 introduce an important academic skill, are a crucial turning point in a topic/subject, combine multiple critical academic skills, and expand on critical academic skills acquired in a previous grade.

Category 2: Include EEs that maintain representative conceptual area or content coverage. The EEs under Category 2 maintain complete coverage of all conceptual areas; provide equivalent coverage, across grades, of similar academic skills in each conceptual area; are the initial or last EE on a topic/subject across grades; address unique skills; and have few critical linkage level skills shared with other EEs.

Category 3: Exclude EEs that have a high degree of similarity with another EE that will remain on the revised blueprints, within or across grades. The EEs under Category 3 provide preferential coverage of the same academic skills in only one conceptual area and do not significantly expand on academic skills acquired in one or more of the surrounding grades.

Category 4: Exclude EEs that allow for additional coverage or more learning opportunities to critical academic skills.

Table 2.7 lists the number of EEs approved for the original and revised ELA blueprint, and Table 2.8 lists all the EEs approved for the original and revised mathematics blueprint.

Table 2.7

Number of Essential Elements Approved for the Original and Revised English Language Arts Blueprint, by Conceptual Area

	Original blueprint						Revised blueprint					
Grade	C1.1	C1.2	C1.3	C2.1	C2.2	Total	C1.1	C1.2	C1.3	C2.1	C2.2	Total
3	7	5	2	2	—	16	4	3	1	2	—	10
4	7	6	1	3	0	17	3	4	1	3	0	11
5	3	8	4	2	0	17	1	5	2	2	0	10
6	1	10	3	2	0	16	1	6	1	3	0	11
7	1	8	4	5	0	18	1	4	3	5	0	13
8	0	9	3	5	0	17	0	6	2	5	0	13
9	0	9	3	3	2	17	0	5	3	4	2	14
10	0	9	3	3	2	17	0	5	3	4	2	14
11	0	8	4	4	2	18	0	5	3	4	2	14

Table 2.8

Number of Essential Elements Approved for the Original and Revised Mathematics Blueprints, by Conceptual Area

	Original blueprint										Revised blueprint									
Grade	C1.1	C1.2	C1.3	C2.1	C2.2	C3.1	C3.2	C4.1	C4.2	Total	C1.1	C1.2	C1.3	C2.1	C2.2	C3.1	C3.2	C4.1	C4.2	Total
3	3	—	1	0	1	2	1	2	1	11	2	—	1	0	1	1	1	1	1	8
4	2	2	1	3	1	3	1	2	1	16	1	0	1	1	1	3	0	0	1	8
5	2	3	2	2	1	3	1	—	1	15	1	2	1	2	1	0	1	—	0	8
6	1	2	2	—	2	—	1	3	—	11	0	2	0	—	2	—	1	2	—	7
7	2	1	3	3	1	—	2	1	1	14	1	0	3	1	1	—	0	1	0	7
8	1	1	2	4	1	—	1	1	3	14	1	0	1	1	1	—	1	1	2	8
9	—	—	3	2	1	0	0	2	0	8	—	—	3	2	0	0	0	2	0	7
10	—	—	1	1	0	1	2	2	2	9	—	—	1	1	0	1	2	1	2	8
11	—	—	2	1	0	0	1	0	5	9	—	—	2	1	0	0	1	0	2	6

EEs were included or excluded from the revised blueprints based on the four categories listed above. In both subjects, test development teams determined which EEs to exclude from the revised blueprint, while maintaining the three guiding principles. Some reasons for excluding EEs from the revised blueprint were a high degree of node overlap with other EEs included in the blueprint or similar skills to other EEs included in the blueprint. Additionally, some EEs were excluded to target other EEs representing skills that receive less coverage in the blueprint than other skills or to allow for additional coverage to be provided to the critical foundational nodes associated with other EEs included in the blueprint. Finally, some EEs were excluded if they did not significantly extend on the skills represented in the standards in the surrounding grades.

Table 2.9 shows how many EEs from the original blueprints in ELA and mathematics were either included or excluded from the revised blueprint, by category. The higher number of exclusions for the mathematics blueprints are due to differences in how writing EEs are assessed in ELA. In each grade and subject, the maximum test length was approximately 40 items. The DLM assessments are administered in testlets, and, with the exception of writing, each testlet measures 1 EE with 3–5 items.[14] Thus, in general, the blueprints contain 7–8 EEs, which allows for up to 40 items. As previously noted, writing is an exception. All writing EEs in a grade are assessed on a single testlet.[15] Thus, it is possible to include more EEs on the ELA blueprint, while staying under the desired test length of 40 items. This can be seen in Table 2.7, where is number of EEs is consistently higher than than the 7–8 EEs that are included on each mathematics blueprint (Table 2.7).

Table 2.9

Number of Essential Elements (EEs) Included or Excluded on the Revised Blueprint, by Categories

Category	English language arts	Mathematics
1 (Included EEs of critical skills)	40	11
2 (Included EEs to maintain coverage)	54	56
3 (Excluded similar EEs)	35	20
4 (Excluded EEs that allow for additional coverage)	6	20

Note. Category 1 and 2 for English language arts (ELA) do not add up to the sum of the ELA EEs in Table 3.3 because Grades 9 and 10 share the same EEs and two EEs were added that were not on the original blueprint.

2.7.2.3. Breadth and Depth of Node Coverage

As described in section 2.1 and section 2.4 of this chapter, EEs can be represented as mini-maps, where each linkage level is comprised of one or more nodes that represent a critical juncture of knowledge or skills on the pathway toward the grade-level learning target(s). This section describes how the revisions to the assessment blueprints maintain a proportional coverage of nodes from the DLM maps compared to the

[14] See Chapter 3 of this manual for a complete description of testlets and assessment content.

[15] See Chapter 3 of this manual for a complete description of writing testlets.

original blueprints. The revision provided consistent breadth and depth of node coverage for the grade-level EEs within each subject.

Table 2.10 and Table 2.11 present the node coverage for the revised blueprints compared to the original blueprints. Test development team revisions retained coverage of between 57% and 79% of nodes in ELA and 50% and 99% of nodes in mathematics outright. Of the nodes that were not retained, between 8% and 21% in ELA and 0% and 55% in mathematics were covered in other grades. The revised blueprints only reduced node coverage by between 2 and 8 nodes in ELA and 0 and 17 nodes in mathematics per grade.

Table 2.10

Node Overlap Between Original and Revised English Language Arts Blueprints and Coverage of Nodes Not Included in the Revised Grade-Level Blueprint

Grade	Nodes in original blueprint (*n*)	Nodes in revised blueprint (*n*)	Nodes retained (%)	Uncovered nodes in grade level (*n*)	Nodes covered in other grades (*n*)	Uncovered nodes (*n*)	Uncovered nodes (%)
3	54	38	70.4	16	10	6	11.1
4	53	36	67.9	17	9	8	15.1
5	61	35	57.4	26	21	5	8.2
6	64	40	62.5	25	17	8	12.5
7	61	40	65.6	21	17	4	6.6
8	53	40	75.5	13	11	2	3.8
9–10	62	42	67.7	20	16	4	6.5
11–12	53	42	79.2	11	8	3	5.7

Note. English language arts Essential Elements for high school are organized into two grade bands (9–10 and 11–12) instead of individual grades.

Table 2.11

Node Overlap Between Original and Revised Mathematics Blueprints and Coverage of Nodes Not Included in the Revised Grade-Level Blueprint

Grade	Nodes in original blueprint (*n*)	Nodes in revised blueprint (*n*)	Nodes retained (%)	Uncovered nodes in grade level (*n*)	Nodes covered in other grades (*n*)	Uncovered nodes (*n*)	Uncovered nodes (%)
3	99	83	83.8	16	16	0	<0.1
4	144	91	63.2	53	47	6	4.2
5	143	71	49.7	72	55	17	11.9
6	98	66	67.3	32	29	3	3.1
7	133	76	57.1	57	50	7	5.3
8	115	71	61.7	44	27	17	14.8
9	92	85	92.4	7	5	2	2.2
10	72	71	98.6	1	0	1	1.4
11	76	61	80.3	15	4	11	14.5

Note. Because the mathematics Essential Elements for high school are not divided into grades, the high school Essential Elements are organized into three grade-level integrated mathematics courses: Math 9, Math 10, and Math 11.

In summary, the revised blueprints provide a connected, continuous delivery of content across all grades. Furthermore, the revisions to the assessment blueprints allowed for an increase in the number of items covered in each EE while simultaneously collecting finer-grained student mastery information necessary to create informative and useful student reports. The complete blueprints for ELA[16] and mathematics[17] are available on the DLM website, and are also included in Appendix B.1 of this manual.

2.8. Alignment

ACERI Partners, LLC, conducted an external alignment study of the DLM operational assessment system (Flowers & Wakeman, 2016a) to investigate the relationships between the content structures in the DLM System. A modification of Links for Academic Learning alignment methodology (Flowers et al., 2009) was used to evaluate the coherence of the DLM System. The alignment study focused on the following relationships (as illustrated by the corresponding numbers in Figure 2.12):

1. College and career readiness content standards and EEs
2. An EE and its target-level node(s)
3. The vertical articulation of the linkage levels associated with an EE
4. DLM learning map nodes within a linkage level and assessment items

For a complete description of the original alignment study, see Flowers and Wakeman (2016b). Following

[16] https://dynamiclearningmaps.org/sites/default/files/documents/Manuals_Blueprints/DLM_YE_ELA_Blueprint.pdf
[17] https://dynamiclearningmaps.org/sites/default/files/documents/Manuals_Blueprints/DLM_YE_Math_Blueprint.pdf

the revision to the assessment blueprint (see section 2.7.2 above), DLM staff reanalyzed the data collected from the original study to reflect the changes to the blueprint EEs and the items administered.

This section describes the results of the first three relationships, reflecting the updated alignment study based on the revised assessment blueprints. For a discussion of the alignment of assessment items, see Chapter 3 of this manual.

In ELA, a total of 79 testlets (25% of the pool that was randomly sampled for the original alignment report) and 304 items (20%) were examined for alignment. In mathematics, 70 (29%) testlets and 192 (21%) items were evaluated.

Content and performance centrality were the primary measures of alignment. Content centrality is a measure of the degree of fidelity between the item/element and the content of the academic grade-level target. Specifically, it measures the degree of fidelity between the college and career readiness standard and the EE or between the EE and the target-level node(s). Panelists rated each pair as having no link, a far link, or a near link. Performance centrality represents the degree to which the operational assessment item and the corresponding academic grade-level content target contain the same performance expectation. Specifically, performance centrality measures the degree to which the performance expectation matches between the college and career readiness standard and the EE or between the EE and the target-level node(s). Panelists rated the degree of performance centrality between each pair as none, some, or all. If panelists identified a relationship that did not meet criteria for alignment (e.g., no link for content centrality), they provided additional feedback. When evaluating items, panelists used the DLM cognitive process dimension taxonomy to identify the category for the highest cognitive dimension required of the student when responding to the item.

The following sections provide a brief summary of findings from the external alignment study. Full results are provided in the updated technical report (Flowers & Wakeman, 2020).

2.8.1. Alignment of College and Career Readiness Standards and Essential Elements

All EEs identified in the assessment blueprints were included in these analyses. Table 2.12 and Table 2.13 display the results of content centrality and performance centrality ratings, respectively. For content centrality, EEs were defined as “Met” if they were given a “Far” or “Near” rating. For performance centrality, EEs were defined as “Met” if they were given a “Some” or “All” rating. The level of acceptable content centrality for alternate assessments is 80% (Flowers et al., 2009). No recommended acceptable level of performance centrality is provided for alternate assessments. In ELA, 82% of EEs were rated as maintaining fidelity to the content and the grade-level college and career readiness standards, and 91% of EEs of were rated as having the same performance expectation. Similarly, in mathematics, 80% of EEs were rated as maintaining fidelity, and 74% were rated as having the same performance expectation. This is an acceptable level of alignment given the rigor of grade-level standards and the need to provide access for all students with the most significant cognitive disabilities. EEs that were rated “No” for maintaining fidelity or “None” for having the same performance expectation were forwarded to the test development teams for further review.

Table 2.12

Content Centrality of College and Career Readiness Standards to Essential Elements

Subject	Total *N*	No (%)	Far (%)	Near (%)	Met (%)
English language arts	96	17 (18)	69 (72)	10 (10)	79 (82)
Mathematics	100	20 (20)	73 (73)	7 (7)	80 (80)

Note. Gray shading indicates acceptable level of alignment.

Table 2.13

Performance Centrality of College and Career Readiness Standards to Essential Elements

Subject	Total *N*	None (%)	Some (%)	All (%)	Met (%)
English language arts	96	9 (9)	75 (78)	12 (13)	87 (91)
Mathematics	100	26 (26)	57 (57)	17 (17)	74 (74)

Note. Gray shading indicates acceptable level of alignment.

2.8.2. Alignment of Essential Element and Target-Level Nodes

Table 2.14 and Table 2.15 display the content and performance centrality of the alignment of EEs to target-level node(s), which are the node(s) that reflect the grade-level expectation in the EE. The number of EEs in Table 2.14 and Table 2.15 is different from Table 2.12 and Table 2.13 because some EEs corresponded to more than one target-level node. All EEs were rated as aligned to the target-level nodes, with most EEs rated as near the target-level node. Similar results were found for performance centrality. All EEs were rated as meeting some or all of the performance expectations found in the target-level node. These findings suggest a strong alignment between EEs and target-level nodes.

Table 2.14

Content Centrality of Essential Elements to Target-Level Nodes

Subject	Total *N*	No (%)	Far (%)	Near (%)	Met (%)
English language arts	96	0 (0)	4 (4)	92 (96)	96 (100)
Mathematics	145	0 (0)	28 (19)	117 (81)	145 (100)

Note. Gray shading indicates acceptable level of alignment.

Table 2.15

Performance Centrality of Essential Elements to Target-Level Nodes

Subject	Total *N*	None (%)	Some (%)	All (%)	Met (%)
English language arts	96	0 (0)	12 (13)	84 (88)	96 (100)
Mathematics	145	0 (0)	32 (22)	113 (78)	145 (100)

Note. Gray shading indicates acceptable level of alignment.

2.8.3. Vertical Articulation of Linkage Levels for Each Essential Element

Panelists evaluated linkage levels to see if they reflected a progression of knowledge, skills, and understandings. Table 2.16 shows the results of the vertical articulation of the linkage levels for the EEs at each grade level for ELA and mathematics. For ELA, 95 linkage levels were reviewed by panelists and 76 (80%) were rated as showing a clear progression from precursor to successor nodes. The low rating for seventh grade was due to panelists reporting that the Initial Precursor was not clearly part of the progression in the ordered nodes. For mathematics, 66 linkage levels were reviewed and 64 linkage levels (97%) were rated as demonstrating a clear progression in the ordered nodes. EEs that were rated as not reflecting a clear progression were forwarded to the test development teams for further review.

Table 2.16

Vertical Articulation of Linkage Levels for Essential Elements

Grade	Total *N*	Clear progression (%)
English language arts		
3	10	9 (90)
4	11	9 (82)
5	10	8 (80)
6	11	9 (82)
7	13	7 (54)
8	13	10 (77)
9–10	14	12 (86)
11–12	13	12 (92)
Total	*95*	*76 (80)*
Mathematics		
3	8	8 (100)
4	8	7 (88)
5	8	8 (100)
6	7	6 (86)
7	6	6 (100)
8	8	8 (100)
9	7	7 (100)
10	8	8 (100)
11	6	6 (100)
Total	*66*	*64 (97)*

2.9. Learning Maps for the Operational Assessment

Table 2.17 includes the overall statistics describing the DLM learning maps as implemented for the operational assessment. This version of the set of ELA, mathematics, and foundational DLM learning maps is the basis for the operational assessments. Foundational nodes support both ELA and mathematics maps.

Table 2.17

Number of Nodes and Connections in the Learning Maps

Node category	Number of nodes	Number of connections
Foundational	150	277
English language arts	1,919	5,045
Mathematics	2,399	5,200
Total	*4,468*	*10,522*

2.10. Conclusion

The DLM assessments are built on learning maps that describe pathways for students to acquire knowledge, skills, and understandings. Different areas of the map are associated with conceptual areas and EEs, which are the learning targets for the DLM assessments. To ensure all students have access to grade-level academic content, each EE is available at five linkage levels, which are small collections of nodes that represent critical junctures on the path toward and beyond the learning target. Both the learning maps and EEs were developed by subject-matter experts synthesizing the research literature with stakeholder feedback. Procedural evidence and preliminary empirical evidence both support the structure of the DLM learning maps. Finally, the external alignment study provides evidence of the connections between DLM System components and the college and career readiness standards, via EEs, learning map nodes, and linkage levels.

3. Assessment Design and Development

Dynamic Learning Maps® (DLM®) assessments measure student knowledge and skills using items organized into short testlets. This chapter describes how assessment content is developed and evaluated.

The chapter describes DLM assessments and how they were developed using principles of Evidence-Centered Design (ECD) and Universal Design for Learning (UDL). The chapter first describes the design of English language arts (ELA) reading and writing testlets, as well as mathematics testlets, and alternate testlets for students who are blind or have visual impairments. The chapter then provides information on the test development process, including item writing, and external reviews for content, bias, and accessibility. The chapter then presents evidence of the alignment of linkage level nodes to assessed items and evidence of students' response processes when engaging with assessment content. The chapter concludes by presenting evidence of item quality, including a summary of field-testing data analysis and associated reviews, a summary of the pool of operational testlets available for administration, and an evaluation of differential item functioning.

3.1. Assessment Structure

As discussed in Chapter 2 of this manual, the DLM Alternate Assessment System uses learning maps as the basis for assessment, which are highly connected representations of how academic skills are acquired as reflected in research literature. Nodes in the maps represent specific knowledge, skills, and understandings in ELA and mathematics, as well as important foundational skills that provide an understructure for the academic skills. The maps go beyond traditional learning progressions to include multiple pathways by which students develop content knowledge and skills.

Four broad claims were developed for ELA and mathematics, which were then subdivided into nine conceptual areas, to organize the highly complex learning maps (see Chapter 2 of this manual). Claims are overt statements of what students are expected to learn and be able to demonstrate as a result of mastering skills within a very large neighborhood of the map. Conceptual areas are nested within claims and comprise multiple conceptually related content standards, and the nodes that support and extend beyond the standards. The claims and conceptual areas apply to all grades in the DLM System.

Essential Elements (EEs) are specific statements of knowledge and skills, analogous to alternate or extended content standards. The EEs were developed by linking to the grade-level expectations identified in the Common Core State Standards (see Chapter 2 of this manual). The purpose of the EEs is to build a bridge from the Common Core State Standards to academic expectations for students with the most significant cognitive disabilities.

For each EE, five linkage levels—small collections of nodes that represent critical junctures on the path toward and beyond the learning target—were identified in the map. Assessments are developed at each linkage level for a particular EE.

Testlets are the basic units of measurement in the DLM System. Testlets are short, instructionally relevant measures of student skills and understandings. Students take a series of testlets to achieve blueprint coverage.

Each testlet is made up of three to nine assessment items. Assessment items were developed based on nodes at the five linkage levels for each EE. Each testlet measures an EE and linkage level, with the exception of writing testlets. The Target linkage level reflects the grade-level expectation aligned directly to the EE. For each EE, small collections of nodes are identified earlier in the map that represent critical junctures on the path toward the grade-level expectation. Nodes are also identified beyond the Target to give students an opportunity to grow toward the grade-level targets for students without significant cognitive disabilities.

There are three levels below the Target and one level beyond the Target.

1. Initial Precursor
2. Distal Precursor
3. Proximal Precursor
4. Target
5. Successor

3.2. Items and Testlets

In reading and mathematics, testlets are based on nodes for one linkage level of one EE. Writing testlets measure multiple EEs and linkage levels. Each testlet contains three to nine items. All testlets begin with a nonscored engagement activity, which includes a stimulus related to the assessment designed to help the student focus on the task at hand. In ELA, the engagement activity for reading testlets is a story or informational text.

Several item types are used in DLM testlets. While most types are used in both ELA and mathematics testlets, some types are used only in testlets for one subject. The following item types are used in DLM testlets:

- Multiple-choice single-select
- Multiple-choice multiple-select
- Select text (ELA only)
- Matching lines (mathematics only)

Most items within the testlets are multiple-choice single-select items with three answer options presented in a multiple-choice format using either text or images. Technology-enhanced items (i.e., not multiple choice with a single correct response) are used on a limited basis due to the additional cognitive load they can introduce. Some assessed nodes in the DLM maps require complex cognitive skills such as sorting or matching that are difficult to assess efficiently in a multiple-choice format while keeping the length of the assessment constrained. In these cases, technology-enhanced items that matched the construct described by the nodes were used to avoid having to use many multiple-choice items to assess the same construct. Evidence for the accessibility and utility of technology-enhanced items was collected from item tryouts and cognitive labs, which are described later in this chapter, in section 3.5.

There are two general modes for DLM testlet delivery: computer-delivered and educator-administered. Computer-delivered assessments are designed so students can interact independently with the computer, using special assistive technology devices such as alternate keyboards, touch screens, or switches as necessary. Computer-delivered testlets emphasize student interaction with the content of the testlet, regardless of the means of physical access to the computer. Therefore, the contents of testlets, including directions, engagement activities, and items, are presented directly to the student. Educators may assist students during these testlets using procedures described in Chapter 4 of this manual.

Educator-administered testlets are designed for the educator to administer outside the system, with the test administrator recording responses in the system rather than the student recording their own responses. These educator-administered testlets include onscreen content for the test administrator that begins by telling, in a general way, what will happen in the testlet. Directions for the test administrator then specify the materials that need to be collected for administration. After the educator direction screen(s), educator-administered testlets include instructions for the engagement activity. After the engagement activity, items are presented. All educator-administered testlets have some common features:

- Directions and scripted statements guide the test administrator through the administration process.
- The engagement activity involves the test administrator and student interacting directly, usually with objects or manipulatives.
- The test administrator enters responses based on observation of the student's behavior.

Testlet organization, the type of engagement activity, and the type and position of items vary depending on the intended delivery mode (computer-administered or educator-administered) and content being assessed (reading, writing, or mathematics). Specific descriptions and examples of the structure of testlets, engagement activities, and different item types are included in the following sections related to reading, writing, and mathematics testlets.

3.2.1. English Language Arts Reading Testlets

ELA reading testlets are built around texts adapted from or related to grade-level–appropriate general education texts. Short narrative texts are constructed from books commonly taught in general education, and short informational texts are written to relate to thematic elements from narratives. All texts are deliberately written to provide an opportunity to assess specific nodes in the maps associated with different EEs and linkage levels. Text complexity for narrative texts is reduced from the grade-level texts for non-DLM students, focusing on core vocabulary, simple sentence structure, and readability.[18]

ELA Claim 1 states, "Students can comprehend text in increasingly complex ways." To provide access to a wide range of student needs, the complexity of the text is held relatively constant, but the complexity of cognitive tasks needed to answer items is increased. Texts are generally very brief, typically between 50 and 200 words in length. Texts are presented with 1–3 sentences on a screen with an accompanying image. One screen is presented at a time. Students and educators can navigate forward and backward between screens. ELA texts contain between 6 and 25 screens.

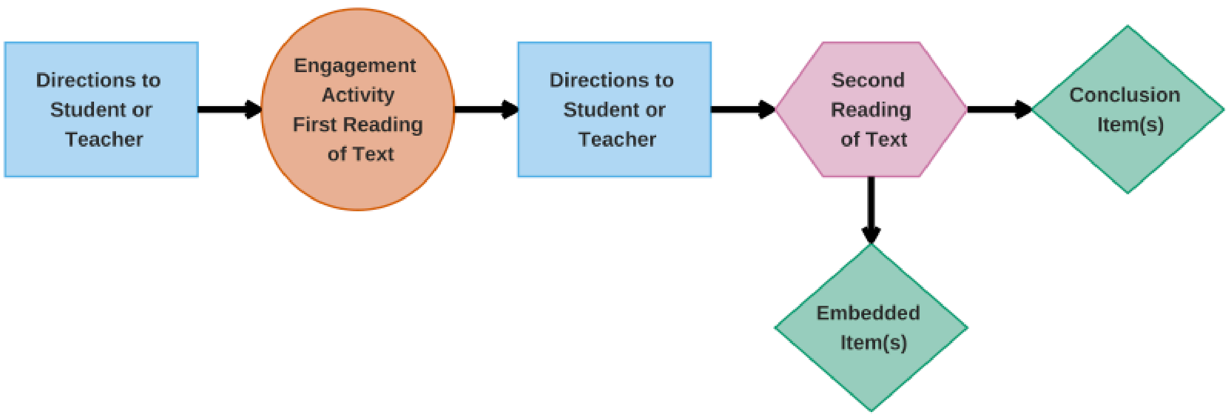
ELA reading testlets follow a basic structure, with variations for some educator-administered testlets or testlets assessing nodes that require students to compare more than one text. Figure 3.1 shows the elements of an ELA reading testlet. An ELA reading testlet begins with directions to the student in computer-delivered testlets, or to the test administrator in educator-administered testlets, followed by an engagement activity. The engagement activity consists of the first reading of the story or text that allows students to read, become familiar with, and comprehend the story or text before responding to any items. After the first reading, directions to the student or educator explain that the story or text is complete and

[18] See section 3.3.4 of this chapter for a complete description of ELA texts.

that next, students will re-read the text and respond to some questions. After these directions, the student begins the second reading. The second reading is presented the same as the first reading, but with assessment items embedded into the text. Embedded items are placed between the screens of the text, and conclusion items appear after the second reading of the text is complete.

Figure 3.1

Elements of an English Language Arts Reading Testlet



3.2.1.1. Engagement Activities

ELA reading testlets include an engagement activity that outlines the structure of the testlet and instructs the student and/or test administrator how to proceed through the testlet. In reading testlets, the first reading of the text is considered a part of the engagement activity. In computer-delivered testlets, the engagement activity instructs students to read the text. Students may read on their own or with read-aloud support as a selected accessibility support (see Chapter 4 of this manual). An example of a computer-delivered engagement activity screen is shown in Figure 3.2.

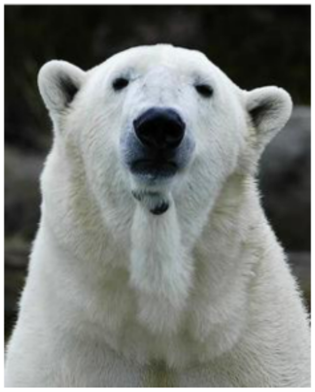
Figure 3.2

Example of an English Language Arts Computer-Delivered Reading Engagement Activity Screen

Educator-administered testlets require the test administrator to assess the student outside the online testing platform and enter responses. In educator-administered testlets, the engagement activity introduces the testlet to the test administrator, who will read the story or text with the student. For educator-administered ELA reading testlets, the engagement activity is also the first reading of the text. The directions for the engagement activity are presented to the test administrator. An example of an educator-administered engagement activity screen is shown in Figure 3.3.

Figure 3.3

Example of an English Language Arts Educator-Administered Reading Engagement Activity Screen



3.2.1.2. Second Reading of Text

The decision to use two readings of the same text in each reading testlet was made in consideration of Cognitive Load Theory. Within the context of instructional and assessment design, the application of Cognitive Load Theory emphasizes decreasing the memory storage demands in order to emphasize processing components of the activity (Chandler & Sweller, 1991). Thompson, Johnstone, and Thurlow (2002) describe a set of strategic processes aligned with UDL that can be seen as a way to reduce the extraneous cognitive load for students with disabilities. The approach adopted for reading testlets was intended to reduce the demands on student working memory by providing an opportunity to read a text and then immediately read it again, embedding items into the second reading between screens that present the text. Examples of the skills and processes assessed by the embedded items include

- identifying features of texts,
- identifying details in texts,
- finding specific words in texts, and
- identifying relationships described in texts.

The use of embedded items means that rather than having students read a story once and then recall how a character felt at some prior point in the story, the embedded question is presented when the character's feeling state is active in working memory.

Conclusion items are presented after the conclusion of the second reading of the text. These items focus on products of comprehension or assessments of elements that depend on a representation of the entire text. Examples of the skills and products that conclusion items focus on include

- identifying the theme and/or main idea(s) of a text,
- identifying structural elements of an entire text (e.g., beginning, middle, end),
- comparing multiple texts, and
- analyzing purpose, evidence, or goals in a text.

Testlets can include a mixture of embedded and/or conclusion items.

3.2.1.3. Items

Computer-delivered ELA reading testlets contain three item types: multiple choice, multiple-choice multiple-select, and select-text. Technology-enhanced items such as select-text are used when nodes at certain linkage levels would be difficult to assess using a multiple-choice item. Items of all three types can be embedded or conclusion items. Educator-administered ELA reading testlets use only multiple-choice items.

For many multiple-choice items, the stem is a question related to the text. For others, the stem includes a line from the story or text followed by a question. Most multiple-choice items contain three answer options, one of which is correct. Students may select only one answer option. Most answer options are words, phrases, or sentences. For items that evaluate certain map nodes, answer options are images. An example of an ELA multiple-choice item with text answer options is shown in Figure 3.4.

Figure 3.4

Example of an English Language Arts Computer-Delivered Multiple-Choice Item

How did Mary and Martha feel about each other?

They lived in a nice house.

They loved living together.

They made a dress every day.

For multiple-choice multiple-select items, the item stem directs the student to select answers from four answer options, where more than one is correct. Answer options are words, phrases, or sentences. Multiple-choice multiple-select items allow students to choose up to four answer options. An example of an ELA multiple-choice multiple-select item is shown in Figure 3.5.

Figure 3.5

Example of an English Language Arts Computer-Delivered Multiple-Choice Multiple-Select Item

Choose two things that can be planted in a garden.

carrots

flowers

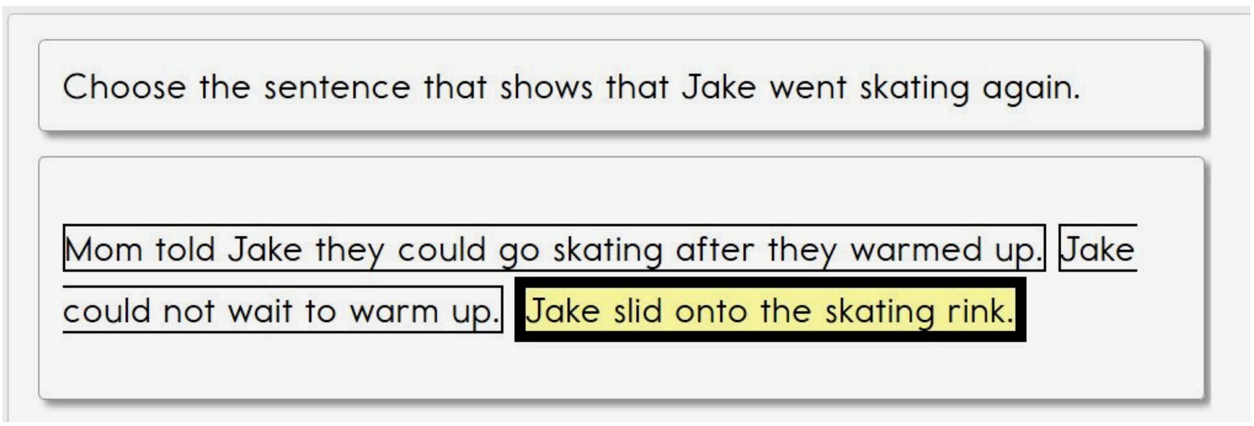
gloves

rocks

Select-text items direct students to select an answer from a passage taken from the story or text. In Figure 3.6, the student chose the appropriate sentence from a short passage. The stem is a directive to the student to select a word, phrase, or sentence from the passage. Certain words have a box around them to indicate they are answer options. When a student selects a word, phrase, or sentence, it becomes highlighted in yellow.

Figure 3.6

Example of an English Language Arts Select-Text Item

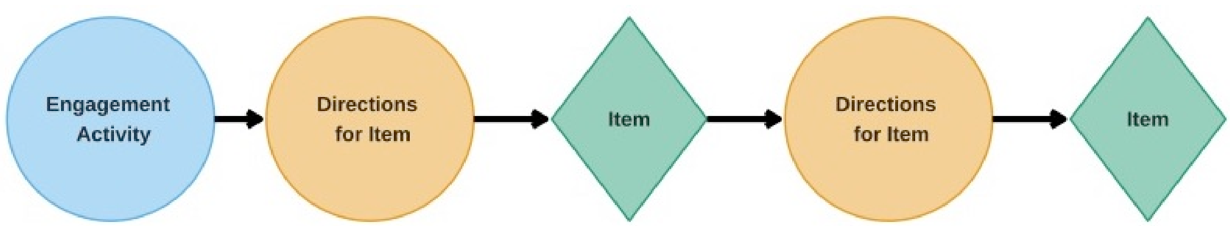


3.2.2. English Language Arts Writing Testlets

Writing testlets measure multiple EEs and linkage levels. All ELA writing testlets are educator-administered. For writing testlets, the test administrator engages in a scripted activity with a student outside the online testing platform and then enters observations and evaluation of the student's writing process and product into the online testing platform. Figure 3.7 shows the structure of a writing testlet. The testlet begins with an engagement activity and provides directions for the test administrator for each item before the item is presented.

Figure 3.7

Elements of an English Language Arts Writing Testlet



Every grade has an Emergent and Conventional writing testlet. Emergent writing describes the marks, scribbles, and random selection of letters seen in beginning writers (Erickson et al., 2010). The DLM EEs focus on having students work toward an understanding of writing as a form of communication and the ability to write about information. Emergent writing testlets focus on nodes in the map that are identified as being important precursor skills on the way toward conventional writing. Conventional writing includes methods of writing that use orthography (letters, words) assembled in ways that are meaningful to others. Key conceptual components of conventional writing include an understanding that words comprise letters, that words have meanings, and that written words can be put together to communicate to others. Key behaviors associated with conventional writing include writing letters and words using a traditional writing tool or alternate pencil.

3.2.2.1. Engagement Activities

Writing testlets begin with a materials screen that lists materials the student will need to complete the testlet, instructions to the test administrator about administering the testlet, and instructions to the test administrator on administrating an engagement activity that outlines how students should choose an object or topic to write about. Test administrators are directed to engage the student in thinking about a topic to encourage recall of relevant prior knowledge before a student begins to write. These instructions provide guidance to the test administrator on allowing the student to select an object to use or topic to write about as they complete the items in the writing testlet. Figure 3.8 shows an example.

Figure 3.8

Example of an English Language Arts Writing Engagement Activity

> Educator Directions:
>
> Give the student time to select an informational topic to write about. Provide examples of informational topics that have been used during instruction. Once the student has selected an informational topic to write about, select "NEXT."

3.2.2.2. Items

In writing testlets, the engagement activity is followed by items that require the test administrator to evaluate the student's writing process. Some writing testlets also evaluate the student's writing product, and these product items occur at the end of the testlet. Process and product items are multiple-choice single-select or multiple-choice multiple-select items with answer choices. Educators evaluate samples for easily perceptible text features requiring minimal inference, such as correct syntax, spelling, capitalization, and punctuation. Both item types ask test administrators to select a response from a checklist of possible responses that best describes what the student did or produced as part of the writing testlet.

Items that assess student writing processes are evaluations of the test administrator's observations of the student as the student completes items in the testlet. Figure 3.9 shows an example of a process item from an emergent writing testlet focused on letter identification in support of writing the student's first name. The construct assessed in this item is the student's ability to identify the first letter of his or her own name. In the example, either "writes the first letter of his or her own name" or "indicates the first letter of his or her own name" are scored as correct responses (Figure 3.9). The inclusion of multiple correct answer options is designed to ensure that this testlet is accessible to emergent writers who are beginning to write letters and emergent writers who have not yet developed writing production skills but are still able to identify the first letter of their name.

Figure 3.9

Example of an English Language Arts Emerging Writing Item Focused on Process

Choose the highest level that describes your observation.

- ☐ Writes the first letter of his or her first name
- ☐ Indicates the first letter of his or her first name
- ☐ Writes or indicates another letter
- ☐ Writes marks or selects symbols other than letters
- ☐ Attends to other stimuli
- ☐ No response

Items that assess writing products are the test administrator's evaluations of the product created by the student as a result of the writing processes completed in the administration of the testlet. Figure 3.10 provides an example of an item that evaluates a student's writing product. For some product items, administrators choose all the responses in the checklist that apply to the student's writing product. The interrater reliability of the writing sample evaluations is described in Chapter 7 of this manual.

Figure 3.10

Example of an English Language Arts Conventional Writing Item Focused on Product

After the student has finished writing, choose the highest level that describes your evaluation of the final product. Correct spelling is not evaluated in this item.

- ☐ Wrote two or more facts or details related to the informational topic
- ☐ Wrote one fact or detail related to the informational topic
- ☐ Wrote facts or details unrelated to the informational topic
- ☐ Wrote marks or selected symbols other than letters
- ☐ Did not write

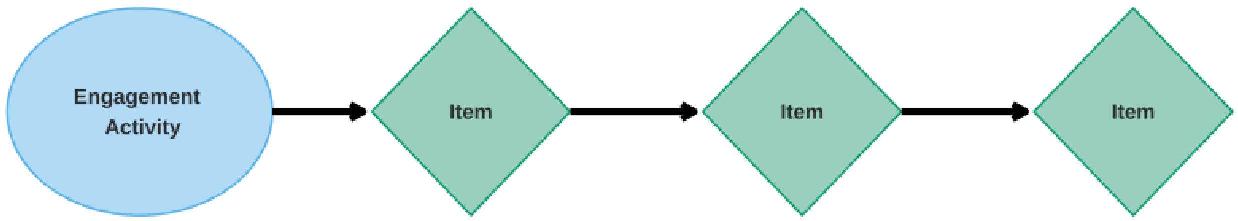
Writing testlets are constructed to provide test administrators with a coherent structure for delivering an instructionally relevant writing task to the student. Each writing testlet provides multiple opportunities for the test administrator to evaluate writing processes, and in some levels and grades, products. Each writing testlet measures multiple EEs. All EEs have five identified linkage level nodes, but writing testlets combine the delivery of assessments into emergent testlets and conventional testlets in Grades 3–8 and high school. The initial and distal precursor levels are combined into an emergent writing testlet. The proximal precursor, target, and successor levels are combined into a conventional writing testlet. Because writing testlets address multiple EEs and linkage levels, they differ from reading and mathematics testlets in that answer choices, rather than item stems, are aligned to nodes. Some items may include answer options associated to different EEs and linkage levels. For example, in Figure 3.9, the first two answer options are associated with a distal precursor linkage level node, while the third answer option is associated with an initial precursor linkage level node for the same EE.

3.2.3. Mathematics Testlets

Mathematics testlets are designed to assess student knowledge and skills by focusing on cognitive processes and reducing extraneous cognitive load by using a common context across all items in the testlet. Figure 3.11 shows the order of presentation of mathematics testlets. The testlet begins with an engagement activity, which is followed by items that assess specific nodes associated with the EE and linkage level.

Figure 3.11

Elements of a Mathematics Testlet



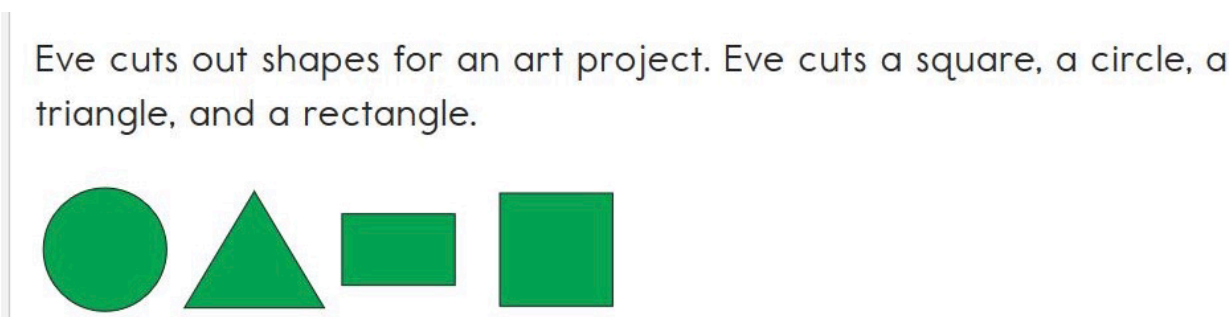
Following the engagement activity, three to five items are presented to the student. Educator-administered testlets, delivered off-screen, require the student to interact with manipulatives and respond to specific questions asked by the educator. Items on computer-delivered testlets are delivered onscreen.

3.2.3.1. Engagement Activities

Mathematics testlets start with an engagement activity that provides context for the questions. The mathematics engagement activity in Figure 3.12 provides a context related to shapes and activates a cognitive process about putting things together. This example is written to be broadly applicable to students who might have personal experiences in art class or another context with putting shapes together. This activity is intended to prepare the student for items about combining shapes.

Figure 3.12

Example of a Mathematics Engagement Activity



3.2.3.2. Items

Computer-delivered mathematics testlets contain three item types: multiple choice, multiple-choice multiple-select, and matching. Technology-enhanced items such as multiple-select and matching are used when nodes at certain linkage levels would be difficult to assess using a multiple-choice item. One example is for students to sort objects based on shape. Educator-administered mathematics testlets use only multiple-choice items.

Multiple-choice items contain three answer options, one of which is correct. Students can select only one answer option. Most mathematics items use a multiple-choice item type. An example multiple-choice mathematics item using text as answer options is shown in Figure 3.13.

Figure 3.13

Example of a Mathematics Multiple-Choice Item with Text

Jay counts $1.00. Jay then counts $0.25. What is the total amount Jay counts?

$0.75

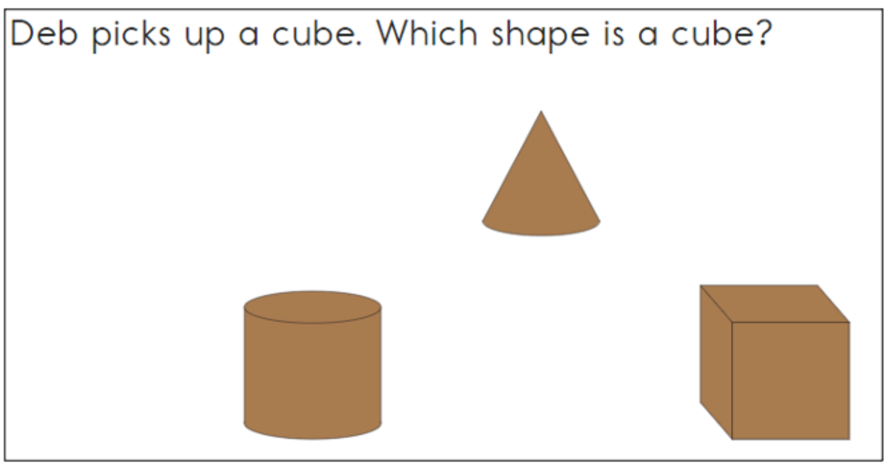
$1.25

$1.75

An example multiple-choice mathematics item using pictures as response options is shown in Figure 3.14.

Figure 3.14

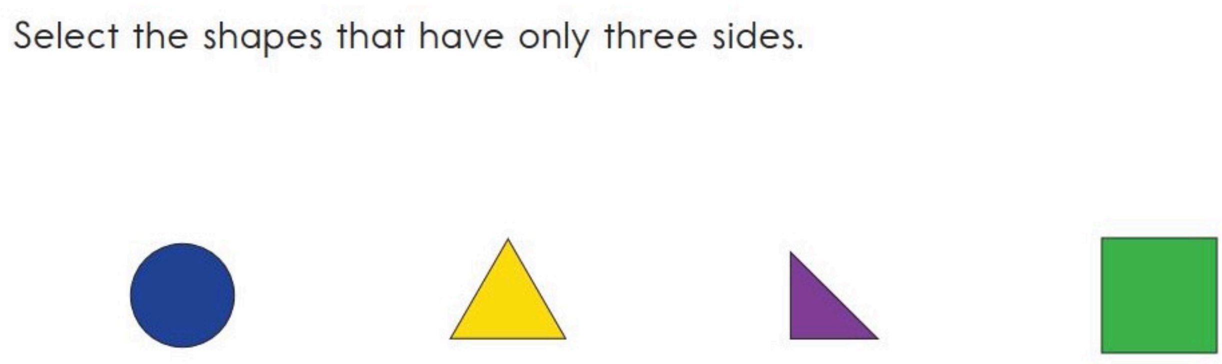
Example of a Mathematics Multiple-Choice Item with Pictures



Multiple-choice multiple-select items provide the student with the opportunity to make more than one answer choice. An example of a multiple-choice multiple-select item is shown in Figure 3.15.

Figure 3.15

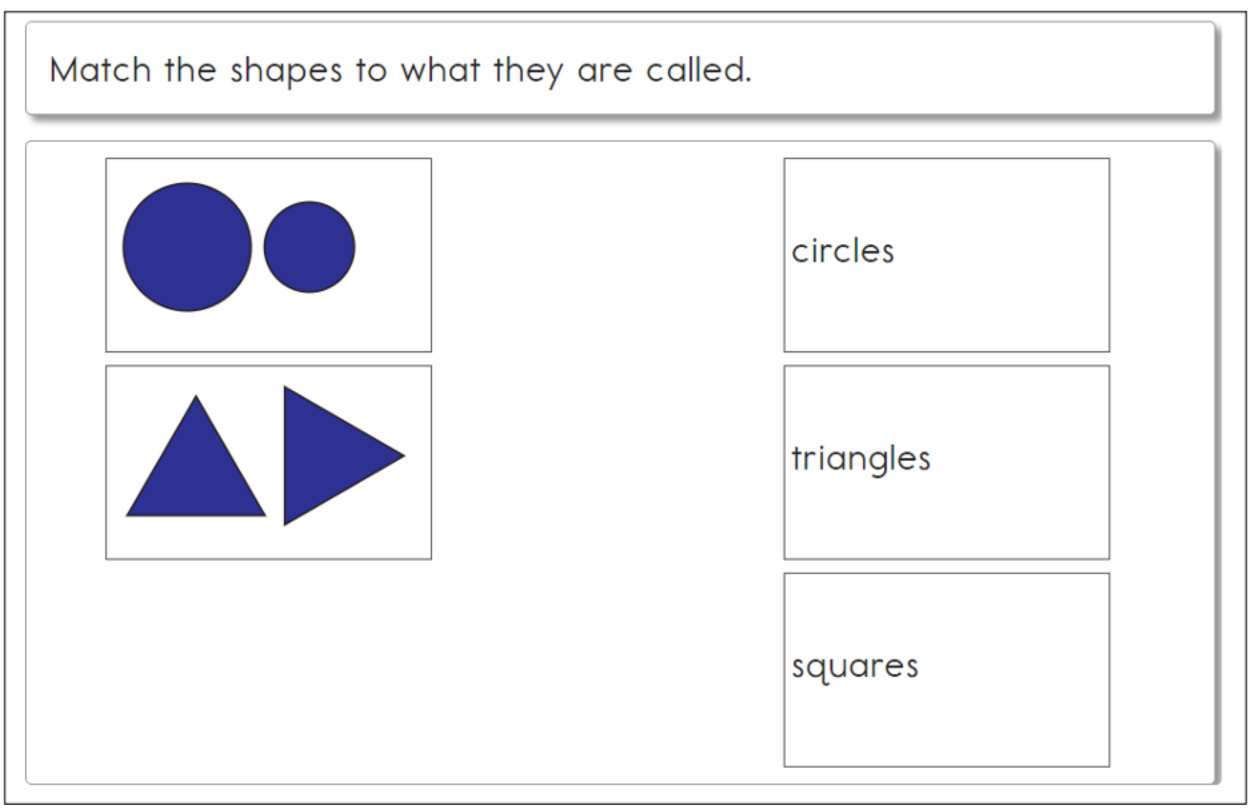
Example of a Mathematics Multiple-Choice Multiple-Select Item



Some mathematics testlets use matching items where students match items from two lists. An example of a matching-lines item is shown in Figure 3.16. In this item type, the student selects a box from the left and then a box from the right. When the option from the right is selected, a line is drawn between the two selected boxes.

Figure 3.16

Example of a Mathematics Matching Item



3.2.4. Alternate Testlets for Students who are Blind or Have Visual Impairments

Two types of alternate testlets were developed for students who are blind or have visual impairments (BVI). Both were designed as alternates to the general testlet form for that EE and linkage level.

The first type was alternate testlets, called BVI forms, which are created when nodes are difficult to assess online for students who have visual impairments, even with features such as read aloud or magnification (such as locating a point on an onscreen coordinate plane). Computer-delivered BVI testlets begin with an instruction screen for the test administrator, then continue with content intended for the student to access. These testlets list materials that the educator may use to represent the onscreen content for the student. Objects are used instead of tactile graphics, which are too abstract for the majority of students with the most significant cognitive disabilities who are also blind. However, educators have the option to use tactile graphics if their student can use them fluently. In educator-administered BVI testlets, test administrators are recommended to use special materials for students who are blind or have visual impairments, but other familiar materials may be substituted. Details about needed materials for testlets delivered in both modes (computer- and educator-delivered) are provided in the Testlet Information Pages (see Chapter 4 of this manual). In 2021, mathematics BVI forms were retired and instead BVI pages were integrated into the Testlet Information Pages. In this case, students receive a standard mathematics form, and the Testlet Information Pages provide information on how to make appropriate adaptations for the student. This makes all general forms accessible to students who are blind or have visual impairments and increased the number of testlets available to those students. In 2022, there were a total of 24 BVI forms available for ELA EE and linkage level combinations, and BVI forms were selected for 855 students (1%).

The second type was Braille forms, which are available for grades 3–5 at the Target and Successor levels and in grades 6 through high school at the Proximal Precursor, Target, and Successor levels. Braille is intentionally limited to these grades and linkage levels as alternate forms. Braille forms are provided when sighted students are expected to read the equivalent content. At the lowest two linkage levels, and occasionally at the third linkage level in the lower grades, the assessed nodes are at levels where students are not yet reading, even on an emerging basis. For example, a student who is asked to differentiate between some and none, or to identify his or her own feelings, is not working on concrete representations of text for the purpose of reading. Because general versions of testlets at those EEs and levels do not require reading, braille is not provided at those levels. In total, Braille forms were made available for 311 EE/linkage level combinations (168 ELA, 143 mathematics), and Braille forms were selected for 75 students (<1%).

3.2.5. Practice Activities and Released Testlets

The DLM System provides educators and students with the opportunity to preview assessments by using practice activities and released testlets. Practice activities and released testlets are accessed through the online testing platform. Using login information provided by the system, both types of activities can be completed as many times as desired.

The practice activities are designed to familiarize users with testlet administration in the online testing platform. One practice activity is for educators, and the other practice activity is for students. The practice activity for educators provides information about the types of testlets that will be administered directly by the educator and is structured like an educator-administered testlet. The first screen introduces the concept of administering the educator-administered testlets. Next, the practice activity familiarizes educators with the navigation icons on the screen, as well as the spoken audio icon (available when students have this feature enabled in their Personal Needs and Preferences profile). The next screens cover the question format and required materials, followed by in-depth walkthrough of an ELA testlet. This walkthrough includes an example engagement activity, an example transition screen, and example sets of questions requiring specific educator-student interactions, requiring materials, using two screens, and using multiple choice with images. After the ELA testlet, there is an in-depth walkthrough of a mathematics testlet that includes mathematics questions that use two screens and mathematics questions requiring student demonstration. The final screen in the educator practice activity explores the answer summary page.

The student practice activity is formatted like a computer-delivered testlet. The first screen in the activity explains the navigation icons in the online testing platform. The second screen explains the two types of testlets that the student will encounter (ELA and mathematics). Next, the different types of questions (e.g., multiple choice, matching, sorting) are explained. Following each explanation, there is a practice question so that students can familiarize themselves with how to respond to each type of question. Two multiple choice questions are used, one that requires one answer selection and one that allows multiple answer selections. The final screens explain the answer summary page indicating if all the questions were answered and giving students a chance to navigate back to the testlet.

A released testlet is a publicly available sample DLM assessment. Released testlets cover the same content and are in the same format as operational DLM testlets. Students and educators can use released testlets as examples or opportunities for practice. Released testlets are developed using the same standards and methods used to develop testlets for the DLM operational assessments. New released testlets are added on a yearly basis. Released testlets are selected from a variety of EEs and linkage levels across Grades 3–12.

Each year, six testlets are selected for release from each subject based on several criteria. Criteria include providing testlets from across all grades/grade bands and linkage levels, that contain three to five items, measure EEs that are assessed in both administration models, and are useful for instruction.

In response to state inquiries about supplemental assessment resources to address the increase in remote or disrupted instruction due to COVID-19, DLM staff published additional ELA and mathematics released testlets during the spring 2020 window. Across all subjects, nearly 50 new released testlets were selected and made available through the online testing platform. To help parents and educators better review the available options for released testlets, DLM staff also provided tables for each subject that display the EEs and linkage levels for which released testlets are available.

The test development team selected new released testlets that would have the greatest impact for remote or disrupted instruction. The team prioritized testlets at the Initial Precursor, Distal Precursor, and Proximal Precursor linkage levels, as those linkage levels are used by the greatest number of students. The test development team selected testlets written to EEs that covered common instructional content, with a consideration for previously released testlets to minimize overlap between the testlets that were already available and new released testlets. The test development team also aimed to provide at least one new released testlet per grade level, where possible.

3.3. Test Development Procedures

This section covers test development principles and the test development process, including item writing, external reviews, and internal procedures for preparing test content for administration.

3.3.1. Test Development Principles

The DLM System uses a variant of ECD to develop processes for item and test development. ECD describes a conceptual framework for designing, developing, and administering educational assessments (Mislevy et al., 1999). The ECD framework supports the creation of well-constructed tests that are valid for their intended purposes by "explicating the relationships among the inferences the assessor wants to make about the student, what needs to be observed to provide evidence for those inferences, and what features of situations evoke that evidence" (Mislevy et al., 1999, p. 1). ECD requires test designers to define the relationships between inferences that they want to make about student skills and understandings and the tasks that can elicit evidence of those skills and understandings in the assessment. The ECD approach is structured as a sequence of test development layers that include (a) domain analysis, (b) domain modeling, (c) conceptual assessment framework development, (d) assessment implementation, and (e) assessment delivery (Mislevy & Riconscente, 2005). Since the original introduction of ECD, the principles, patterns, examples, common language, and knowledge representations for designing, implementing, and delivering educational assessment using the processes of ECD have been further elaborated for alternate assessment (DeBarger et al., 2011; Flowers et al., 2015).

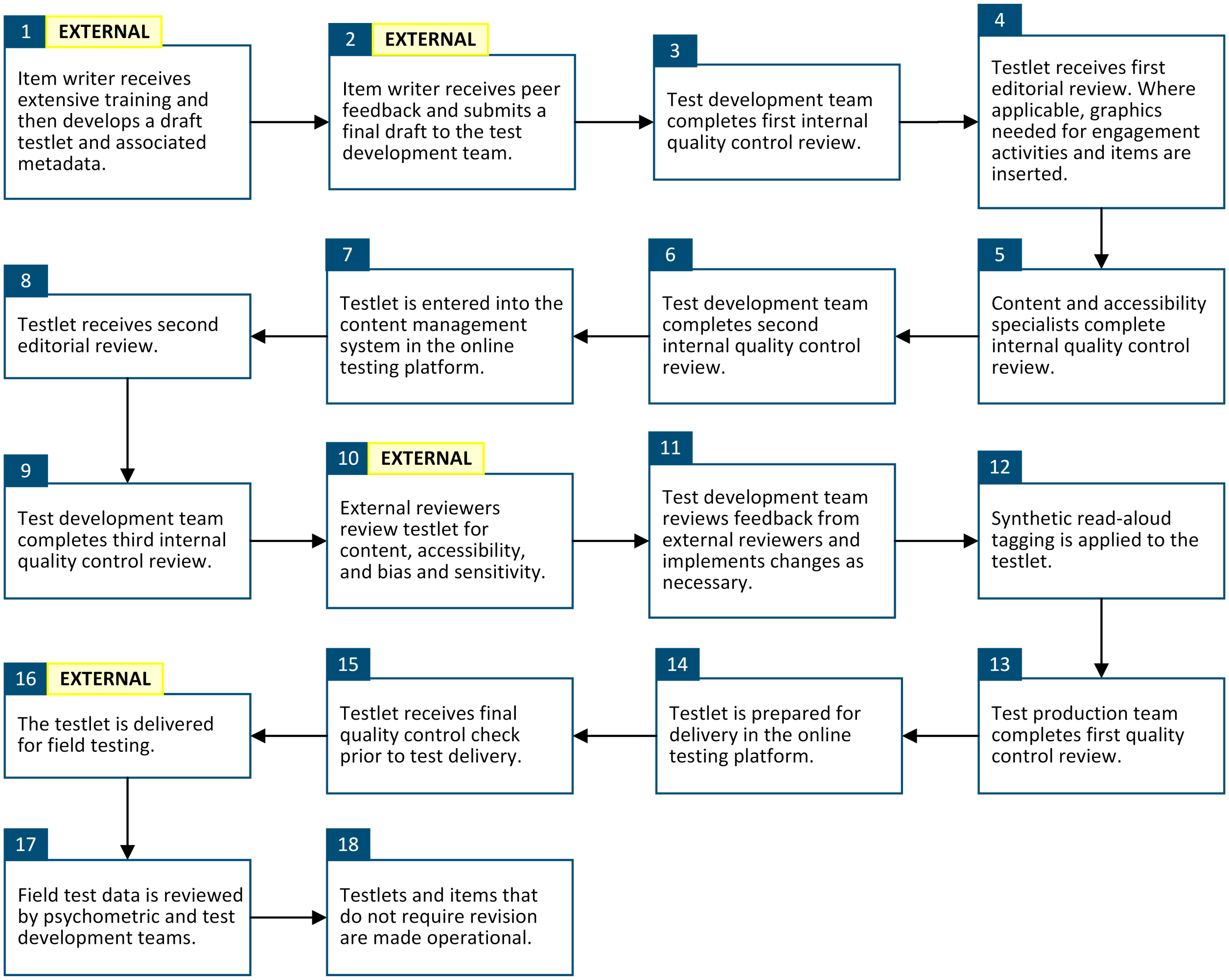
The DLM System uses ECD procedures to develop test specifications and task templates for item creation that also incorporate UDL principles (Bechard et al., 2019). Incorporating principles of UDL allows students to respond to items free of barriers while emphasizing accessibility and offering multiple ways to demonstrate skills. The DLM task templates are called Essential Element Concept Maps (EECMs) and are described in section 3.3.3.1.3.

3.3.2. Overview of the Testlet Development Process

Items are developed by highly qualified item writers from across DLM states. After extensive training (see section 3.3.3.1.2 below), item writers draft testlets and receive peer feedback during item writing. The items undergo internal quality reviews and editorial reviews in several iterations before being reviewed externally by panels for content, bias and sensitivity, and accessibility for students with the most significant cognitive disabilities. After external reviews, testlets are prepared for field testing. Items are field tested by DLM students prior to being promoted to the operational pool of content. The full set of test development steps are outlined in Figure 3.17.

Figure 3.17

Steps in the Test Development Process



Each item writer is paired with another item writer who is assigned the same grade and subject. The item writer creates a first draft of the item. Item writers provide peer feedback to each other once first drafts are complete, and they take that feedback into account when completing their final draft. Item writers then submit final drafts to the test development team. The test development team conducts the first internal quality control review, which includes checks for alignment of the content to map nodes. Staff then conduct an editorial review and create any necessary images.

Staff internally review testlets for content, accessibility, language, and bias and sensitivity and address any feedback. The testlets are then externally reviewed by panels of educators for content, accessibility, and bias and sensitivity (see section 3.3.5). The test development team synthesizes and addresses this feedback. Items are prepared for field-test delivery, which includes adding synthetic read loud and preparing test delivery resources (see Chapter 4 of this manual for a description of Testlet Information Pages).

Test production, test development, and psychometric teams then complete a final quality control check for accessibility, display, content, and associated test delivery resources. Testlets are scheduled for field testing (see section 3.6.1). Following field-test data collection, staff review the field-test data to determine which testlets meet quality standards and are ready for operational assessment.

Security of materials is maintained throughout the test development process. Paper materials are kept in locked facilities. Electronic transfers are made on a secure network drive or within the secure content management system in the online testing platform.

3.3.3. Testlet and Item Writing

Highly qualified item writers are recruited annually from across DLM states to develop DLM items and testlets. Item writers are recruited based a multitude of characteristics, including teaching experience and experience with the DLM alternate assessment, with priority given to those with subject matter expertise, population expertise, and prior DLM item-writing experience. The number of item writers per year is determined based on the number of testlets needed, with a range of 4 to 117 participants per year from 2013–2022. Item writers are typically assigned one EE and linkage level to write for per round, with a range of 5 to 14 rounds across years.

3.3.3.1. 2022 Testlet and Item Writing

Highly qualified, experienced item writers were selected to participate in a 2-day virtual item-writing event that was held on January 25–26, 2022. Item writer training included instruction on the item-writing process and peer review process. During the event, item-writing pairs collaborated and began to develop testlets. Following the virtual event, item writers continued producing and peer reviewing testlets virtually via a secure online platform through June 2022. A total of 427 testlets were produced, including 227 ELA testlets and 200 mathematics testlets.

3.3.3.1.1. Participants

Item writers were selected from the Accessible Teaching, Learning, and Assessment Systems (ATLAS) MemberClicks database. The database is a profile-based recruitment tool hosted in MemberClicks. The MemberClicks database includes individuals actively recruited via recommendations from governance board members and social media, individuals who have previously participated in an event, and individuals who created profiles via the “sign up to participate in DLM events” link on the DLM homepage. Interested individuals create and update their participant profile. Participant profiles include demographic, education, and work experience data.

A total of 245 individual profiles were pulled from the ATLAS MemberClicks Database for 2022 item writing. Minimum eligibility criteria included at least 1 year of teaching experience, teaching in a DLM state, and experience with the DLM alternate assessment. Prior DLM event participation, subject matter expertise, population expertise, and distribution of experience in each grade band was also considered in selection and assignment to a subject area. Of the 245 individuals, 77 individuals registered, completed advanced training, and committed to attend the event. In total, 69 item writers from 16 states attended both days of the training event and at least rounds 1 and 2 of the item-writing event. Of these item writers, 39 developed ELA testlets and 30 developed mathematics testlets.

The median and range of years of teaching experience is shown in Table 3.1. The median years of experience was at least 12 years for item writers of both ELA and mathematics testlets in pre-K–12.

Table 3.1

Item Writers' Years of Teaching Experience

Teaching experience	Median	Range
Pre-K–12	15.0	2–37
English language arts	12.5	0–28
Mathematics	12.0	1–28

Grade 6 was most commonly taught by item writers (n = 34; 49%). See Table 3.2 for a summary.

Table 3.2

Item Writers' Grade-Level Teaching Experience

Grade level	*n*	%
Grade 3	29	42.0
Grade 4	33	47.8
Grade 5	32	46.4
Grade 6	34	49.3
Grade 7	32	46.4
Grade 8	32	46.4
High school	25	36.2

The 69 item writers represented a highly qualified group of professionals representing both content and special education perspectives. The level and most common types of degrees held by item writers are shown in Table 3.3 and Table 3.4, respectively. The item writers with complete MemberClicks profiles held at least a bachelor's degree. A majority (n = 59; 86%) also held a master's degree, for which the most common field of study was special education (n = 28; 41%).

Table 3.3

Item Writers' Level of Degree

Degree	*n*	%
Bachelor's	9	13.0
Master's	59	85.5
Not Specified	1	1.4

Table 3.4

Item Writers' Degree Type

Degree	*n*	%
Bachelor's degree		
Education	19	27.5
Content specific	1	1.4
Special education	14	20.3
Other	28	40.6
Missing	6	8.7
Not specified	1	1.4
Master's degree		
Education	17	28.8
Content specific	4	6.8
Special education	28	47.5
Other	9	15.3
Missing	1	1.7

Most item writers had experience working with students with disabilities (93%), and 97% had experience with the administration of alternate assessments. The variation in percentages suggest some item writers may have had experience with administration of alternate assessments but perhaps did not regularly work with students with disabilities.

Item writers reported a range of experience working with students with different disabilities, as summarized in Table 3.5. Item writers collectively had the most experience working with students with a significant cognitive disability, a mild cognitive disability, or multiple disabilities.

Table 3.5

Item Writers' Experience with Disability Categories

Disability category	*n*	%
Blind/low vision	31	44.9
Deaf/hard of hearing	32	46.4
Emotional disability	42	60.9
Mild cognitive disability	53	76.8
Multiple disabilities	54	78.3
Orthopedic impairment	30	43.5
Other health impairment	50	72.5
Significant cognitive disability	53	76.8
Specific learning disability	47	68.1
Speech impairment	42	60.9
Traumatic brain injury	30	43.5

The professional roles reported by the 2022 item writers are shown in Table 3.6. Roles included educators, instructional coaches, district staff, state education agency staff, and other (i.e., university staff, program coordinators, supervisors of instruction).

Table 3.6

Professional Roles of Item Writers

Role	*n*	%
Classroom educator	43	62.3
Other	17	24.6
Instructional coach	5	7.2
District staff	3	4.3
State education agency staff	1	1.4

Among the ELA and mathematics item writers, 16 DLM partner states were represented. ELA item writers were from 13 different states and the District of Columbia and mathematics item writers were from 13 different states. Population density of schools in which item writers taught or held a position is reported in Table 3.7. Rural was defined as a population living outside settlements of 1,000 or fewer inhabitants, suburban was defined as an outlying residential area of a city of 2,000–49,000 or more inhabitants, and urban was defined as a city of 50,000 inhabitants or more. The demographics for the item writers are presented in Table 3.8.

Table 3.7

Population Density for Schools of Item Writers

Population density	*n*	%
Rural	30	43.5
Suburban	21	30.4
Urban	18	26.1

Table 3.8

Demographics of the Item Writers

	n	%
Gender		
Female	65	94.2
Male	3	4.3
Chose not to disclose	1	1.4
Race		
White	64	92.8
African American	3	4.3
Chose not to disclose	2	2.9
Hispanic ethnicity		
Non-Hispanic	64	92.8
Hispanic	3	4.3
Chose not to disclose	2	2.9

3.3.3.1.2. Item Writer Training

Training for item writing consisted of independent asynchronous advance training via a secure online platform as well as synchronous training on the first day of the virtual event. The advance training modules consisted of an overview module focused on the DLM assessment system and population of students, subject-specific information related to ELA or mathematics, and information on UDL, accessibility considerations, and bias and sensitivity considerations. There was a brief posttest at the end of each module that item writers were required to pass with 80% accuracy (item writers were allowed to take the quiz as many times as necessary to reach the 80% requirement). The virtual event training consisted of targeted instruction regarding the structure and development of DLM ELA and mathematics testlets and items from an ECD perspective, including information on accessibility and bias and sensitivity considerations.

3.3.3.1.3. Essential Element Concept Maps

Item and testlet writing are based on EECMs. These graphic organizers are provided as guides to item writers. EECMs use principles of ECD and UDL to define ELA and mathematics content specifications for assessment development. For more information about ECD, see section 3.3.1 of this chapter.

ELA and mathematics test development teams developed the EECMs (Bechard & Sheinker, 2012). Staff with student population expertise also reviewed EECM contents. The templates were specifically designed for clarity and ease of use, as the project engages nonprofessional item writers from states administering DLM assessments who needed to create a large number of items in a constricted timeframe.

The EECMs provide item writers with a content-driven guide on how to develop content-aligned and accessible items and testlets for the DLM student population. Each EECM defines the content framework of a target EE with five levels of complexity (i.e., linkage levels) and identifies key concepts and vocabulary at each level. Developers selected nodes from the learning maps to be assessed at different linkage levels based on an analysis of the map structure. The EECM also describes and defines common misconceptions, common questions to ask, and prerequisite and requisite skills. Finally, the EECM identifies accessibility issues related to particular concepts and tasks.

The EECM has seven functions:

- Identify the targeted standard by claim, conceptual area, Common Core State Standards, and EE;
- Identify key vocabulary to use in testlet questions;
- Describe and define a range of skill development (five levels);
- Describe and define misconceptions;
- Identify requisite and prerequisite skills;
- Identify questions to ask; and
- Identify content through the use of accessibility flags that may require an alternate approach to assessment for some students.

An example EECM that was used for item development is shown in Figure 3.18.

Figure 3.18

Example Essential Element Concept Map Graphic Organizer for ELA.EE.RI.6.2

Claim: ELA.C1 Students can comprehend text in increasingly complex ways.
Conceptual Area: ELA.C1.2 Construct Understandings of Text
General Education Content Standard: ELA.RI.6.2 Determine a central idea of a text and how it is conveyed through particular details; provide a summary of the text distinct from personal opinions or judgments.
Essential Element: ELA.EE.RI.6.2 Determine the main idea of a passage and details or facts related to it.

DYNAMIC® LEARNING MAPS

Essential Questions
- Can the student identify the main idea of a passage?
- Does the student recognize that details and facts can relate to the main idea of a passage?

Vocabulary	(a) Initial Precursor	(b) Distal Precursor	(c) Proximal Precursor	(d) Target	(e) Successor
Concepts	environment, object and person identification, object/picture association	concrete detail identification	detail identification, key details	main idea, detail identification, main idea/detail association	central idea, key details, key detail/central idea association
Words	naming words (dog, ball, girl, etc.), wh words (who, what, which, where)	find, wh words (who, which, what, where, when)	find, wh words (who, which, what, where, when), detail	main idea, details, wh words (who, which, what, where, when)	important, detail, support, wh words (who, which, what, where, when), how, main idea

(a) Initial Precursor Nodes	Node Descriptions	Node Observations	# Items
F-154-Can demonstrate understanding of property words corresponding to the objects used during familiar routines	Can demonstrate a receptive understanding of the property words that describe the objects that accompany familiar games or routines.	During a shared reading activity with the student, the student is able to identify items based on their property descriptions.	3-5 ☒ TA ☐ Blind/VI (B)

(a) Questions to Ask
- Does the student recognize property words?
- Show me the (property word) one.

(a) Misconceptions
- The student indicates a different object.
- The student indicates multiple objects.
- The student attends to other stimuli.
- The student does not respond.

(b) Distal Precursor Node	Node Description	Node Observation	# Items
ELA-1141 Can identify concrete details in familiar informational texts	Can identify the concrete details, such as individuals, events, or ideas in familiar informational texts.	When asked to recall a concrete detail from a familiar informational text, the student is able to identify the correct detail from the text.	5 ☐ TA ☐ Blind/VI (B)

(b) Questions to Ask
- Does the student recognize that informational texts contain concrete details?
- Can the student identify the correct detail to answer a question?
- Who is John?

(b) Misconceptions
- The student chooses a detail unrelated to the particular question.
- The student attempts to use the illustration to answer a question about a concrete detail rather than the text

(c) Proximal Precursor Node	Node Description	Node Observation	# Items
ELA-1462 Can identify the key details in a paragraph of an informational text	Can determine which details in a paragraph of an informational text are important.	After reading an informational text, the student can identify each of the key details in the paragraph.	5 ☐ TA ☐ Blind/VI (B)

(c) Questions to Ask
- Can the student identify key details in text?
- Who sailed across the ocean?
- Who crossed the river?

(c) Misconceptions
- The student identifies details that are only of minor importance to the paragraph.
- The student identifies details that are not from the paragraph.

(d) Target Node	Node Description	Node Observation	# Items
ELA-1463 Can identify the key details that support the main idea of a paragraph in an informational text	Can determine which details contained within a paragraph of an informational text provides an important contribution to the paragraph's main idea.	After reading an informational text, the student is able distinguish between key details that support the paragraph's main idea and details from the paragraph that are of less importance.	5 ☐ TA ☐ Blind/VI (B)

(d) Questions to Ask
- Does the student recognize that key details in a text relate to and support the main idea?
- Which sentence supports the main idea that rainforests should be saved?

(d) Misconceptions
- The student chooses details from the text that are unrelated to the main idea of the text.
- The student is unable to identify the main idea of the text.

(e) Successor Node	Node Description	Node Observation	# Items
ELA-973 Can identify the key details that support the main ideas of an informational text	Can determine which key details in an informational text support the main idea of the whole text or a section of it.	After reading an informational text, the student can determine the main ideas of the text, and can distinguish between key details that support the main ideas and details from the text that are of less importance.	5 ☐ TA ☐ Blind/VI (B)

(e) Questions to Ask
- Does the student recognize that some details are more significant to the central idea of an informational text than others?
- Why does the owl hunt at night?
- Why does the cat have good night vision?

(e) Misconceptions
- The student chooses minor details from the text that are unrelated to the main ideas of the text.
- The student chooses key details that are unrelated to the main ideas of the text.

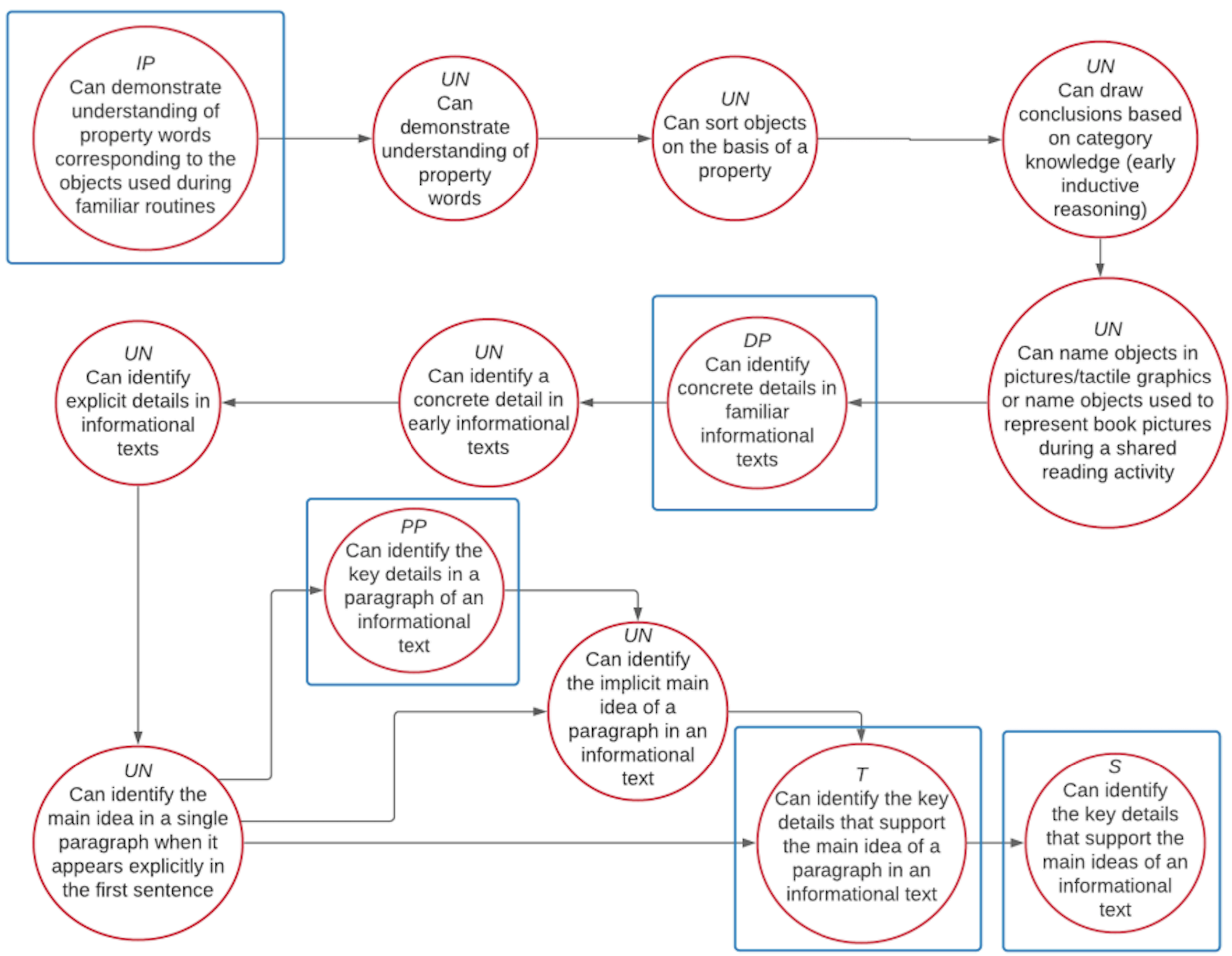
Next Essential Element

ELA.EE.RI.7.2 Determine two or more central ideas in a text.

In addition to text descriptions, EECMs include a small view of the nodes associated with the EE. These mini-maps were provided as a visual means of formally identifying the relationships between skills so that item writers would be able to consider them during the design of testlets. Figure 3.19 shows an example of a mini-map used during test development.

Figure 3.19

Example Mini-Map for ELA.EE.RI.6.2—Determine the main idea of a passage and details or facts related to it



IP = Initial Precursor; DP = Distal Precursor; PP = Proximal Precursor; T = Target; S = Successor; UN = Untested. Boxes indicate tested nodes.

3.3.3.1.4. Other Item-Writing Resource Materials

In addition to the EECMs, item writers used material developed by test development teams to support the creation of testlets. All item writers used the *DLM Core Vocabulary List*. Core vocabulary is made up of words used most commonly in expressive communication (Yorkston et al., 1988). DLM Core Vocabulary is a comprehensive list of words that reflects the research in core vocabulary in Augmentative and Alternative Communication and words needed to successfully communicate in academic settings when the EEs are being taught (Dennis et al., 2013). Additionally, all item writers used an item-writing manual containing general information about DLM testlets and items, instructions for populating testlet templates, and subject-specific information related to such topics as the order of testlet elements and item types unique to each subject (e.g., select text for ELA or matching lines for mathematics). Item writers were also provided with examples of completed testlet templates in both subjects and instructions on testlet development for students who are blind or have visual impairments. The ELA test development team also shared guides to writing texts.

3.3.3.1.5. Item-Writing Process

Because DLM assessments consist of short, coherent, instructionally relevant testlets, item writers produced entire testlets rather than stand-alone items. Item writers reviewed the vocabulary (concepts and words) on the EECM appropriate for each testlet level. Item writers assumed that students were expected to understand, but not necessarily use, these terms and concepts. Item writers were also responsible for writing testlets at increasing levels of complexity, from less complex to more complex linkage levels, when they were writing across linkage levels. Using the EECMs, item writers selected specific vocabulary for the testlet that matched the cognitive complexity of the node(s) being assessed at that linkage level.

Item writers used the EECM “questions to ask” and “misconceptions” information when writing testlets. The questions describe what evidence is needed to show that the student can move from one level to the next, more complex level, and the possible misconceptions or errors in thinking that could be a barrier to students demonstrating their understanding. These EECM sections assisted the item writers to create stems (i.e., the item prompts) and answer options for items in testlets.

Item writers were instructed to write testlet content to be accessible for all students who might receive each testlet. The goal for the item writer was to create testlets that were accessible to the greatest number of students possible. Item writers were also directed to consider barriers that may be present due to the sensitive nature of the content or bias that may occur, which could advantage or disadvantage a particular subgroup group of students.

During item development, item writers and DLM staff maintained the security of materials. Item writers signed security agreements.

3.3.3.1.6. Item Writer Evaluations

An evaluation survey of the item-writing experience was sent to all participating item writers after the item-writing event. Item writers were asked to provide feedback on the perceived effectiveness of training and the overall experience of the item-writing event, as well as narrative comments on their experience and suggestions for future DLM item-writing events.

The majority of item writers responded to the post-event survey (n = 33 for ELA, n = 23 for mathematics). Table 3.9 and Table 3.10 detail responses to the perceived effectiveness questions from the survey for ELA and mathematics item writers, respectively.

Table 3.9

*Perceived Effectiveness of Training, English Language Arts Item Writers (*n *= 33)*

	Very effective		Somewhat effective		Not at all effective	
Degree	*n*	%	*n*	%	*n*	%
Advanced training	26	78.8	7	21.2	0	0.0
Virtual training	26	78.8	7	21.2	0	0.0
Practice activities	26	78.8	6	18.2	1	3.0
Online resources	31	93.9	1	3.0	1	3.0
Discussion with other item writers	26	78.8	7	21.2	0	0.0
Feedback from DLM staff	28	84.8	5	15.2	0	0.0

Table 3.10

*Perceived Effectiveness of Training, Mathematics Item Writers (*n *= 23)*

	Very effective		Somewhat effective		Not at all effective	
Degree	*n*	%	*n*	%	*n*	%
Advanced training	17	73.9	6	26.1	0	0.0
Virtual training	15	65.2	8	34.8	0	0.0
Practice activities	17	73.9	6	26.1	0	0.0
Online resources	22	95.7	1	4.3	0	0.0
Discussion with other item writers	20	87.0	2	8.7	1	4.3
Feedback from DLM staff	23	100.0	0	0.0	0	0.0

Overall, item writers rated the training they received, the practice activities, the online resources, the feedback from the DLM staff, and the discussion with other item writers as very effective. They expressed appreciation of the knowledge gained through the item-writing event and the opportunities to collaborate with peers.

3.3.4. ELA Text Development

This section describes the development of ELA texts. After these texts undergo text-specific external review (see section 3.3.5.2), they are incorporated into testlets that are externally reviewed through the standard testlet review process.

3.3.4.1. Original Development of Texts

The test development team originally created ELA texts by adapting from or relating to grade-level–appropriate general education texts. The team constructed short narrative texts from books commonly taught in general education and wrote short informational texts to relate to thematic elements from narratives. The team deliberately wrote all texts to provide an opportunity to assess specific nodes in the maps associated with different EEs and linkage levels. They reduced text complexity in the ELA texts from the grade-level texts for students without significant cognitive disabilities. The team also developed texts using clear language that reduced any need for prior knowledge. They favored simple sentences, reduced the use of pronouns, and favored consistency in sentence structure within a text. DLM texts are short and consist of 50–250 words, including high-frequency, easily decodable words, such as those found on the *DLM Core Vocabulary List*.

3.3.4.2. Recent Development of Texts

New ELA texts were most recently developed in 2019–2020. Throughout 2019–2020, the ELA test development team created new reading literature and reading informational texts. To determine the number and types of new texts needed, EEs and linkage levels were identified in which two or fewer texts were available, and/or available texts could not support five items per testlet. Following the identification of needed ELA texts, the specifications for each new text to be developed—including nodes, text type (reading literature or reading informational text), and the exemplar source book—were identified.

ELA test development staff wrote, selected images for, and peer reviewed each text. Thirty new texts were written. The number and types of texts created for each grade are summarized in Table 3.11. Just over half of the new ELA texts (n = 16; 53%) were reading informational texts. There were 14 (47%) new reading literature texts written.

Table 3.11

Text Needs by Grade and/or Grade Band and Text Type

Grade/Grade band	Reading literature	Reading informational	Total
3	2	2	4
4	2	2	4
5	2	2	4
6	2	2	4
7	2	2	4
8	2	2	4
9–10	0	2	2
11–12	2	2	4
Total	*14*	*16*	*30*

Following review by test development staff, the texts and images were reviewed by Special Education experts and senior DLM staff to evaluate whether each ELA text met DLM guidelines and supported the assigned EEs, linkage levels, and nodes. The ELA texts received an editorial review before being finalized for external review.

3.3.5. External Reviews

3.3.5.1. Items and Testlets

The purpose of external reviews of items and testlets is to evaluate whether items and testlets measure the intended content, are accessible, and are free of bias or sensitive content. Panelists use external review criteria established for DLM alternate assessments to rate items and testlets as “accept”, “revise”, or “reject” and provide recommendations for “revise” ratings or an explanation for “reject” ratings. The test development team uses collective feedback from the panelists to inform decisions about items and testlets before they are field-tested.

The external review process was piloted in a face-to-face meeting in Kansas City, Missouri, in August 2013 before being implemented in the secure, online content management system in the testing platform. Educators nominated by DLM Governance Board members, including several governance board members themselves, participated as panelists. The pilot event was used to evaluate the effectiveness of reviewer training, clarity and appropriateness of the review criteria for each panel type, and the options available for rating and providing feedback on items and testlets.

Subsequent external review events have followed the structure established in the initial pilot and minimally revised over the operational years of the program. Panelists receive training regarding the review criteria and structure of the DLM assessment, and they continue to complete reviews using the online content management system in the testing platform.

3.3.5.1.1. Overview of Review Process

External review for 2021 was held as a 2-day virtual event. The virtual advance training, training for the panel meetings, and facilitator and co-facilitator training were updated to meet the needs of virtual panel meetings. Previously, one facilitator led the feedback discussion for each panel, whereas for the virtual event, a facilitator and co-facilitator led the feedback discussions and recorded decisions for each panel meeting.

External reviews were conducted by members of three distinct review panels: content, accessibility, and bias and sensitivity. Reviewers were assigned to one type of review panel based on their expertise and used the criteria for that panel to conduct reviews. For each item and each testlet, reviewers made one of three decisions: “accept”, “revise”, or “reject.” Reviewers made decisions independently and without discussion with other reviewers before joining with the facilitator and co-facilitator for a consensus discussion and decision. Both independent and consensus reviews were completed using an application in the secure content management system in the online testing platform. Resulting ratings were compiled and submitted to DLM staff, and DLM staff made final decisions regarding whether the testlet should be rejected, accepted as is, or revised before field testing.

3.3.5.1.2. Review Assignments and Training

Panelists were selected from the ATLAS MemberClicks database and were assigned to content, accessibility, or bias and sensitivity panels based on their qualifications.

In 2021, there were 47 panelists. Of those, 16 were content-specific panelists; 8 were ELA panelists and 8 were mathematics panelists. There were also 16 accessibility panelists and 15 bias and sensitivity panelists who reviewed items and testlets from each subject.

Prior to participating in the virtual panel meetings, panelists completed an advance training course that included an External Review Procedures module and a module that specifically aligned to their assigned panel type. The content modules were subject-specific, while the accessibility and bias and sensitivity modules were universal for all subjects. The content modules described the review criteria for items and testlets and included examples to illustrate the concepts. The accessibility module covered accessibility criteria, including examples for items and testlets. The item accessibility criteria specifically focused on accessibility of text and graphics. The testlet accessibility criteria specifically focused on instructional relevance and ensuring that the testlet is barrier-free. The bias and sensitivity module covered item bias and testlet sensitivity criteria. The item bias criteria included items having a fair construct, a representation of diversity, using people-positive language, avoiding language bias, and avoiding content likely to cause an extreme emotional response. The testlet sensitivity criteria included testlets avoiding sensitive content and language bias. Examples were provided to illustrate the concepts covered in the bias and sensitivity module. After each module, panelists completed a posttest and were required to score 80% or higher to pass advance training; panelists could attempt the posttest as many times as necessary to reach a score of 80%. At the beginning of the virtual event, facilitators reviewed the procedures for how panelists would complete their ratings. Then facilitators reviewed panel-specific information, based on the panel type. This included a security and confidentiality reminder, background information, and an overview of the panel-specific criteria.

During the virtual event, panelists first completed asynchronous reviews of a small calibration set of testlets, followed by a synchronous review of the calibration set focused on items and testlets flagged during the asynchronous review. Subsequent collections of testlets were larger but adhered to this same process of asynchronous reviews followed by synchronous reviews of items and testlets flagged during the asynchronous review. Each panel had two virtual panel meetings led by facilitators and co-facilitators to obtain collective feedback about the items and testlets. Content panels had fewer testlets per collection because these panels reviewed only subject-specific testlets, whereas the bias and sensitivity and accessibility panels had more testlets per collection because these panelists reviewed testlets from all subjects.

The median and range of years of teaching experience is shown in Table 3.12. The median years of experience for external reviewers was 15 years in pre-K–12, 13 years in ELA, and 11 years in mathematics.

Table 3.12

External Reviewers' Years of Teaching Experience

Teaching experience	Median	Range
Pre-K–12	15.0	5–38
English language arts	13.0	2–38
Mathematics	11.0	1–35

High school was the most commonly taught grade level by the external reviewers (n = 42; 33%). See Table 3.13 for a summary.

Table 3.13

External Reviewers' Grade-Level Teaching Experience

Grade level	*n*	%
Grade 3	10	21.3
Grade 4	9	19.1
Grade 5	12	25.5
Grade 6	16	34.0
Grade 7	20	42.6
Grade 8	20	42.6
High school	42	89.4

The 47 external reviewers represented a highly qualified group of professionals. The level and most common types of degrees held by external reviewers are shown in Table 3.14 and Table 3.15, respectively. A majority (*n* = 42; 89%) also held a master's degree, for which the most common field of study was special education (*n* = 13; 28%).

Table 3.14

External Reviewers' Level of Degree

Degree	*n*	%
Bachelor's	5	10.6
Master's	42	89.4

Table 3.15

External Reviewers' Degree Type

Degree	*n*	%
Bachelor's degree		
Education	13	27.7
Special education	11	23.4
Other	20	42.6
Missing	3	6.4
Master's degree		
Education	10	23.8
Content specific	3	7.1
Special education	13	31.0
Other	16	38.1

Most external reviewers had experience working with students with disabilities (74%), and 74% had experience with the administration of alternate assessments.

External reviewers reported a range of experience working with students with different disabilities, as summarized in Table 3.16. External reviewers collectively had the most experience working with students with a significant cognitive disability, specific learning disability, other health impairments, or multiple disabilities.

Table 3.16

External Reviewers' Experience with Disability Categories

Disability category	*n*	%
Blind/low vision	13	27.7
Deaf/hard of hearing	10	21.3
Emotional disability	24	51.1
Mild cognitive disability	24	51.1
Multiple disabilities	30	63.8
Orthopedic impairment	15	31.9
Other health impairment	27	57.4
Significant cognitive disability	27	57.4
Specific learning disability	29	61.7
Speech impairment	25	53.2
Traumatic brain injury	16	34.0

Panelists had varying experience teaching students with the most significant cognitive disabilities. ELA panelists had a median of nine years of experience teaching students with the most significant cognitive disabilities, with a minimum of four years and a maximum of nine years of experience. Mathematics panelists had a median of nine years of experience teaching students with the most significant cognitive disabilities, with a minimum of four years and a maximum of nine years of experience.

The professional roles reported by the 2021 reviewers are shown in Table 3.17. Panelists who reported "other" roles included administrators, program directors, assessment coordinators, and individuals identifying with multiple categories.

Table 3.17

Professional Roles of External Reviewers

Role	*n*	%
Classroom educator	36	76.6
Instructional coach	2	4.3
Other	6	12.8
State education agency staff	2	4.3
Not specified	1	2.1

Among the ELA and mathematics panelists, nine DLM partner states were represented. ELA panelists were from seven different states and the District of Columbia and mathematics panelists were from three different states. Population density of schools in which panelists taught or held a position is reported in Table 3.18. Rural was defined as a population living outside settlements of 1,000 or fewer inhabitants, suburban was defined as an outlying residential area of a city of 2,000–49,000 or more inhabitants, and urban was defined as a city of 50,000 inhabitants or more. The demographics for the external reviewers are presented in Table 3.19.

Table 3.18

Population Density of School of English Language Arts and Mathematics Content Panelists

Population density	*n*	%
Rural	19	40.4
Suburban	12	25.5
Urban	16	34.0

Table 3.19

Demographics of the External Reviewers

	n	%
Gender		
Female	36	76.6
Male	11	23.4
Race		
White	36	76.6
African American	5	10.6
Asian	2	4.3
American Indian	1	2.1
Native Hawaiian or Pacific Islander	1	2.1
Chose not to disclose	2	4.3
Hispanic ethnicity		
Non-Hispanic	44	93.6
Hispanic	2	4.3
Chose not to disclose	1	2.1

3.3.5.1.3. Review Responsibilities

The primary responsibility for reviewers was to review items and testlets using established standards and guidelines. These standards and guidelines are found in the *Guide to External Review of Testlets* (DLM Consortium, 2014). Reviewers completed a security agreement before reviewing and were responsible for maintaining the security of materials at all times.

3.3.5.1.4. Decisions and Criteria

External reviewers looked at testlets and made decisions about both the items in the testlet, and the testlet overall. An overview of the decision-making process is described below.

General Review Decisions. For DLM assessments, "acceptability" at the external review phase was defined as meeting minimum standards to be ready for field testing. Reviewers made one of three general decisions: "accept", "revise", or "reject." The definition of each decision is summarized in Table 3.20.

Table 3.20

General Review Decisions for External Reviews

Decision	Definition
Accept	Item or testlet is within acceptable limits. It may not be perfect, but it can be field tested.
Critical revision required (revise)	Item or testlet violates one or more criteria. It has some potential merits and can be acceptable for field testing after revisions to address the criteria.
Reject	Item or testlet is fatally flawed. No revision could bring this item/testlet to within acceptable limits.

Judgments about items were made separately from judgments about testlets because different criteria are used for items and testlets. Therefore, it is possible to recommend revisions or rejections to items without automatically having to recommend revision or rejection to the testlet as a whole. If reviewers recommended revision or rejection, they were required to provide an explanation that included identification of the problem and, in the case of revision, a proposed solution.

Review Criteria. In all external reviews, the criteria for each type of panel (i.e., content, accessibility, bias and sensitivity) were different. All three panel types had criteria to consider for items and other criteria for testlets as a whole. Training on the criteria was provided in the online training modules and in the on-site training. There were specific criteria for external reviewers of content, accessibility, and bias and sensitivity.

The content, accessibility, and bias and sensitivity review criteria are presented in Table 3.21, Table 3.22, and Table 3.23, respectively.

Table 3.21

Content Review Criteria

Criteria
Item
1. The item assesses the content of the targeted node.
2. The level of cognitive process dimension (CPD) required in the node matches the CPD identified for the item. CPD is listed as Cognitive Category in Content Builder.
3. The content of the item is technically correct (wording and graphics).
4. Item answer options contain only one correct answer (the key), distractors are incorrect and not misleading, and nothing in the item cues the correct response.
5. The item type is logical and appropriate for the assessed content and the graphics (if used) contribute to the quality of the item.
Testlet
1. The testlet is instructionally relevant to students for whom it was written and is grade-level appropriate.
2. Embedded items appear within the text at logical places and conclusion items appear at the end (English language arts only).

Table 3.22

Accessibility Review Criteria

Criteria
Item
1. The text within the item provides an appropriate level of challenge and maintains a link to grade-level content without introducing unnecessary, confusing, or distracting verbiage. The text uses clear language and minimizes the need for inferences or prior knowledge to comprehend the content.
2. Graphics are clear and do not cause confusion. It is possible to present graphics in tactile form and describe in alternate text.
Testlet
1. The testlet is instructionally relevant to students for whom it was written and is grade-level appropriate.
2. The testlet does not introduce barriers for students with (a) limited working memory, (b) communication disorders dependent on spoken English grammatical structures, or (c) limited implicit understandings of others' emotions and intentions.

Table 3.23

Bias and Sensitivity Review Criteria

Criteria
Item
1. The item does not require prior knowledge outside the bounds of the targeted content.
2. Where applicable, the item has a fair representation of diversity in race, ethnicity, gender, disability, and family composition.
3. The item avoids stereotypes. The item uses appropriate labels for groups of people and uses people-first language for individuals with disabilities.
4. Language used does not prevent or disadvantage any group from demonstrating what they know about the measurement target.
5. Item does not focus on material that is likely to cause an extreme emotional response.
Testlet
1. The testlet is free of content that is controversial, disturbing, or likely to cause an extreme emotional response due to issues of culture, region, gender, religion, ethnicity, socio-economic status, occupation, or current events.
2. The language in the testlet neither prevents nor disadvantages any regional or cultural group from demonstrating what they know about the targeted content. The testlet uses people-first language for individuals with disabilities and does not depict populations in a stereotypical manner.

3.3.5.1.5. Results of Reviews

For ELA, the percentage of items rated as accept across grades, panels, and rounds of review ranged from 70% to 99%. The percentage of testlets rated as accept across grades, panels, and rounds of review ranged from 66% to 96%. The percentage of items and testlets rated as revise ranged from 1% to 29% and 3% to 33%, respectively. The rate at which items and testlets were recommended for rejection ranged from 0% to <1% and 0%, respectively.

For mathematics, the percentage of items and testlets rated as accept ranged from 60% to 99% and 68% to 100%, respectively across grades, panels, and rounds of review. The percentage of items and testlets rated as revise ranged from 1% to 39% and 0% to 30%, respectively. The rate at which both items and testlets were recommended for rejection ranged from 0% to 1%.

3.3.5.1.6. Test Development Team Decisions

Because each item and testlet is examined by three distinct panels, ratings were compiled across panel types, following a process last updated in 2017–2018. The test development team reviewed the collective feedback provided by the panelists for each item and testlet. There are five decision options for the test development team to apply to each item and testlet: (a) accept, no pattern of similar concerns, accept as is; (b) revise minor, pattern of minor concerns, will be addressed; (c) revise major, major revision needed; (d) reject; and (e) more information needed. Once the test development team viewed each item and testlet and considered the feedback provided by the panelists, they assigned a decision. While panelist recommendations are carefully considered, the test development team does not make decisions solely on those recommendations.

The ELA test development team accepted as is 91% of items. Of the items that were revised, most required major changes (e.g., stem or response option replaced) as opposed to minor changes (e.g., minor rewording but concept remained unchanged). The ELA test development team made two (3%) minor revisions and 75 (97%) major revisions to items, and they rejected zero testlets.

The mathematics test development team accepted as is 47% of items. Of the items and testlets that were revised, most required major changes (e.g., stem or response option replaced) as opposed to minor changes (e.g., minor rewording but concept remained unchanged). The mathematics test development team made 72 (17%) minor revisions and 351 (83%) major revisions to items, and they rejected two testlets.

3.3.5.2. External Review of ELA Texts

The purpose of the external review of texts is to evaluate whether they are measuring the intended content, are accessible, are free of biased or sensitive content, and include appropriate imagery. Panelists also provide recommendations for revisions or an explanation for a “reject” rating. The ELA test development team uses the collective feedback from the panelists to inform decisions about texts and images before they are used in item and testlet development.

Following their finalization, an external review of texts for 2020 was conducted as a 2-day virtual event with panel meetings. There were four panels of between 3 and 6 individuals per panel. The facilitator and co-facilitator trainings were updated to meet the needs of virtual panel meetings, and the panelist advance training was revised to be more comprehensive. Panelists completed one advance training module designed to provide background information on the DLM alternate assessment and DLM ELA testlets and texts and received more rigorous training from DLM staff at the beginning of the event. Panelists completed independent reviews before a facilitator and co-facilitator led the feedback discussions and recorded decisions for each panel meeting.

3.3.5.2.1. Recruitment, Training, Panel Meetings, and Results

Panelists were selected from the ATLAS MemberClicks database based on predetermined qualifications for each panel type. Individuals first qualified by having more than 3 years of teaching experience, teaching in a DLM state, experience with the DLM alternate assessments, and having no item writer experience in 2019 or 2020. Potential panelists were sorted by grade and subject. Panelists were then assigned to content, accessibility, bias and sensitivity, or text image panels based on their qualifications. For example, those with experience teaching students with disabilities (other than students with significant cognitive disabilities) were prioritized to the accessibility panels, while those with experience teaching students with significant cognitive disabilities were prioritized to the bias and sensitivity panel.

In fall 2020, 57 panelists who had experience with ELA content and/or experience with students with significant cognitive disabilities were recruited to participate. Panelists represented 17 partner states. Three panelists did not indicate their state.

The median and range of years of teaching experience is shown in Table 3.24. The median years of experience for the ELA text panelists was 16 years in pre-K–12 and 15 years in ELA.

Table 3.24

English Language Arts Text Panelists' Years of Teaching Experience

Teaching experience	Median	Range
Pre-K–12	16.0	5–38
English language arts	15.0	2–43

The 57 ELA text panelists represented a highly qualified group of professionals. The level and most common types of degrees held by panelists are shown in Table 3.25 and Table 3.26, respectively. A majority (*n* = 53; 93%) held a master's degree, for which the most common field of study was special education (*n* = 21; 30%).

Table 3.25

English Language Arts Text Panelists' Level of Degree

Degree	*n*	%
Bachelor's	3	5.3
Master's	53	93.0
Other	1	1.8

Table 3.26

English Language Arts Text Panelists' Degree Type

Degree	*n*	%
Bachelor's degree		
Education	19	33.3
Content specific	1	1.8
Special education	6	10.5
Other	28	49.1
Not specified	3	5.3
Master's degree		
Education	13	24.5
Content specific	5	9.4
Special education	21	39.6
Other	14	26.4

ELA text panelists reported a range of experience working with students with different disabilities, as summarized in Table 3.27. ELA text panelists collectively had the most experience working with students with a significant cognitive disability, a mild cognitive disability, or multiple disabilities.

Table 3.27

English Language Arts Text Panelists' Experience with Disability Categories

Disability category	*n*	%
Blind/low vision	30	52.6
Deaf/hard of hearing	22	38.6
Emotional disability	38	66.7
Mild cognitive disability	40	70.2
Multiple disabilities	40	70.2
Orthopedic impairment	20	35.1
Other health impairment	38	66.7
Significant cognitive disability	40	70.2
Specific learning disability	38	66.7
Speech impairment	36	63.2
Traumatic brain injury	19	33.3

Panelists had varying experience teaching special education, with a median of 10 years of experience, a minimum of 3 years of experience, and a maximum of 30 years of experience.

The professional roles of the ELA text panelists are shown in Table 3.28. Roles include classroom educators, district staff members, state education staff agency, and other (i.e., instructional coach, item developer, university faculty).

Table 3.28

Professional Role of the English Language Arts Text Panelists

Role	*n*	%
Classroom educator	38	66.7
District staff member	1	1.8
State education agency staff	5	8.8
Other	13	22.8

Population density of schools in which panelists taught or held a position is reported in Table 3.29. Rural was defined as a population living outside settlements of 1,000 or fewer inhabitants, suburban was defined as an outlying residential area of a city of 2,000–49,000 or more inhabitants, and urban was defined as a city of 50,000 inhabitants or more. The demographics for the ELA text panelists are presented in Table 3.30.

Table 3.29

Population Density of School of the English Language Arts Text Panelists

Population density	*n*	%
Rural	28	49.1
Suburban	14	24.6
Urban	15	26.3

Table 3.30

Demographics of the English Language Arts Text Panelists

	n	%
Gender		
Female	52	91.2
Male	4	7.0
Chose not to disclose	1	1.8
Race		
White	49	86.0
Asian	2	3.5
African American	2	3.5
American Indian	1	1.8
Native Hawaiian or Pacific Islander	1	1.8
Chose not to disclose	2	3.5
Hispanic ethnicity		
Non-Hispanic	54	94.7
Hispanic	1	1.8
Chose not to disclose	2	3.5

Advanced training included panel-specific modules covering the external review criteria and a module containing information about DLM ELA text and testlets. The week before the event, panelists completed an initial review of the texts for their respective grade band assignments. The information covered in the advance training course laid the groundwork for the on-site training. The on-site training objectives were to understand the consensus discussion structure and process, the panel-specific criteria, and the resources used in the review.

Following the completion of the advance training module, the texts were provided to the panelists via a secure file-sharing platform. Panelists used the criteria for their assigned panel type to complete asynchronous reviews. Following asynchronous review, facilitators hosted panel meetings. Additional training on the structure and process of consensus discussions, panel-specific criteria, and resources were provided during the first panel meeting. During the panel meetings, panelists engaged in criteria-based discussion of each text to provide a consensus rating of the text. Panelists also made comments and suggested revisions to words and images. The co-facilitator recorded consensus ratings and recommendations for revision on text-rating sheets. In cases in which panelists recommended revisions, texts were revised to enhance language clarity, cohere with images, or better align with the text criteria. As shown in Table 3.31, panelists either rated texts as "Accept as is" or "Accept pending revisions." In some cases, revisions were made to texts that were rated as "Accept as is" by the panelists. This is because certain aspects are meant to be consistent across all texts (e.g., an edit to a character image that appears in multiple texts). Thus, if a revision is recommended for one text, a parallel edit may be made for other texts, even if the other text was rated as "Accept."

Table 3.31

Summary of Panel Ratings and Final Decisions

Panel ratings	Texts accepted as is	Texts revised per panel suggestion	Texts revised per DLM staff suggestion	Texts rejected per panel suggestion	Texts rejected per staff suggestion
Accept as is	4	0	13	0	0
Revise	0	26	0	0	0
Reject	0	0	0	0	0

Note. Texts can be revised per panel suggestion and per staff suggestion.

3.4. Alignment of Learning Map Nodes within a Linkage Level and Assessment Items

To ensure the developed testlets are assessing the intended construct, an external alignment study was conducted in 2014–2015 to evaluate the relationship between the DLM assessment contents and the assessed constructs. Briefly, ELA and mathematics testlets from the operational pool were evaluated for alignment with nodes by external partners. The primary measures of alignment were content and performance centrality. Content centrality is a measure of the degree of fidelity between the content of the learning map nodes within a linkage level and the assessment items. Panelists rated each pair as having no link, a far link, or a near link. Performance centrality represents the degree to which the operational assessment item and the corresponding academic grade-level content target contain the same performance expectation. The panelists rated the degree of performance centrality between each pair as none, some, or all. The external alignment study was updated in 2019–2020 to account for the adjusted Year-End model blueprint and the sole administration of single EE testlets.

This section provides a summary of findings from the external alignment study. Full results are provided in the separate technical report (Flowers & Wakeman, 2020).

Table 3.32 and Table 3.33 report the content and performance centrality ratings for the linkage level nodes to the assessment items, respectively. Overall, 96% of ELA and 100% of mathematics items were rated as having far or near content centrality to the corresponding linkage level. Similarly, the performance centrality ratings indicated that almost all items maintained the performance expectations found in the corresponding linkage level node.

Table 3.32

Content Centrality of Linkage Level Nodes to Assessment Items

Subject	Total *N*	No		Far		Near		Met*	
		n	%	*n*	%	*n*	%	*n*	%
English language arts	304	11	4	25	8	268	88	293	96
Mathematics	192	0	0	13	7	179	93	192	100

Note. Gray shading indicates acceptable level of alignment.

Table 3.33

Performance Centrality of Linkage Level Nodes to Assessment Items

Subject	Total *N*	None		Some		All		Met*	
		n	%	*n*	%	*n*	%	*n*	%
English language arts	304	8	3	33	11	263	87	296	97
Mathematics	186	0	0	6	3	180	97	186	100

Note. Gray shading indicates acceptable level of alignment.

Table 3.34 shows the percentage of items rated with each of the CPDs used by the DLM assessments for ELA and mathematics. Most ELA items were rated at the respond or understand level, and the mathematics items rated mostly at the remember to analyze CPD levels. Most items were located in the middle of the CPD distribution. These results suggest that the items cover a wide range of cognitive complexity.

Table 3.34

Cognitive Process Dimensions for English Language Arts and Mathematics Items

Cognitive Process Dimension	English language arts % (*n* = 278)	Mathematics % (*n* = 248)
Pre-intentional	0	0
Attend	2	0
Respond	26	0
Replicate	0	2
Remember	24	20
Understand	47	34
Apply	1	25
Analyze	1	15
Evaluate	0	4
Create	0	0

3.5. Evidence of Students’ Response Process

A cognitive lab study was conducted in 2014 to better understand how students interact with technology-enhanced items. The study focused on students’ experience engaging with test content for various item types in computer-administered testlets.

With a move to computer-based testing, many assessment programs have introduced technology-enhanced items. When designing the DLM assessments, the DLM project staff considered the potential trade-offs of these new item types. On one hand, these items offer a means of assessing student knowledge using fewer items, which minimizes the testing burden on a population that has difficulty with long tests. For example, a student’s ability to classify objects could be assessed through a series of multiple-choice items or through one item that involves sorting objects into categories. However, one concern about technology-enhanced item types was that they would be challenging for students with the most significant cognitive disabilities in terms of cognitive demands of the items, lack of familiarity, and the physical access barriers related to students’ fine motor skills.

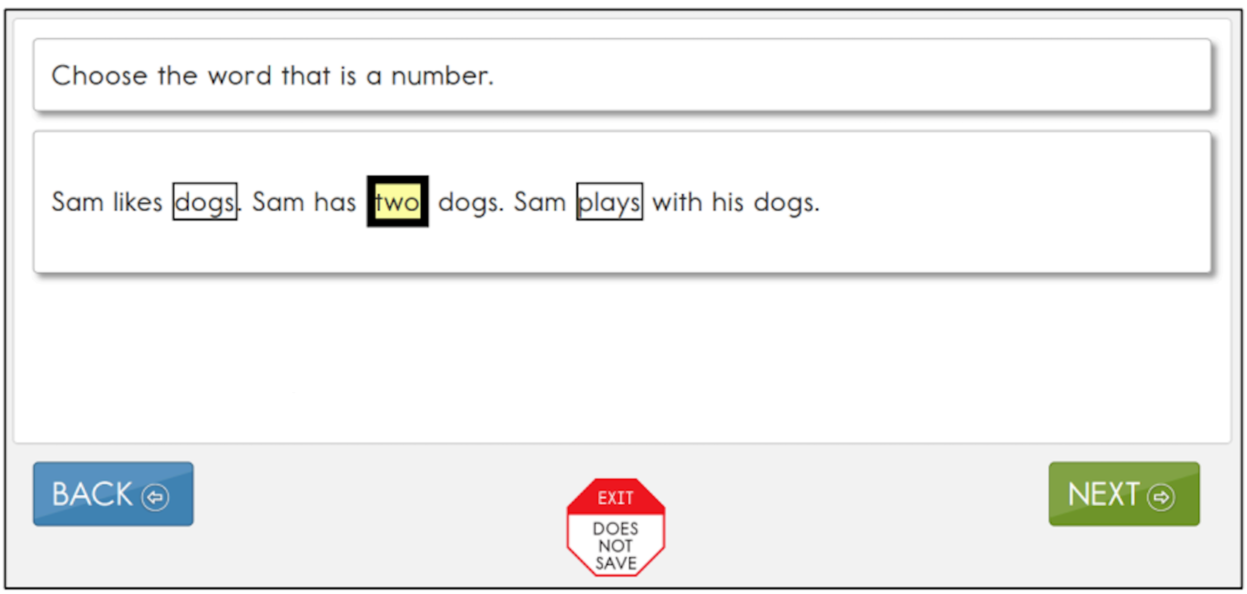
The purpose of this study was to evaluate whether the construct-irrelevant item response demands presented barriers during the response process. Cognitive labs are typically used to elicit statements that allow the observer to know whether the item is tapping the intended cognitive process (Ericsson & Simon, 1993). Due to the challenges in getting students with the most significant cognitive disabilities to verbalize in this manner (Altman et al., 2010), the study included both observational data collection and post-hoc interview questions.

Labs were conducted with 27 students from multiple states in spring 2014 and spring 2015. Eligible students were from tested grades (Grades 3–8 and high school) and had sufficient symbolic communication systems to be able to interact with the content of onscreen items without physical assistance through keyboard and mouse, tablet, or other assistive technology. Inclusion criteria also required that the students have some verbal expressive communication and were able to interact with the testing device without physical assistance.

Labs focused on student interaction with two types of technology-enhanced items, including select-text and multiple-choice multiple-select item types. The first select-text item types were designed specifically for DLM assessments and are delivered through an online testing platform designed for this population. The select-text item type is only used in some ELA assessments. In a select-text item, answer options are marked in a text selection with boxes around words, phrases, or sentences. When a student makes a selection, the word, phrase, or sentence is highlighted in yellow. To clear a selection, the student clicks it again. Multiple-choice multiple-select items were also constructed to access a response process requiring the student to select all of the answer options that matched a category. To avoid relying on items that might be too difficult and therefore inappropriate for use in cognitive labs (Johnstone et al., 2011), the labs used four-item testlets with content that did not rely on prior academic knowledge. Figure 3.20 shows a select-text item that was constructed to minimize the need for prior knowledge.

Figure 3.20

Sample Select-Text Item



Each testlet contained one type of item. For select-text items, the number of objects to sort and the number of categories varied, with more complex versions of the item type appearing later in the testlet. Each student completed two testlets (one per item type) and the order of presentation of testlet assignments was counterbalanced. Eight students completed select-text testlets and 11 students completed multiple-choice multiple-select testlets. The eight students who completed select-text testlets also completed a testlet that used the same content as the select-text items but presented the content in a traditional, single-select multiple-choice format.

For each item type, the examiner looked for evidence of challenge with each step of the item completion process and for evidence indicating whether the student experienced challenges based on the number of objects to be manipulated per item. For all item types, the examiner also looked for evidence of the student's understanding of the task. If the student was not able to complete the task without assistance, the examiner provided additional instructions on how to complete the task.

Students were not asked to talk while they completed the items because of to the potential to increase cognitive load. Instead, they were asked questions at the end of each testlet and after the session. These questions were simpler than those described by Altman et al. [-Altman et al. (2010); e.g., "What makes you believe that answer is the right one?") and only required yes/no responses (e.g., "Did you know what to do?"]. Students were asked the same four questions in the same sequence each time. The yes/no response requirement and identical sequence requirement parallel instructional practice for many students who are eligible for alternate assessments based on alternate achievement standards.

Video recordings of the administrations were reviewed to confirm that the ratings of potential sources of challenge were correctly recorded by the observer. Results reported in this section consist of descriptive statistics for items in the observation protocol and frequency distributions for students' responses to interview questions.

Sources of challenge in responding to multiple-choice multiple-select items were examined by observing student difficulty with the selection of the first object and the subsequent object(s), the concept of needing to make more than one selection, and need of assistance to complete the item. A summary of the sources of challenge in responding to multiple-choice multiple-select items is shown in Table 3.35. On 41% of the items, students had difficulty with the concept of making multiple selections.

Table 3.35

Sources of Challenge in Response to Multiple-Choice Multiple-Select Items

Source of challenge	*n*	%
Difficulty with selection of first object	4	9.0
Difficulty with selection of subsequent objects	6	13.6
Difficulty with multiple-select concept	18	40.9
Needed assistance to complete	9	20.5

Note. *N* = 11 students, 44 items. One testlet was not completed.

The select-text item type required less manipulation of onscreen content and only one selection to respond to the item. Across eight students and 32 items, there were only two items (6.3%) for which the student had difficulty selecting the box and two items (6.3%) for which the student needed assistance to complete the item.

Finally, Table 3.36 summarizes student responses to post-hoc interview questions. For both select-text and multiple-choice multiple-select items, students liked these item types, perceived them as easy, and understood the response process required, with at least 73% of students endorsing each of these responses. Student interview responses were consistent with evaluations of item effectiveness based on sources of challenge noted by the observers.

Table 3.36

Affirmative Student Responses to Post-Hoc Interview Questions

	Multiple select (*N* = 11)		Select text (*N* = 8)	
Question	*n*	%	*n*	%
Did you like it?	9	81.8	8	100.0
Was it easy?	10	90.9	8	100.0
Was it hard?	1	9.0	1	12.5
Did you know what to do?	8	72.7	8	100.0

3.6. Evidence of Item Quality

Testlets are the fundamental unit of the DLM alternate assessments. Each year, testlets are added to and removed from the operational pool to maintain a pool of high-quality testlets. The following sections describe evidence of item quality, including evidence supporting field-test testlets available for administration, a summary of the operational pool, and evidence of differential item functioning (DIF).

3.6.1. Field Testing

Field-test testlets are administered in the DLM assessments to conduct a preliminary evaluation of item quality for EEs assessed at each grade level for ELA and mathematics. In addition to evaluating item quality, field testing is also conducted to deepen operational pools so that multiple testlets are available in the spring window, including making more content available at EEs and linkage levels that educators administer to students the most. By deepening the operational pool, testlets can also be evaluated for retirement in instances in which other testlets perform better. Additionally, assigning field-test testlets at adjacent linkage levels helps support future evaluation of the linkage level ordering (see Chapter 2 of this manual).

There are multiple item quality indicators that are reviewed for items on field-test testlets. Items are expected to be appropriately difficult and to function similarly to items measuring the same EE and linkage level. Items are also expected to be consistent with DLM item-writing guidelines and aligned with the assessed node, and the test development team makes decisions of whether to accept or reject the items on the field-test testlets.

For the spring field-test window, the ELA and mathematics test development teams selected field-test testlets to be assessed for Grades 3–12. In this section, we describe the field tests administered in 2021–2022 and the associated review activities.

3.6.1.1. Description of Field Tests Administered in 2021–2022

The Instructionally Embedded and Year-End assessment models share a common item pool, and testlets field tested during the instructionally embedded assessment window may be eventually promoted to the spring assessment window. Therefore, field testing from both assessment windows is described.

Testlets were made available for field testing based on the availability of field-test content for each EE and linkage level. Because field-test testlets are assigned after completion of the operational assessment and Year-End model students only test in the spring, testing during the instructionally embedded window is optional, so no field tests were administered to students from states adopting the Year-End model during the fall window. However, field tests were completed by students from states participating in the Instructionally Embedded model.

During spring assessment window, field-test testlets were administered after completion of the operational assessment. A field-test testlet was assigned for an EE that was assessed during the operational assessment at a linkage level equal or adjacent to the linkage level of the operational testlet.

Table 3.37 summarizes the number of field-test testlets available during 2021–2022. A total of 477 were available across grades, subjects, and windows.

Table 3.37

2021–2022 Field-Test Testlets, by Subject

	Instructionally embedded assessment window		Spring assessment window	
Grade	English language arts (*n*)	Mathematics (*n*)	English language arts (*n*)	Mathematics (*n*)
3	11	9	18	12
4	9	10	9	15
5	6	7	12	12
6	12	12	19	14
7	9	8	14	13
8	10	8	15	19
9	10	8	19	17
10	10	8	19	17
11	7	5	11	20
12	7	5	11	20

A summary of the demographic breakdown of students completing field-test testlets during 2021–2022 is presented by subject in Table 3.38. Consistent with the DLM population, approximately 67% of students completing field-test testlets were male, approximately 60% were white, and approximately 75% were non-Hispanic. The vast majority of students completing field-test testlets were not English learner eligible or monitored. The students completing field-test testlets were split across the four complexity bands, with most students assigned to Band 1 or Band 2.[19]

[19] See Chapter 4 of this manual for a description of the student complexity bands.

Table 3.38

Demographic Summary of Students Participating in Field Tests

Demographic group	English language arts		Mathematics	
	n	%	*n*	%
Gender				
Male	43,700	67.4	45,829	67.4
Female	21,098	32.5	22,062	32.5
Nonbinary/undesignated	72	0.1	71	0.1
Race				
White	38,170	58.8	39,932	58.8
African American	13,503	20.8	14,161	20.8
Two or More Races	6,976	10.8	7,309	10.8
Asian	3,671	5.7	3,867	5.7
American Indian	2,032	3.1	2,141	3.2
Native Hawaiian or Pacific Islander	371	0.6	395	0.6
Alaska Native	147	0.2	157	0.2
Hispanic ethnicity				
Non-Hispanic	48,855	75.3	51,076	75.2
Hispanic	16,015	24.7	16,886	24.8
English learning (EL) participation				
Not EL eligible or monitored	60,314	93.0	63,242	93.1
EL eligible or monitored	4,556	6.9	4,720	6.9
English language arts complexity band				
Foundational	8,533	13.2	9,492	14.0
Band 1	26,289	40.5	26,799	39.4
Band 2	23,907	36.9	25,046	36.9
Band 3	6,141	9.5	6,625	9.7
Mathematics complexity band				
Foundational	8,992	13.9	9,827	14.5
Band 1	26,247	40.5	26,920	39.6
Band 2	24,611	37.9	25,761	37.9
Band 3	5,020	7.7	5,454	8.0

Note. See Chapter 4 of this manual for a description of student complexity bands.

Participation in field testing was not required, but educators were encouraged to administer all available testlets to their students. Field-test participation rates for ELA and mathematics in the instructionally embedded and spring assessment windows are shown in Table 3.39. Note that because the Instructionally Embedded and Year-End models share an item pool, participation numbers are combined across all states.

In total, 63% of students in ELA and 66% of students in mathematics completed at least one field-test testlet. In the instructionally embedded assessment window, 76% of field-test testlets had a sample size of at least 20 students (i.e., the threshold for item review). In the spring assessment window, 94% of field-test testlets had a sample size of at least 20 students.

Table 3.39

Field-Test Participation, by Subject and Window

	Instructionally embedded assessment window		Spring assessment window		Combined	
Subject	*n*	%	*n*	%	*n*	%
English language arts	4,087	27.2	62,972	62.3	64,870	62.9
Mathematics	3,461	23.4	66,539	66.0	67,962	66.1

3.6.1.2. Field-Test Data Review

Data collected during each field test are compiled, and statistical flags are implemented ahead of test development team review. Flagging criteria serve as a source of evidence for test development teams in evaluating item quality; however, final judgments are content based, taking into account the testlet as a whole, the underlying nodes in the DLM maps that the items were written to assess, and pool depth.

Review of field-test data occurs annually during February and March. This includes data from the immediately preceding instructionally embedded and spring assessment windows. That is, the review in February and March of 2022 includes field-test data collected during the spring 2021 assessment window and the 2021–2022 instructionally embedded assessment window. Data that were collected during the 2022 spring assessment window will be reviewed in February and March of 2023, with results included in the 2022–2023 technical manual update.

Test development teams for each subject make four types of item-level decisions as they review field-test items flagged for either a *p*-value or a standardized difference value beyond the threshold:

1. No changes made to item. Test development team decided item can go forward to operational assessment.
2. Test development team identified concerns that required modifications. Modifications were clearly identifiable and were likely to improve item performance.
3. Test development team identified concerns that required modifications. The content was worth preserving rather than rejecting. Item review may not have clearly pointed to specific edits that were likely to improve the item.
4. Rejected item. Test development team determined the item was not worth revising.

For an item to be accepted as is, the test development team had to determine that the item was consistent with DLM item-writing guidelines and that the item was aligned to the node. An item or testlet was rejected completely if it was inconsistent with DLM item-writing guidelines, if the EE and linkage level were covered by other testlets that had better-performing items, or if there was no clear content-based revision to improve the item. In some instances, a decision to reject an item resulted in the rejection of the testlet, as well.

Common reasons for flagging an item for modification included items that were misaligned to the node, distractors that could be argued as partially correct, or unnecessary complexity in the language of the stem. After reviewing flagged items, the reviewers looked at all items rated as three or four within the testlet to help determine whether to retain or reject the testlet. Here, the test development team could elect to keep the testlet (with or without revision) or reject it. If a revision was needed, it was assumed the testlet needed field testing again. The entire testlet was rejected if the test development team determined the flagged items could not be adequately revised.

3.6.1.3. Results of Item Analysis

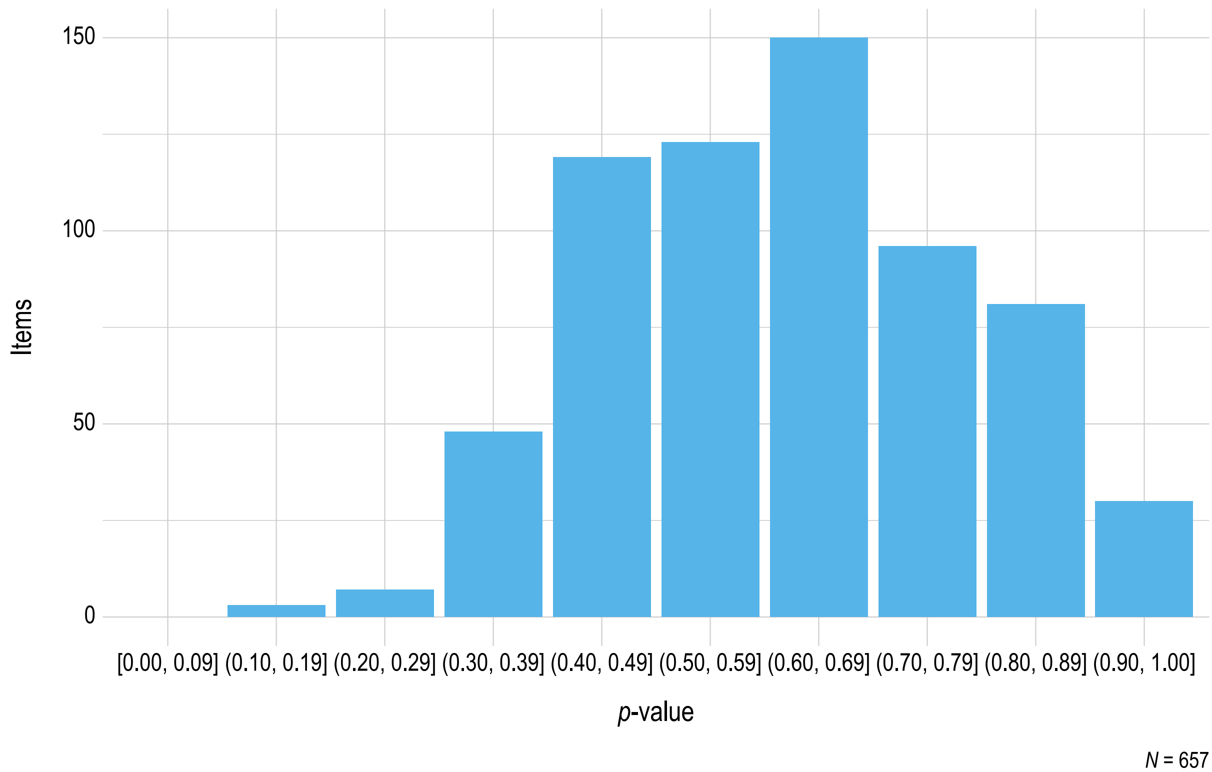
All items are reviewed by test development teams following field testing. Items were specifically flagged if they met either of the following statistical criteria:

- The item was too challenging, as indicated by a *p*-value of less than .35. This value was selected as the threshold for flagging because most DLM assessment items offer three response options, so a value of less than .35 may indicate less than chance selection of the correct response option.
- The item was significantly easier or harder than other items assessing the same EE and linkage level, as indicated by a weighted standardized difference greater than two standard deviations from the mean *p*-value for that EE and linkage level combination.

Figure 3.21 and Figure 3.22 summarize the *p*-values for items that met the minimum sample size threshold of 20. Most items fell above the .35 threshold for flagging. In ELA, 632 items (96%) were above the .35 flagging threshold. In mathematics, 402 items (84%) were above the .35 flagging threshold. Test development teams for each subject reviewed items below the threshold, which was 25 items (4%) for ELA and 74 items (16%) for mathematics.

Figure 3.21

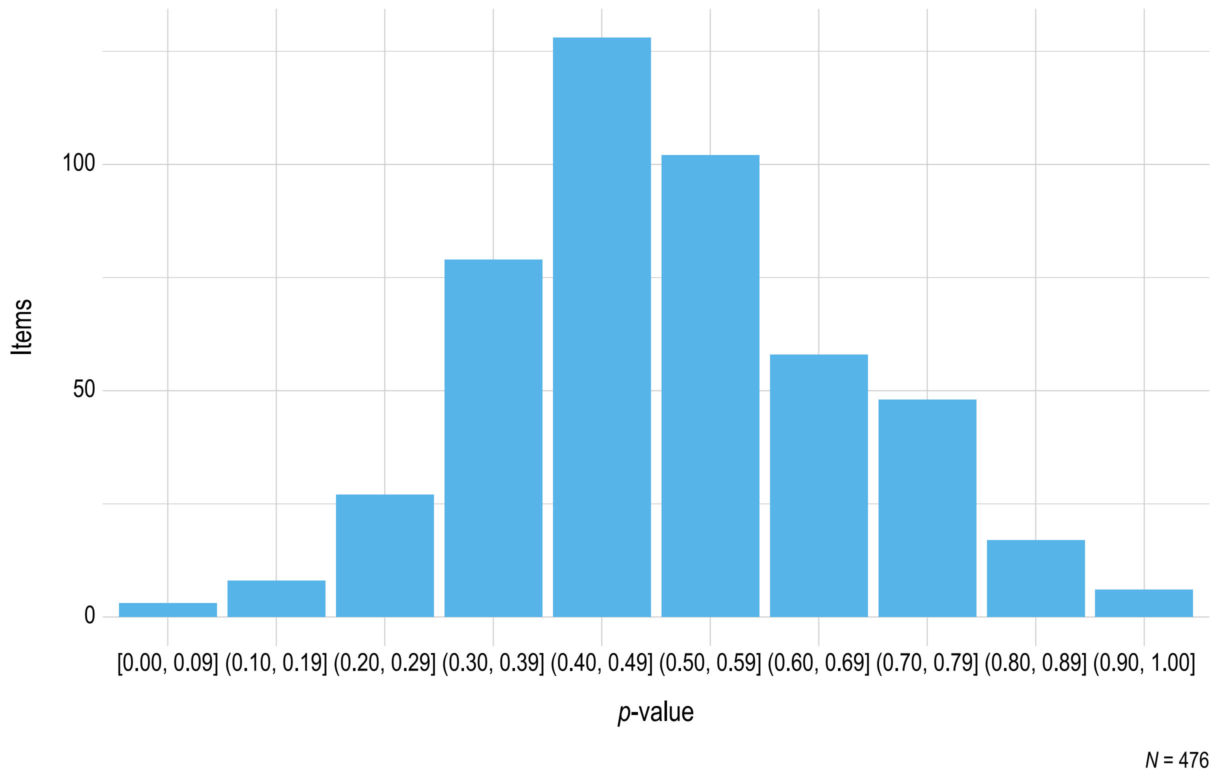
p-*values for English Language Arts Field-Test Items*



Note. Items with a sample size of less than 20 were omitted.

Figure 3.22

p-*values for Mathematics Field-Test Items*

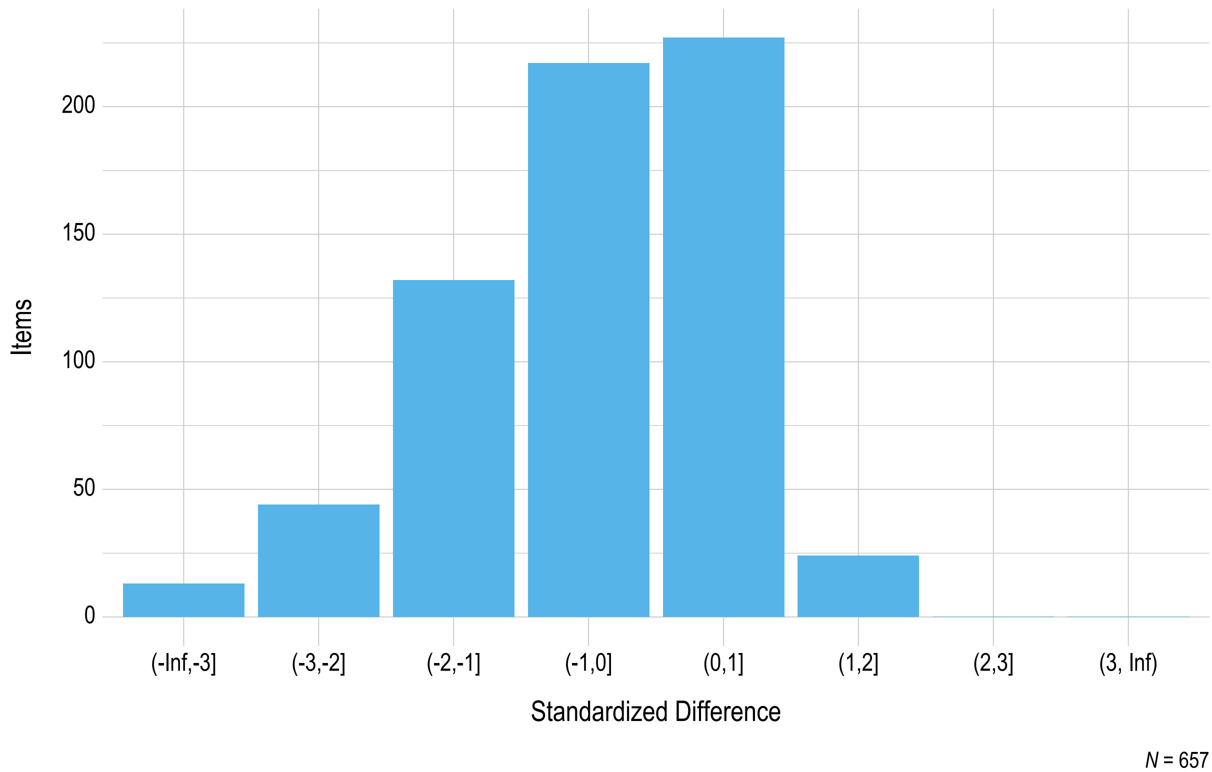


Note. Items with a sample size of less than 20 were omitted.

Figure 3.23 and Figure 3.24 summarize the standardized difference values for items field tested during the instructionally embedded window for ELA and mathematics, respectively. Most items fell within two standard deviations of the mean for the EE and linkage level. Items beyond the threshold were reviewed by test development teams for each subject.

Figure 3.23

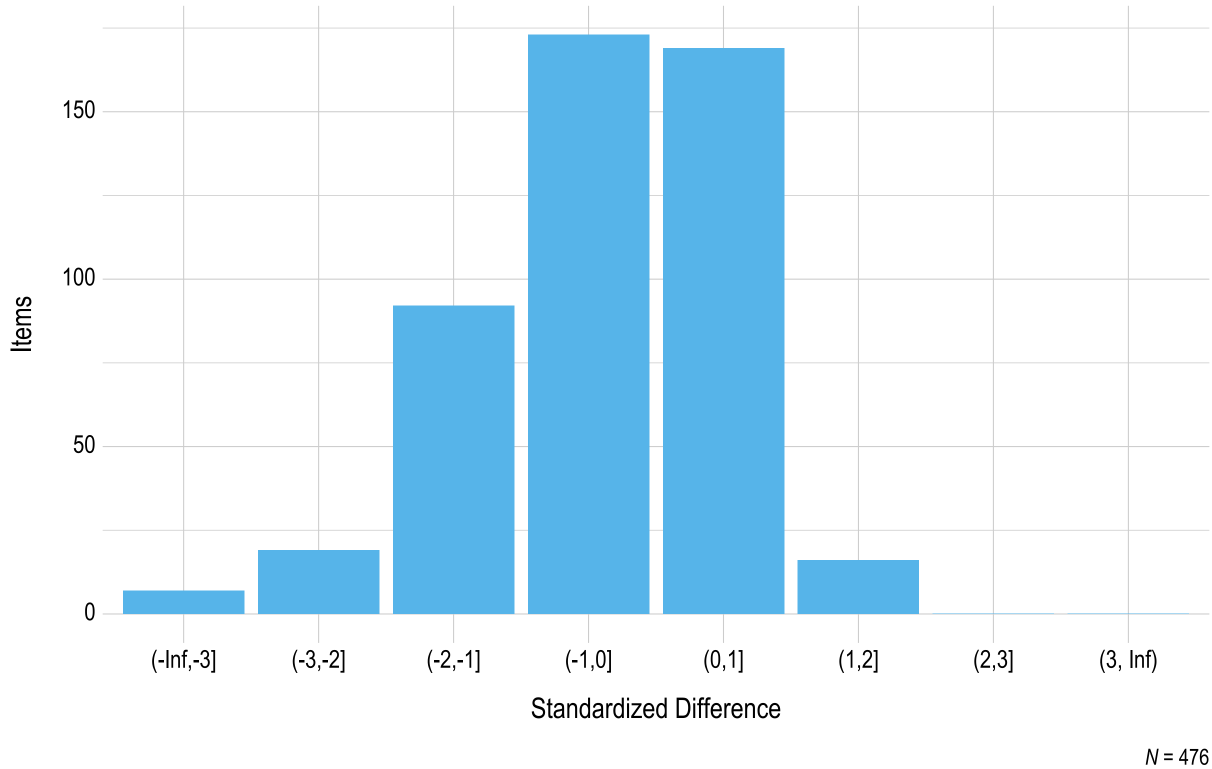
Standardized Difference Z-Scores for English Language Arts Field-Test Items



Note. Items with a sample size of less than 20 were omitted.

Figure 3.24

Standardized Difference Z-Scores for Mathematics Field-Test Items



Note. Items with a sample size of less than 20 were omitted.

A total of 46 ELA testlets (36%) and 44 mathematics testlets (47%) had at least one item flagged due to their *p*-value and/or standardized difference value. Test development teams reviewed all flagged items and their context within the testlet to identify possible reasons for the flag and to determine whether an edit was likely to resolve the issue.

Of the 81 ELA testlets that were not flagged, 15 (19%) were edited and reassigned to the field-test pool for content-based reasons (e.g., changes to item wording), 65 (80%) were promoted to the operational pool, and one (1%) was sent back to the field-test pool with no edits for additional data collection to get estimates of item difficulty that are based on larger samples. Of the 46 ELA testlets that were flagged, 36 (78%) were edited and reassigned to the field-test pool, nine (20%) were sent back to the field-test pool with no edits for additional data collection to get estimates of item difficulty that are based on larger samples, and one (2%) was rejected and retired. Of the 50 mathematics testlets that were not flagged, 10 (20%) were edited and reassigned to the field-test pool for content-based reasons, 36 (72%) were promoted to the operational pool, and four (8%) were rejected and retired. Of the 44 mathematics testlets that were flagged, 22 (50%) were edited and reassigned to the field-test pool, 12 (27%) were promoted to the operational pool to maintain pool depth given content-based testlet retirement, four (9%) were sent back to the field-test pool with no edits for additional data collection to get estimates of item difficulty that are based on larger samples, and six (14%) were rejected and retired.

3.6.2. Operational Assessment Items for 2021–2022

The DLM assessments include a total of 901 operational testlets, with 515 operational ELA testlets and 386 operational mathematics testlets. Because the operational pool needs to be deepened, particularly for content at the EEs and linkage levels that are administered to students the most, updates are made to the operational pool each year. The primary updates are promoting testlets to the operational pool and removing testlets from the operational pool.

Testlets are promoted to the operational pool via field testing, with students who completed the operational assessment in the spring. Field-test testlets are eligible for review after they have been completed by at least 20 students. As mentioned in the field testing section above (section 3.6.1), there are multiple item quality indicators that are considered when deciding whether to promote an item to the operational pool. Statistically, items are expected to be appropriately difficult and to function similarly to items measuring the same EE and linkage level. To review these statistical item quality indicators, the difficulty and internal consistency of items on field-test testlets are evaluated. Items are also expected to be consistent with DLM item-writing guidelines and aligned with the assessed node. To review these content-based item quality indicators, the quality of the eligible items on the field-test testlets is evaluated, and the test development team makes decisions of whether to accept or reject the items on the field-test testlets. For a full description of field testing, see above in section 3.6.1.

Testlets are removed from the operational pool via retirement based on item quality standards. There are several processes that can lead an item or testlet to be prioritized for retirement. Items are evaluated for evidence of model fit, and the results of these evaluations may be used to prioritize items and testlets for retirement. Items are also evaluated for evidence of DIF, and these results may be used to prioritize items and testlets for retirement. This process is described in section 3.6.3. Finally, the test development team periodically reviews the content pool and prioritizes testlets for retirement. These reviews refresh the operational pool by removing older content when newer content is available.

For 2021–2022, 113 testlets were promoted to the operational pool from field testing in 2020–2021, including 65 ELA testlets and 48 mathematics testlets.

Testlets were made available for operational testing in 2021–2022 based on the 2020–2021 operational pool and the promotion of testlets field-tested during 2020–2021 to the operational pool following their review. Table 3.40 summarizes the total number of operational testlets for 2021–2022. In total, there were 901 operational testlets available. This total included 336 EE/linkage level combinations (192 ELA, 144 mathematics) for which both a general version and a version for students who are blind or visually impaired or read braille were available.

Table 3.40

2021–2022 Operational Testlets, by Subject (N = 901)

Grade	English language arts (*n*)	Mathematics (*n*)
3	66	42
4	72	47
5	68	49
6	65	45
7	65	39
8	58	44
9–10	59	120
11–12	62	*

* In mathematics, high school is banded in grades 9–11.

3.6.2.1. Educator Perception of Assessment Content

Each year, test administrators are asked two questions about their perceptions of the assessment content;[20] Table 3.41 describes their responses in 2021–2022. Questions pertained to whether the DLM assessments measured important academic skills and reflected high expectations for their students.

Test administrators generally responded that content reflected high expectations for their students (85% agreed or strongly agreed) and measured important academic skills (77% agreed or strongly agreed). While the majority of test administrators agreed with these statements, 15%–23% disagreed. DLM assessments represent a departure from the breadth of academic skills assessed by many states' previous alternate assessments. Given the short history of general curriculum access for this population and the tendency to prioritize the instruction of functional academic skills (Karvonen et al., 2011), test administrators' responses may reflect awareness that DLM assessments contain challenging content. However, test administrators were divided on its importance in the educational programs of students with the most significant cognitive disabilities. Feedback from focus groups with educators focusing on score reports reflected similar variability in educator perceptions of assessment content (Clark et al., 2018, 2022).

Table 3.41

Educator Perceptions of Assessment Content

	Strongly disagree		Disagree		Agree		Strongly agree	
Statement	*n*	%	*n*	%	*n*	%	*n*	%
Content measured important academic skills and knowledge for this student.	4,297	8.7	7,236	14.7	28,687	58.3	9,016	18.3
Content reflected high expectations for this student.	2,285	4.7	4,901	10.0	28,806	58.9	12,914	26.4

[20] Participation in the test administrator survey is described in Chapter 4 of this manual.

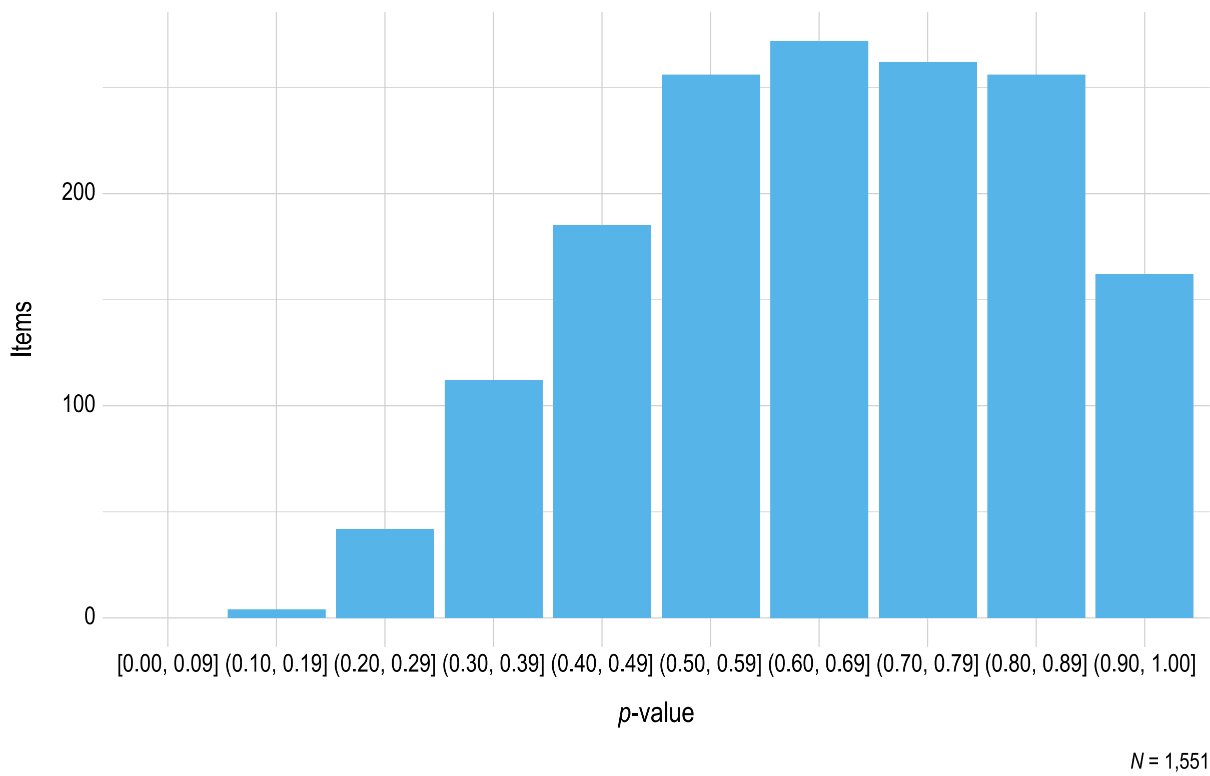
3.6.2.2. Psychometric Properties of Operational Assessment Items for 2021–2022

The proportion correct (*p*-value) was calculated for all operational items to summarize information about item difficulty.

Figure 3.25 and Figure 3.26 include the *p*-values for each operational item for ELA and mathematics, respectively. To prevent items with small sample sizes from potentially skewing the results, the sample size cutoff for inclusion in the *p*-value plots was 20. In total, 38 items (1% of all items) were excluded due to small sample size, where 18 of the items were ELA items (1% of all ELA items) and 20 of the items were mathematics items (1% of all mathematics items). In general, ELA items were easier than mathematics items, as evidenced by the presence of more items in the higher bin (*p*-value) ranges.

Figure 3.25

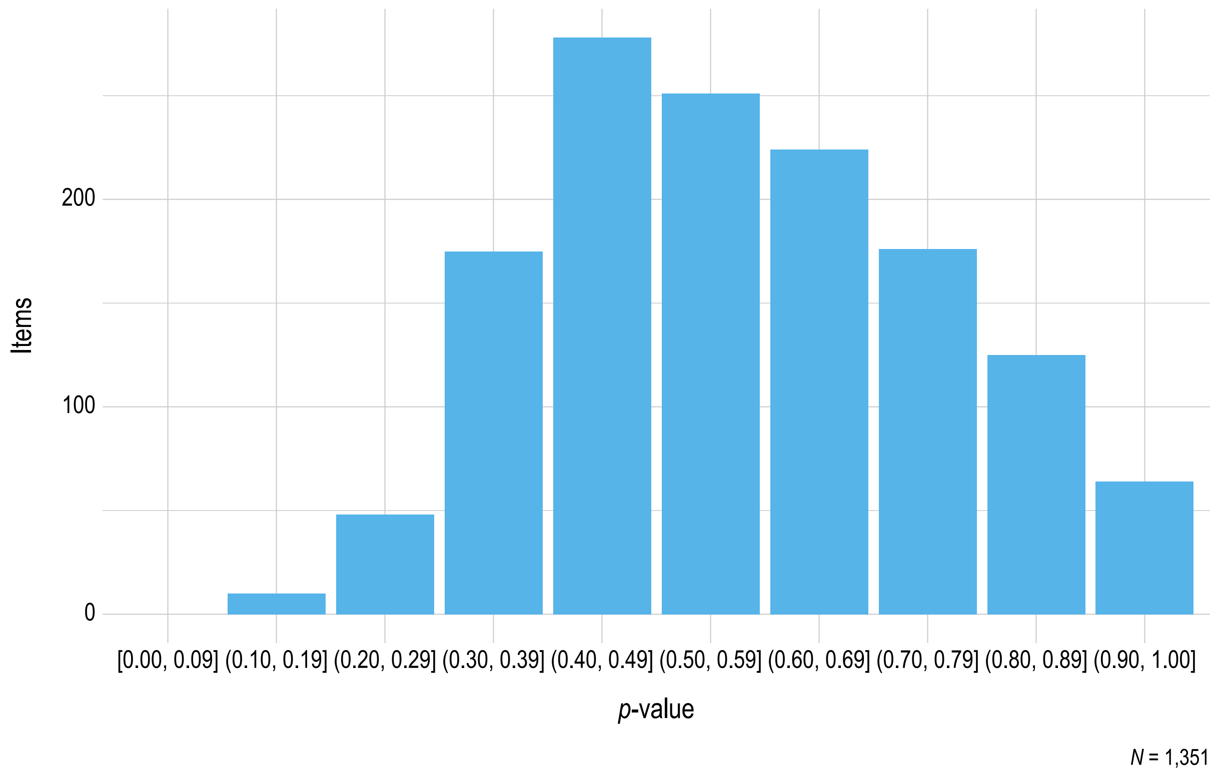
p-*values for English Language Arts 2021–2022 Operational Items*



Note. Items with a sample size of less than 20 were omitted.

Figure 3.26

p-*values for Mathematics 2021–2022 Operational Items*



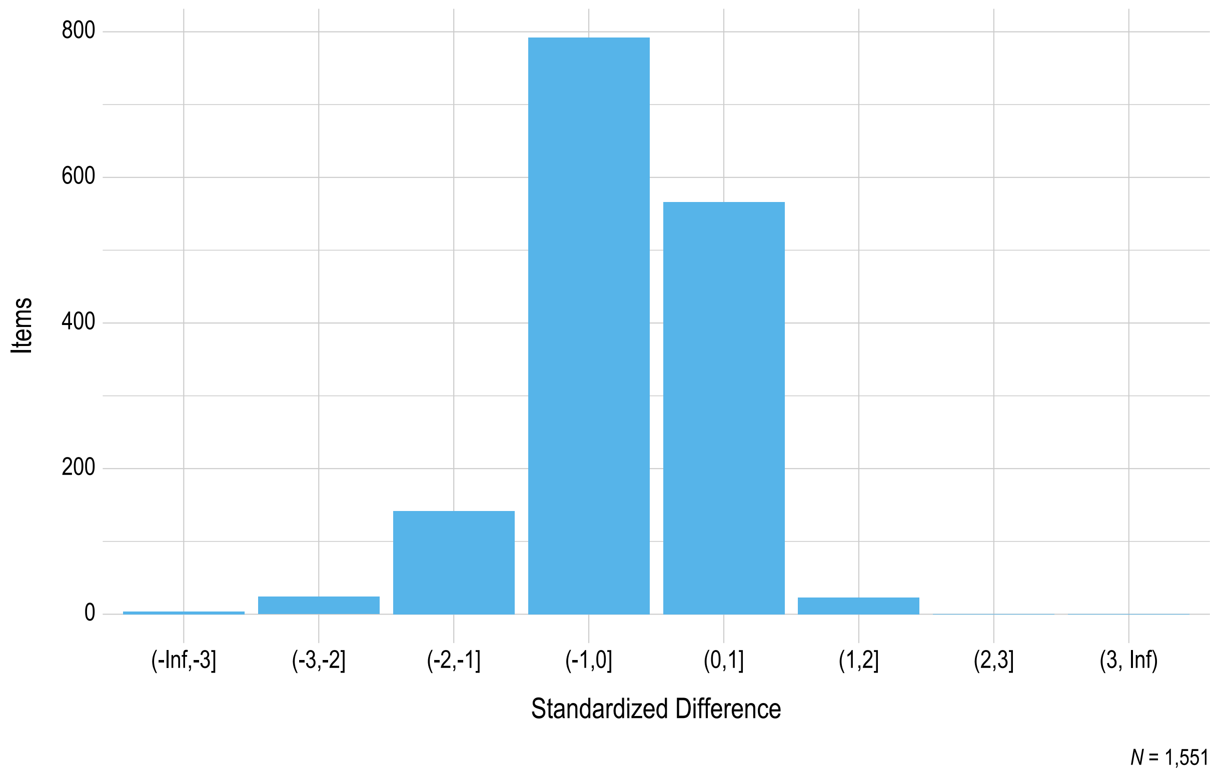
Note. Items with a sample size of less than 20 were omitted.

Items in the DLM assessments are fungible (i.e., interchangeable) within each EE and linkage level, meaning that the items are expected to function identically to the other items measuring the same EE and linkage level. To evaluate the fungibility assumption, standardized difference values were also calculated for all operational items, with a student sample size of at least 20 required to compare the *p*-value for the item to all other items measuring the same EE and linkage level. If an item is fungible with the other items measuring the same EE and linkage level, the item is expected to have a nonsignificant standardized difference value. The standardized difference values provide one source of evidence of internal consistency.

Figure 3.27 and Figure 3.28 summarize the standardized difference values for operational items for ELA and mathematics, respectively. Of all items measuring the EE and linkage level, 98% of ELA items and 99% of mathematics items fell within two standard deviations of the mean. As additional data are collected and decisions are made regarding item pool replenishment, test development teams will consider item standardized difference values, along with item misfit analyses, when determining which items and testlets are recommended for retirement.

Figure 3.27

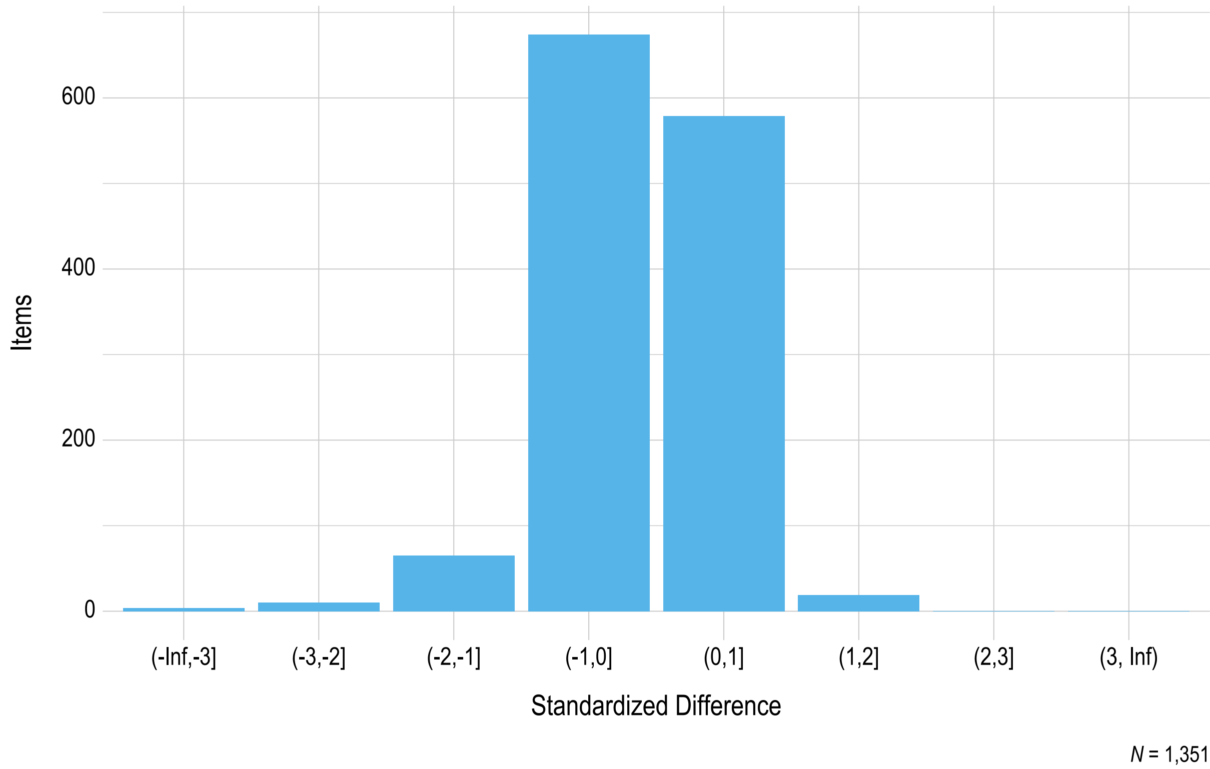
Standardized Difference Z-Scores for English Language Arts 2021–2022 Operational Items



Note. Items with a sample size of less than 20 were omitted.

Figure 3.28

Standardized Difference Z-Scores for Mathematics 2021–2022 Operational Items

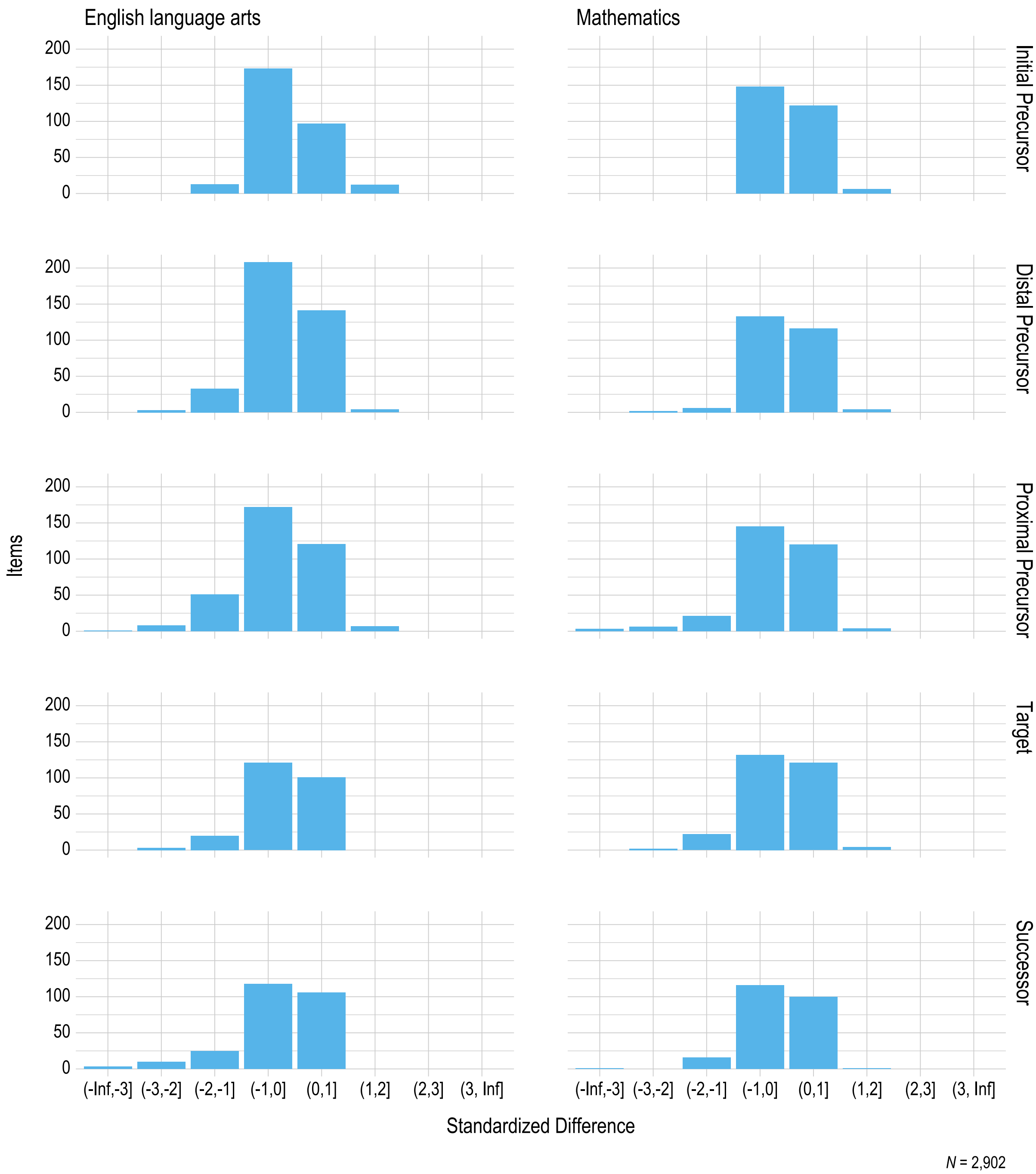


Note. Items with a sample size of less than 20 were omitted.

Figure 3.29 summarizes the standardized difference values for operational items by linkage level. Most items fell within two standard deviations of the mean of all items measuring the respective EE and linkage level, and the distributions are consistent across linkage levels.

Figure 3.29

Standardized Difference Z-Scores for 2021–2022 Operational Items by Linkage Level



Note. Items with a sample size of less than 20 were omitted.

3.6.3. Evaluation of Item-Level Bias

Analyses comparing how items function across subgroups of students indicate one source of evidence for item quality. Given the heterogeneous nature of the student population, statistical analyses can examine whether particular items function differently for specific subgroups (e.g., male versus female). Each year, DLM assessment items are reviewed for evidence of DIF for gender and ethnicity subgroups. The following sections provide a summary of findings from the evaluation of item-level bias.

DIF addresses the challenges created when some test items are more difficult for some groups of examinees despite these examinees having knowledge and understanding of the assessed concepts (Camilli & Shepard, 1994). DIF analyses can uncover internal inconsistency if particular items are functioning differently in a systematic way for identifiable subgroups of students (AERA et al., 2014). While identification of DIF does not always indicate a weakness in the test item, it can point to construct-irrelevant variance, posing considerations for validity and fairness.

3.6.3.1. Method

DIF analyses examined race in addition to gender. Analyses included data from 2015–2016 through 2020–2021[21] to flag items for evidence of DIF. Items were selected for inclusion in the DIF analyses based on minimum sample-size requirements for the two gender subgroups (male and female) and for race subgroups: white, African American, Asian, American Indian, Native Hawaiian or Pacific Islander, Alaska Native, and multiple races.

The DLM student population is unbalanced in both gender and race. The number of female students responding to items is smaller than the number of male students by a ratio of approximately 1:2. Similarly, the number of nonwhite students responding to items is smaller than the number of white students by a ratio of approximately 1:2. Therefore, on advice from the DLM Technical Advisory Committee, the threshold for item inclusion requires that the focal group must have at least 100 students responding to the item. The threshold of 100 was selected to balance the need for a sufficient sample size in the focal group with the relatively low number of students responding to many DLM items. Writing items were excluded from the DIF analyses described here because they include nonindependent response options.

Additional criteria were included to prevent estimation errors. Items with an overall proportion correct (*p*-value) greater than .95 or less than .05 were removed from the analyses. Items for which the *p*-value for one gender or racial group was greater than .97 or less than .03 were also removed from the analyses.

For each item, logistic regression was used to predict the probability of a correct response, given group membership and performance in the subject. Specifically, the logistic regression equation for each item included a matching variable comprised of the student's total linkage levels mastered in the subject of the item and a group membership variable, with the reference group (i.e., males for gender, white for race) coded as 1 and the focal group (i.e., females for gender; African American, Asian, American Indian, Native Hawaiian or Pacific Islander, Alaska Native, or two or more races for race) coded as 0. An interaction term was included to evaluate whether nonuniform DIF was present for each item (Swaminathan & Rogers, 1990); the presence of nonuniform DIF indicates that the item functions differently because of the interaction between total linkage levels mastered and the student's group (i.e., gender or racial group).

[21] DIF analyses are conducted on the sample of data used to update the model calibration, which uses data through the previous operational assessment. See Chapter 5 of this manual for more information.

When nonuniform DIF is present, the group with the highest probability of a correct response to the item differs along the range of total linkage levels mastered; thus, one group is favored at the low end of the spectrum and the other group is favored at the high end.

Three logistic regression models were fitted for each item:

$$\text{M}_0\text{: logit}(\pi_i) = \beta_0 + \beta_1\text{X} \tag{3.1}$$

$$\text{M}_1\text{: logit}(\pi_i) = \beta_0 + \beta_1\text{X} + \beta_2\text{G} \tag{3.2}$$

$$\text{M}_2\text{: logit}(\pi_i) = \beta_0 + \beta_1\text{X} + \beta_2\text{G} + \beta_3\text{XG} \tag{3.3}$$

where π_i is the probability of a correct response to item *i*, X is the matching criterion, G is a dummy coded grouping variable (0 = reference group, 1 = focal group), β_0 is the intercept, β_1 is the slope, β_2 is the group-specific parameter, and β_3 is the interaction term.

Because of the number of items evaluated for DIF, Type I error rates were susceptible to inflation. The incorporation of an effect-size measure can be used to distinguish practical significance from statistical significance by providing a metric of the magnitude of the effect of adding group and interaction terms to the regression model.

For each item, the change in the Nagelkerke pseudo R^2 measure of effect size was captured, from M_0 to M_1 or M_2, to account for the effect of the addition of the group and interaction terms to the equation. All effect-size values were reported using both the Zumbo and Thomas (1997) and Jodoin and Gierl (2001) indices for reflecting a negligible, moderate, or large effect. The Zumbo and Thomas thresholds for classifying DIF effect size are based on Cohen's (1992) guidelines for identifying a small, medium, or large effect. The thresholds for each level are .13 and .26; values less than .13 have a negligible effect, values between .13 and .26 have a moderate effect, and values of .26 or greater have a large effect. The Jodoin and Gierl thresholds are more stringent, with lower threshold values of .035 and .07 to distinguish between negligible, moderate, and large effects.

3.6.3.2. Results

Using the above criteria for inclusion, 2,385 (83%) items were selected for gender, and 1,950 (67%) items were selected for at least one racial group comparison. The number of items evaluated by grade and subject for gender ranged from 30 in grades 9–10 ELA to 233 in grades 9–10 mathematics. The number of items evaluated by grade and subject for race ranged from nine in grades 9–10 ELA to 149 in grade 4 ELA. Because students taking DLM assessments represent seven possible racial groups,[22] there are up to six comparisons that can be made for each item, with the white group as the reference group and each of the other six groups (i.e., African American, Asian, American Indian, Native Hawaiian or Pacific Islander, Alaska Native, two or more races) as the focal group. Across all items, this results in 40,446 possible comparisons. Using the inclusion criteria specified above, 7,686 (19%) item and focal group comparisons were selected for analysis. Overall, 508 items were evaluated for one racial focal group, 1,006 items were evaluated for two racial focal groups, 421 items were evaluated for three racial focal groups, and 15 items were evaluated for four racial focal groups. One racial focal group and the white reference group were

[22] See Chapter 7 of this manual for a summary of participation by race and other demographic variables.

used in each comparison. Table 3.42 shows the number of items that were evaluated for each racial focal group. Across all gender and race comparisons, sample sizes for each comparison ranged from 271 to 6,720 for gender and from 412 to 5,708 for race.

Table 3.42

Number of Items Evaluated for Each Race

Focal group	Items (*n*)
African American	1,949
American Indian	19
Asian	433
Two or more races	1,442

Of the 504 items (17% of the operational item pool) that were not included in the DIF analysis for gender, 379 (75%) had a focal group sample size of less than 100, 70 (14%) had an item *p*-value greater than .95, and 55 (11%) had a subgroup *p*-value greater than .97. A total of 939 items were not included in the DIF analysis for race for any of the subgroups. Of the 13,491 item and focal group comparisons that were not included in the DIF analysis for race, 13,154 (98%) had a focal group sample size of less than 100, 105 (1%) had an item *p*-value greater than .95, and 232 (2%) had a subgroup *p*-value greater than .97. Table 3.43 and Table 3.44 show the number and percentage of items that did not meet each inclusion criteria for gender and race, respectively, by subject and the linkage level the items assess. The majority of nonincluded comparisons come from ELA for both gender (*n* = 349; 69%) and race (*n* = 7,274; 54%).

Table 3.43

Comparisons Not Included in Differential Item Functioning Analysis for Gender, by Subject and Linkage Level

	Sample size		Item proportion correct		Subgroup proportion correct	
Subject and Linkage Level	*n*	%	*n*	%	*n*	%
English language arts						
Initial Precursor	62	23.4	0	0.0	0	0.0
Distal Precursor	97	36.6	0	0.0	1	2.7
Proximal Precursor	64	24.2	0	0.0	7	18.9
Target	6	2.3	10	21.3	13	35.1
Successor	36	13.6	37	78.7	16	43.2
Mathematics						
Initial Precursor	5	4.4	0	0.0	0	0.0
Distal Precursor	5	4.4	0	0.0	2	11.1
Proximal Precursor	17	14.9	12	52.2	3	16.7
Target	19	16.7	7	30.4	7	38.9
Successor	68	59.6	4	17.4	6	33.3

Table 3.44

Comparisons Not Included in Differential Item Functioning Analysis for Race, by Subject and Linkage Level

	Sample size		Item proportion correct		Subgroup proportion correct	
Subject and Linkage Level	*n*	%	*n*	%	*n*	%
English language arts						
Initial Precursor	1,153	16.4	0	0.0	0	0.0
Distal Precursor	1,813	25.7	0	0.0	15	9.7
Proximal Precursor	1,722	24.4	1	1.4	38	24.5
Target	1,142	16.2	17	23.6	43	27.7
Successor	1,217	17.3	54	75.0	59	38.1
Mathematics						
Initial Precursor	1,014	16.6	0	0.0	2	2.6
Distal Precursor	1,089	17.8	0	0.0	5	6.5
Proximal Precursor	1,322	21.6	16	48.5	14	18.2
Target	1,384	22.7	10	30.3	35	45.5
Successor	1,298	21.3	7	21.2	21	27.3

3.6.3.2.1. Uniform Differential Item Functioning Model

A total of 250 items for gender were flagged for evidence of uniform DIF when comparing M_0 to M_1. Additionally, 364 item and focal group combinations across 335 items for race were flagged for evidence of uniform DIF. Table 3.45 and Table 3.46 summarize the total number of combinations flagged for evidence of uniform DIF by subject and grade for gender and race, respectively. The percentage of combinations flagged for uniform DIF ranged from 2% to 14% for gender and from 3% to 20% for race.

Table 3.45

Combinations Flagged for Evidence of Uniform Differential Item Functioning for Gender

Grade	Items flagged (*n*)	Total items (*N*)	Items flagged (%)	Items with moderate or large effect size (*n*)
English language arts				
3	8	148	5.4	0
4	17	162	10.5	0
5	17	154	11.0	0
6	12	141	8.5	0
7	13	148	8.8	0
8	19	135	14.1	0
9	2	22	9.1	0
10	1	8	12.5	0
11	21	148	14.2	0
9–10	12	111	10.8	0
Mathematics				
3	19	136	14.0	0
4	22	153	14.4	0
5	15	158	9.5	0
6	12	148	8.1	0
7	16	126	12.7	0
8	17	147	11.6	0
9	2	121	1.7	0
10	12	112	10.7	0
11	13	107	12.1	0

Table 3.46

Combinations Flagged for Evidence of Uniform Differential Item Functioning for Race

Grade	Items flagged (*n*)	Total items (*N*)	Items flagged (%)	Items with moderate or large effect size (*n*)
English language arts				
3	28	279	10.0	0
4	29	291	10.0	0
5	25	253	9.9	0
6	26	238	10.9	0
7	18	253	7.1	0
8	29	250	11.6	0
9	1	5	20.0	0
11	16	189	8.5	0
9–10	13	118	11.0	0
Mathematics				
3	34	275	12.4	0
4	27	306	8.8	0
5	22	269	8.2	0
6	21	259	8.1	0
7	21	217	9.7	0
8	23	279	8.2	0
9	6	105	5.7	0
10	2	65	3.1	0
11	23	186	12.4	0

For gender, using the Zumbo and Thomas (1997) effect-size classification criteria, all combinations were found to have a negligible effect-size change after the gender term was added to the regression equation. When using the Jodoin and Gierl (2001) effect-size classification criteria, all combinations were found to have a negligible effect-size change after the gender term was added to the regression equation.

The results of the DIF analyses for race were similar to those for gender. When using the Zumbo and Thomas (1997) effect-size classification criteria, all combinations were found to have a negligible effect-size change after the race term was added to the regression equation. Similarly, when using the Jodoin and Gierl (2001) effect-size classification criteria, all combinations were found to have a negligible effect-size change after the race term was added to the regression equation.

3.6.3.2.2. Combined Model

A total of 305 items were flagged for evidence of DIF when both the gender and interaction terms were included in the regression equation, as shown in Equation 3.3. Additionally, 424 item and focal group combinations across 381 items were flagged for evidence of DIF when both the race and interaction terms were included in the regression equation. Table 3.47 and Table 3.48 summarize the number of combinations flagged by subject and grade. The percentage of combinations flagged ranged from 6% to 18% for gender and from 5% to 20% for race.

Table 3.47

Items Flagged for Evidence of Differential Item Functioning for the Combined Model for Gender

Grade	Items flagged (*n*)	Total items (*N*)	Items flagged (%)	Items with moderate or large effect size (*n*)
English language arts				
3	9	148	6.1	0
4	21	162	13.0	0
5	16	154	10.4	0
6	17	141	12.1	0
7	19	148	12.8	0
8	14	135	10.4	0
9	3	22	13.6	0
10	1	8	12.5	0
11	19	148	12.8	0
9–10	15	111	13.5	0
Mathematics				
3	22	136	16.2	0
4	24	153	15.7	0
5	21	158	13.3	0
6	26	148	17.6	0
7	18	126	14.3	0
8	24	147	16.3	0
9	10	121	8.3	0
10	14	112	12.5	0
11	12	107	11.2	0

Table 3.48

Items Flagged for Evidence of Differential Item Functioning for the Combined Model for Race

Grade	Items flagged (*n*)	Total items (*N*)	Items flagged (%)	Items with moderate or large effect size (*n*)
English language arts				
3	19	279	6.8	0
4	30	291	10.3	0
5	29	253	11.5	0
6	28	238	11.8	0
7	29	253	11.5	0
8	27	250	10.8	0
9	1	5	20.0	0
11	26	189	13.8	0
9–10	15	118	12.7	0
Mathematics				
3	29	275	10.5	0
4	46	306	15.0	0
5	25	269	9.3	0
6	18	259	6.9	0
7	42	217	19.4	0
8	24	279	8.6	0
9	9	105	8.6	0
10	3	65	4.6	0
11	24	186	12.9	0

Using the Zumbo and Thomas (1997) effect-size classification criteria, all combinations were found to have a negligible effect-size change after the gender and interaction terms were added to the regression equation. When using the Jodoin and Gierl (2001) effect-size classification criteria, all combinations were found to have a negligible effect-size change after the gender and interaction terms were added to the regression equation.

The results of the DIF analyses for race were similar to those for gender. When using the Zumbo and Thomas (1997) effect-size classification criteria, all combinations were found to have a negligible effect-size change after the race and interaction terms were added to the regression equation. Similarly, when using the Jodoin and Gierl (2001) effect-size classification criteria, all combinations were found to have a negligible effect-size change after the race and interaction terms were added to the regression equation.

3.7. Conclusion

Content in the DLM assessments undergoes multiple rounds of internal and external review before it is promoted into the operational pool. Item writers are trained and given resource materials prior to developing items and testlets. The created content is first reviewed internally by the test development team, the editing team, and content and accessibility panelists. The created content is then reviewed externally by content, accessibility, and bias and sensitivity panelists, and the test development team uses the panelist feedback to revise the items as necessary. After these internal and external reviews are complete, the content is field tested and the results of the field test are reviewed by the test development and psychometric teams. Testlets and items that do not require revision can be promoted to the operational pool.

During the 2021–2022 academic year, the test development teams conducted virtual events for both item writing and external review. Overall, 427 testlets were written for ELA and mathematics. Following external review, the test development team retained 100% and 47% of ELA and mathematics testlets, respectively. Of the content already in the operational pool, most items had *p*-values within two standard deviations of the mean for the EE and linkage level. Field testing in 2021–2022 focused on collecting data to refresh the operational pool of testlets.

4. Assessment Delivery

Chapter 4 presents the processes and procedures used to deliver the Dynamic Learning Maps® (DLM®) Alternate Assessment System in 2021–2022. As described in earlier chapters, the DLM System uses adaptive computer-delivered alternate assessments that provide the opportunity for students with the most significant cognitive disabilities to show what they know and can do in English language arts (ELA) and mathematics. DLM assessments are administered in small groups of items called testlets. The DLM assessment system incorporates accessibility by design and is guided by the core beliefs that all students should have access to challenging, grade-level content and that educators adhere to the highest levels of integrity in providing instruction and administering assessments based on this challenging content.

This chapter begins with an overview of the general features of assessment administration, including the Kite® Suite used to assign and deliver assessments, testlet formats (computer-delivered and educator-administered), and accessibility features. Next, we describe the key features of the Year-End assessment model. We explain how a student's First Contact survey is used to assign the first testlet in each subject and the adaptive routing algorithm that is used to assign subsequent testlets. We also describe administration resources and materials available to test administrators and district users, followed by test administrator responsibilities and procedures and test security. We then provide evidence from the DLM System, including administration time, device usage, linkage level selection, evaluation of blueprint coverage, and accessibility support selections. We also present evidence from assessment administration monitoring, including test administration observations, formative monitoring, and data forensics reports. Finally, we present evidence from test administrators, including user experience with the DLM System, students' opportunity to learn, ratings of items on the First Contact survey, and educator cognitive labs.

4.1. Overview of General Administration Features

Based on students' support needs, DLM assessments are designed to be administered in a one-on-one, student/test administrator format. Most test administrators are the special education educators of the students, as they are best equipped to provide the most conducive conditions to elicit valid and reliable results. Assessment administration processes and procedures also reflect the priorities of fairness and validity through a broad array of accessibility tools and features that are designed to provide access to assessment content and materials as well as limit construct-irrelevant variance.

This section describes the key, overarching features of DLM assessment administration, including the online testing platform, the Kite Suite, the two assessment delivery modes, and accessibility features.

4.1.1. The Kite Suite

The DLM alternate assessments are managed and delivered using the Kite Suite, which was designed and developed to meet the needs of the next generation of large-scale assessments for students with significant cognitive disabilities. Educators and students use the following applications: Kite Educator Portal and Kite Student Portal. The Kite Suite was developed with IMS Global Question and Test Interoperability item structures and Accessible Portable Item Protocol tagging on assessment content to support students' Personal Needs and Preferences (PNP) Profiles (see the Accessibility section below) and World Wide Web Consortium Web Content Accessibility Guidelines. Kite Student Portal and supported browsers for Kite Educator Portal are published on the DLM website and in the Technology Specifications

Manual (DLM Consortium, 2022b) linked on each state's DLM webpage.

4.1.1.1. Kite Educator Portal

Kite Educator Portal is the administrative application where district staff and educators manage student data, assign optional instructionally embedded assessments, access resources needed for each assigned testlet, and retrieve reports.

- Assessment administrators, who are usually educators, use Kite Educator Portal to manage all student data. They are responsible for checking class rosters of the students who are assigned to take DLM testlets and for completing the PNP and First Contact surveys for each student (see the respective Accessibility and Linkage Level sections below for more information on the PNP and First Contact surveys, respectively).
- Essential Elements (EEs) are administered in a pre-determined, fixed sequence. The linkage level for the first testlet is determined by responses to the First Contact survey, and subsequent testlets are determined by an adaptive routing algorithm. After the EE and linkage level are assigned, the test administrator retrieves information to support instruction on the associated nodes. See section 4.2 on key administration features of the Year-End model for more information on testlet assignment.
- After each testlet is assigned to a student, the system delivers a Testlet Information Page (TIP) through Kite Educator Portal. The TIP, which is unique to the assigned testlet, is a PDF that contains any instructions necessary to prepare for testlet administration. See section 4.3.1.2.1 of this chapter for more information.
- During optional instructionally embedded assessments, the Instruction and Assessment Planner displays information about student mastery for assessed EEs and linkage levels. Educators can also download or print reports on demand, including the student's history of instructional plans created in the Instruction and Assessment Planner as well as a report that shows the EEs and linkage levels for which the student has completed a testlet or a testlet assignment is pending.

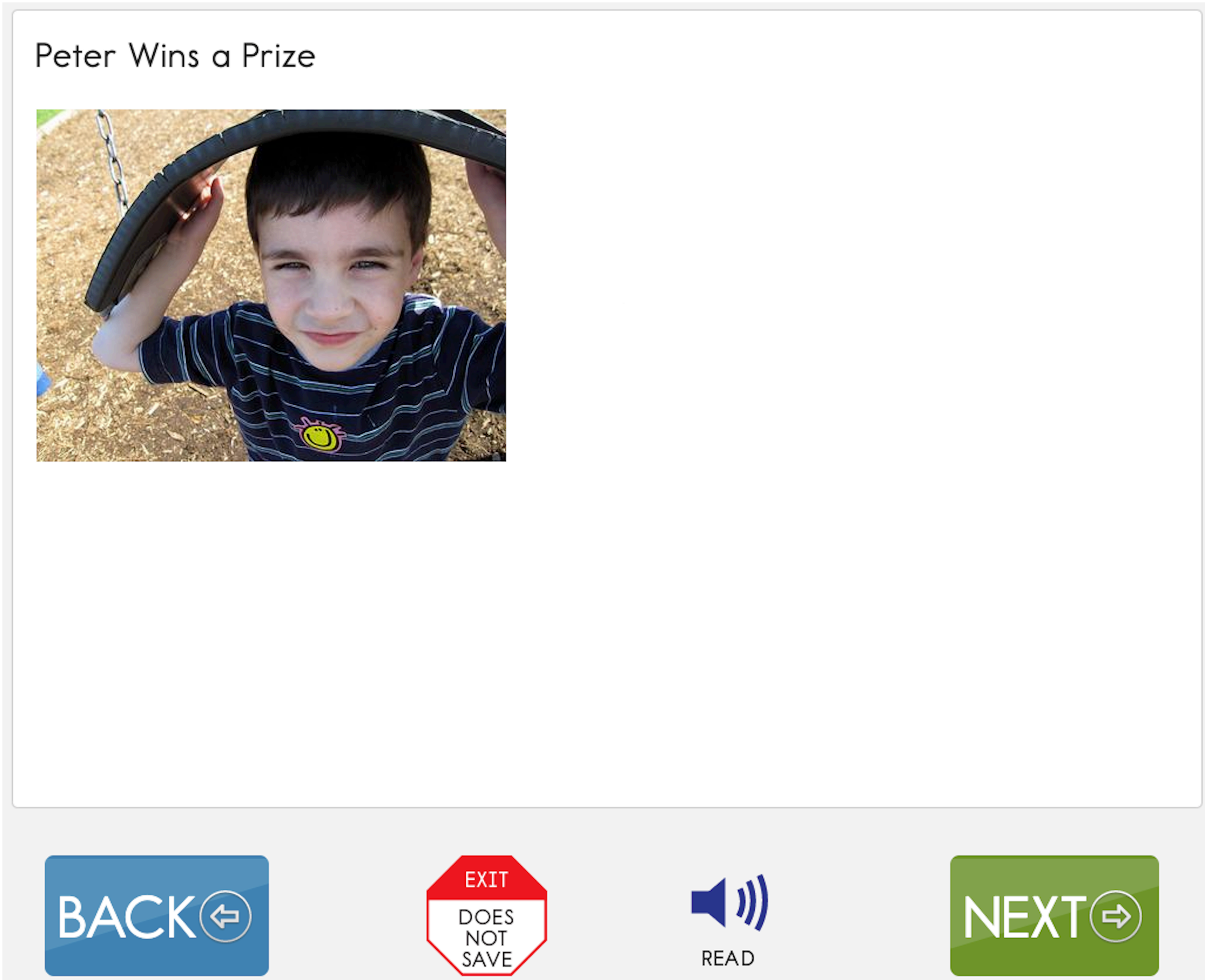
4.1.1.2. Kite Student Portal

Kite Student Portal is the platform that allows students to log in and complete assigned testlets. Practice activities and released testlets are also available to students and test administrators through Kite Student Portal (see Chapter 3 of this manual for more information). Kite Student Portal prevents students from accessing unauthorized content or software while taking assessments. Kite Student Portal is supported on devices running Windows or macOS (OSX), on Chromebooks, and on iPads.

Kite Student Portal provides students with a simple, web-based interface with student-friendly and intuitive graphics. The student interface used to administer the DLM assessments was designed specifically for students with the most significant cognitive disabilities. It maximizes space available to display content, decreases space devoted to tool-activation buttons (i.e., read aloud), and minimizes the cognitive load related to test navigation and response entry. An example of a screen used in an ELA testlet is shown in Figure 4.1. The blue **BACK** and green **NEXT** buttons are used to navigate between screens. The octagonal **EXIT DOES NOT SAVE** button allows the user to exit the testlet without recording any responses. The **READ** button plays an audio file of synthetic speech for the content on screen. Synthetic read aloud is the only accessibility feature with a tool directly enabled through each screen in the testlet. Further information regarding accessibility is provided in section 4.1.3 of this chapter.

Figure 4.1

An Example Screen From the Student Interface in Kite Student Portal



4.1.1.3. Local Caching Server

During DLM assessment administration, schools with unreliable network connections have the option to use the Local Caching Server (LCS). The LCS is a specially configured machine that resides on the local network and communicates between the testing machines at the testing location and the main testing servers for the DLM System. The LCS stores testing data from Kite Student Portal in an internal database; if the upstream network connection becomes unreliable or variable during testing, students can still continue testing, and their responses are transmitted to the Kite servers as bandwidth allows. The LCS submits and receives data to and from the DLM servers while the students are taking tests. The LCS must be connected to the internet between testlets to deliver the next testlet correctly.

4.1.2. Assessment Delivery Modes

The DLM System includes testlets designed to be delivered via computer directly to the student and testlets designed for the test administrator to administer outside the system and record responses in the system. The majority of testlets were developed for the computer-delivered mode because evidence suggested the majority of students with the most significant cognitive disabilities are able to interact directly with the computer or are able to access the content of the assessment on the computer with navigation assistance from a test administrator (Nash et al., 2016). Educator-administered testlets include all testlets at the Initial Precursor linkage level, some higher linkage level mathematics testlets requiring manipulatives, some alternate forms for students who are blind or who have visual impairments, and all writing testlets. A brief overview of the two types of testlets is included in the following sections. See Chapter 3 of this manual for a complete description of DLM testlets.

4.1.2.1. Computer-Delivered Assessments

Most DLM alternate assessments are delivered directly to students by computer through the Kite Suite. Computer-delivered assessments were designed so students can interact independently with the computer, using special assistive technology devices such as alternate keyboards, touch screens, or switches as necessary.

The computer-delivered testlets include various item types, including single-select multiple choice with three response options and text or images as response options, multiple choice multi-select with text or images as response options, matching items from two lists, sorting objects into categories, and highlighting selected text.

4.1.2.2. Educator-Administered Assessments

Some testlets were designed to be administered directly by the test administrator outside the Kite Suite. The Kite Suite delivers the testlet, but the test administrator is responsible for setting up the assessment, delivering it to the student, and recording student responses in Kite.

There are three general categories of educator-administered testlets.

1. Testlets with content designed for students who are developing symbolic understanding or who may not yet demonstrate symbolic understanding (Initial Precursor and some Distal Precursor).
2. Some mathematics testlets at higher linkage levels for which representing the content online would make the task too abstract and introduce unnecessary complexity to the item. Manipulatives are often used in this case, especially for students with blindness or visual impairment.
3. All writing assessments.

All three types of educator-administered testlets have some common features, which are described in Chapter 3 of this manual.

4.1.3. Accessibility

The DLM System was designed to be optimally accessible to diverse learners through accessible content (see Chapter 3 of this manual) as well as through initialization and routing driven by the First Contact survey and prior performance (see section 4.2 of this chapter for details). The interface in the Kite Suite was also designed to be easy to use to support accessibility. Consistent with the DLM learning map and item and test development practices described in earlier chapters (see Chapter 2 and Chapter 3, respectively), principles of universal design for assessment were applied to administration procedures and platforms. Decisions were largely guided by universal design for assessment principles of flexibility of use and equitability of use through multiple means of engagement, multiple means of representation, and multiple means of action and expression.

In addition to these considerations, a variety of accessibility supports are made available in the DLM assessment system. The *Accessibility Manual* (DLM Consortium, 2021b) outlines a six-step process for test administrators and Individualized Education Program (IEP) teams to use in making decisions about accessibility supports. This process begins with confirming the student meets the DLM participation guidelines and continues with the selection, administration, and evaluation of the effectiveness of accessibility supports. Test administrators select supports for each student in the PNP. The PNP can be completed any time before beginning testing. It can also be changed during testing as a student's needs change. Once updated, the changes appear the next time the student is logged in to the Kite Suite. All test administrators are trained in the use and management of these features.[23]

4.1.3.1. Overview of Accessibility Supports

Accessibility supports considered appropriate to use during administration of computer-delivered and educator-administered testlets are listed in the *Accessibility Manual* (DLM Consortium, 2021b). A brief description of the supports is provided here (see the *Accessibility Manual* for a full description of each support and its appropriate use). Supports are grouped into three categories: those provided through the PNP, those requiring additional tools or materials, and those provided outside the system. Additional techniques that are traditionally thought of as accommodations are considered allowable practices in the DLM assessment system. These are described in a separate section below.

4.1.3.1.1. Category 1: Supports Provided Within the DLM System via the PNP

Online supports include magnification, invert color choice, color contrast, and overlay color. Educators can test these options in advance to make sure they are compatible and provide the best access for students. Test administrators can adjust the PNP-driven accessibility during the assessment, and the selected options are then available the next time the student logs in to Kite Student Portal.

- **Magnification**. Magnification allows educators to choose the amount of screen magnification during testing.
- **Invert color choice**. In invert color choice, the background is black and the font is white.
- **Color contrast**. The color contrast allows educators to choose from several background and lettering color schemes.
- **Overlay color**. The overlay color is the background color of the test.

4.1.3.1.2. Category 2: Supports Requiring Additional Tools or Materials

These supports include braille, switch system preferences, iPad administration, and use of special equipment and materials. These supports are all recorded in the PNP even though the one-switch system is the only option actually activated by the PNP.

- **Uncontracted braille**. Uncontracted braille testlets are available during the testing window for

[23] See Chapter 9 for a complete description of test administrator training.

grades 3–5 at the Target and Successor levels and for grades 6 through high school at the Proximal Precursor, Target, and Successor levels. The standard delivery method is to deliver braille-ready files electronically to the school or district for local embossing as each testlet is assigned. The Kite Suite also delivers the identical general testlet form. After the student takes the testlet in its embossed form, the test administrator transfers the student's answers into Kite Student Portal.

- **Single-switch system**. Single-switch scanning is activated using a switch set up to emulate the Enter key on the keyboard. Scan speed, cycles, and initial delay may be configured.
- **Two-switch system**. Two-switch scanning does not require any activation in the PNP. Kite Student Portal automatically supports two-switch step scanning.
- **Administration via iPad**. Students may take the assessment via iPad.
- **Adaptive equipment used by student**. Test administrators may use any familiar adaptive equipment needed for the student.
- **Individualized manipulatives**. Individualized manipulatives are suggested for use with students rather than requiring educators to have a standard materials kit. Recommended materials and rules governing materials selection or substitution are described in the TIP (see section 4.3.1.2.1 of this chapter for more information on TIPs). Having a familiar concrete representation ensures that students are not disadvantaged by objects that are unfamiliar or that present a barrier to accessing the content.
- **BVI forms**. Alternate forms for students who are blind or have visual impairments (BVI) but do not read braille were developed for certain EEs and linkage levels. BVI testlets are educator-administered, requiring the test administrator to engage in an activity outside the system and enter responses into Kite Student Portal. The general procedures for administering these forms are the same as with other educator-administered testlets. Additional instructions include the use of several other supports (e.g., human read aloud, test administrator response entry, individualized manipulatives) as needed. When onscreen materials are being read aloud, test administrators are instructed to (1) present objects to the student to represent images shown on the screen, and (2) change the object language in the testlet to match the objects being used. Objects are used instead of tactile graphics, which are too abstract for the majority of students with the most significant cognitive disabilities who are also blind. However, test administrators have the option to use tactile graphics if their student can use them fluently.

4.1.3.1.3. Category 3: Supports Provided Outside the DLM System

These supports require actions by the test administrator, such as reading the test, signing or translating, and assisting the student with entering responses.

- **Human read aloud**. The test administrator may read the assessment to the student. Test administrators are trained to follow guidance to ensure fidelity in the delivery of the assessment. This guidance includes the typical tone and rate of speech, as well as avoiding emphasizing the correct response or important information that would lead the student to the correct response. Test administrators are trained to avoid facial expressions and body language that may cue the correct response and to use exactly the words on screen, with limited exceptions to this guideline, such as the use of shared reading strategies on the first read in ELA testlets. Finally, guidance includes ensuring that answer choices are always read in the same order as presented on the screen, with comprehensive examples of all item types. For example, when answer choices are in a triangle order, they are read in the order of top center, bottom left, and bottom right. In most cases, test administrators are allowed to describe graphics or images to students who need those described. Typically, this additional support is provided to students who are blind or have visual impairments. Alternate text for graphics and images in each testlet is included in the TIP as an attachment after the main TIP information. Test administrators who need to read alternate text have the Kite Suite open and the TIP in front of them while testing so they can accurately read the alternate text provided on the TIP with the corresponding screen. Human read aloud is allowed in either subject. The reading EEs included in the blueprints focus on comprehension of narratives and informational texts, not decoding. The read aloud support is available to any student who can benefit from decoding support in order to demonstrate the comprehension skills in the tested EEs.

- **Sign interpretation of text**. If the student requires sign language to understand the text, items, or instructions, the test administrator is allowed to use the words and images on the screen as a guide while signing for the student using American Sign Language, Signed Exact English, or any individualized signs familiar to the student. The test administrator is also allowed to spell unfamiliar words when the student does not know a sign for that word and accept responses in the student's sign language system. Sign is not provided via human or avatar video because of the unique sign systems used by students with the most significant cognitive disabilities who are also deaf/hard of hearing.
- **Language translation of text**. The DLM assessment system does not provide translated forms of testlets because of the unique cognitive and communication challenges for students taking DLM alternate assessments and because students who are English learners speak such a wide variety of languages; providing translated forms appropriate for all DLM-eligible students to cover the entire blueprint would be nearly impossible. Instead, test administrators are supplied with instructions regarding supports they can provide based on (1) each student's unique combination of language-related and disability-related needs, and (2) the specific construct measured by a particular testlet. For students who are English learners or who respond best to a language other than English, test administrators are allowed to translate the text for the student. The TIP includes information about exceptions to the general rule of allowable translation. For example, when an item assesses knowledge of vocabulary, the TIP includes a note that the test administrator may not define terms for the student on that testlet. Unless exceptions are noted, test administrators are allowed to translate the text for the student, simplify test instructions, translate words on demand, provide synonyms or definitions, and accept responses in either English of the student's native language.
- **Test administrator enters responses for student**. During computer-delivered assessments, if students are unable to physically select their answer choices themselves due to a gap between their accessibility needs/supports and the Kite Suite, they are allowed to indicate their selected responses to the test administrator through their typical communication modes (e.g., eye gaze, verbal). The test administrator then enters the response. The *Test Administration Manual* provides guidance on the appropriate use of this support to avoid prompting or misadministration. For example, the test administrator is instructed not to change tone, inflection, or body language to cue the desired response or to repeat certain response options after an answer is provided. The test administrator is also instructed to ensure the student continues to interact with the content on the screen.
- **Partner-assisted scanning**. Partner-assisted scanning is a commonly used strategy for students who do not have access to or familiarity with an augmentative or communication device or other

communication system. These students do not have verbal expressive communication and are limited to response modes that allow them to indicate selections using responses such as eye gaze. In partner-assisted scanning, the communication partner (the test administrator in this case) "scans" or lists the choices that are available to the student, presenting them in a visual, auditory, tactual, or combined format. For test items, the test administrator might read the stem of an item to the student and then read the answer choices aloud in order. In this example, the student could use a variety of response modes to indicate a response. Test administrators may repeat the presentation of choices until the student indicates a response.

4.1.3.2. Additional Allowable Practices

The Kite Student Portal user interface was specially designed for students with the most significant cognitive disabilities. Testlets delivered directly to students via computer were designed to facilitate students' independent interaction with the computer, using special devices such as alternate keyboards, touch screens, or switches as necessary. However, because computerized testing was new to many students using the DLM alternate assessment, the DLM Governance Board recognized that students would need various levels of support to interact with the computer. Test administrators are provided general principles for the allowable practices when the supports built into the system do support a student's completely independent interaction with the system.

To help make decisions about additional supports for computer-delivered testlets, test administrators receive training to follow two general principles. First, students are expected to respond to the content of the assessment independently. No matter which additional supports IEP teams and test administrators selected, all should be chosen with the primary goal of student independence at the forefront. Even if more supports are needed to provide physical access to the computer-based system, students should be able to interact with the assessment content and use their normal response modes to indicate a selection for each item. Second, test administrators are to ensure that students are familiar with the chosen supports. Ideally, any supports used during assessment are also used consistently during routine instruction. Students who have never received a support prior to the testing day are unlikely to know how to make the best use of the support.

In order to select the most appropriate supports during testing, test administrators are encouraged to use their best professional judgment and to be flexible while administering the assessment. Test administrators are allowed to use additional supports beyond PNP options. The supports detailed below in Table 4.1 are allowed in all computer-delivered and educator-administered testlets unless exceptions are noted in the TIP.

Table 4.1

Additional Allowable Practices

Practice	Explanation
Breaks as needed	Students can take breaks during or between testlets. Test administrators are encouraged to use their best judgment about the use of breaks. The goal should be to complete a testlet in a single session, but breaks are allowed if the student is fatigued, disengaged, or having behavioral problems that can interfere with the assessment. Kite Student Portal allows for up to 90 minutes of inactivity without timing out so that test administrators and students can pause for breaks during testlet administration. In cases in which administration begins but a short break is not sufficient for the student, the **EXIT DOES NOT SAVE** button can be used to exit the testlet (see Figure 4.1). The test administrator and student can then return to it and start over at another time.
Individualized student response mode†	The nodes assessed in the educator-administered testlets do not limit responses to certain types of expressive communication; therefore, all response modes are allowed. Test administrators can represent answer choices outside the system to maximize the student's ability to respond. For example, for students who use eye gaze to communicate, test administrators can represent the answer choices in an alternate format or layout to ensure the student can indicate a clear response.
Use of special equipment for positioning	For students who need special equipment to access the test material such as a slant board for positioning or Velcro objects on a communication board, test administrators are encouraged to use the equipment to maximize the student's ability to provide a clear response.
Navigation across screens	For students who have limited experience with, motor skills for, and/or devices for interacting directly with the computer, the test administrator can assist students to navigate across screens or enter the responses.
Use of interactive whiteboard	If the student has a severe visual impairment and needs larger presentation of content than the highest magnification setting provides, the test administrator can use an interactive whiteboard or projector or a magnification device that works with the computer screen to enlarge the assessment to the needed size.
Represent the answer options in an alternate format	Representing the answer options in an alternate format is allowed as long as the representation does not favor one answer choice over another. For instance, if the test administrator is presenting the answer choices to a student on a communication board or using objects to represent the answer choices, the correct answer choice cannot always be closest to the student or in the same position each time.
Use of graphic organizers	If the student is accustomed to using specific graphic organizers, manipulatives, or other tools during instruction, the use of those tools is allowable during the DLM alternate assessment.

Table 4.1

Additional Allowable Practices (continued)

Practice	Explanation
Use of blank paper	If the student requires blank, lined, or unlined paper, this can be provided. Once there is any writing on the paper, it becomes a secure testing document and needs to be disposed of and shredded at the conclusion of the testing session.
Generic definitions	If the student does not understand the meaning of a word used in the assessment, the test administrator can define the term generically and allow the student to apply that definition to the problem or question in which the term is used. Exceptions to this general rule are noted in the TIP for specific testlets.

† Allowed using speech, sign, or language translation unless prohibited for a specific testlet.

Although there are many supports and practices allowable for computer-delivered and educator-administered testlets, there are also practices that test administrators are trained to avoid, including the following:

- Repeating the item activity again after a student has responded or in any other way prompting the student to choose a different answer
- Using physical prompts or hand-over-hand guidance to the correct answer
- Removing answer choices or giving hints to the student
- Rearranging objects to prompt the correct answer—for example, putting the correct answer closer to the student

Test administrators are encouraged to ask any questions regarding whether a support is allowable via the DLM Service Desk or through their state education agency.

4.2. Key Features of the Year-End Assessment Model

As briefly described in Chapter 1, the DLM assessment system has two available models. This manual describes the Year-End assessment model. Consistent with the DLM Theory of Action described in Chapter 1, the DLM assessment administration features reflect multidimensional, non-linear, and diverse ways that students learn and demonstrate their learning. Test administration procedures therefore use multiple sources of information to assign testlets, including student characteristics, prior performance, and educator judgment.

In the Year-End model, the DLM System is designed to assess student learning at the end of the year. All testlets are administered in the spring assessment window; however, optional instructionally embedded testlets are available throughout the fall and winter. The instructionally embedded assessments, if administered, do not contribute to summative scoring. This assessment model yields summative results based only on testlets completed during the spring assessment window.

With the exception of writing testlets, each testlet contains items for one EE and one linkage level. In reading and mathematics, items in a testlet are aligned to nodes at one of five linkage levels for a single

EE. Writing testlets cover multiple EEs and are delivered at one of two levels: emergent (which corresponds with Initial Precursor and Distal Precursor linkage levels) or conventional (which corresponds with Proximal Precursor, Target, and Successor linkage levels).

This section describes the features of the Year-End assessment model, including initialization, adaptive routing, and test administration windows.

4.2.1. Testlet Assignment

This section describes how testlets are assigned to students during the spring assessment window. Educators complete the First Contact survey, which is used to assign the linkage level of the first testlets in each subject. The linkage level for subsequent testlets is determined by an adaptive routing algorithm.

4.2.1.1. First Contact Survey

The First Contact survey is a survey of learner characteristics that covers a variety of areas, including communication, academic skills, attention, and sensory and motor characteristics. A completed First Contact survey is required for each student prior to the assignment of testlets.

The items on the First Contact survey are categorized into the following sections:

- Special Education
- Sensory Capabilities
- Motor Capabilities and Health
- Computer Instruction
- Communication (Expressive and Receptive)
- Language
- Academics

Four sections of the First Contact survey are used to assign students to complexity bands in reading, mathematics, and writing: Expressive Communication, Reading Skills, Mathematics Skills, and Writing Skills. For expressive communiction, reading, and mathematics, there are four complexity bands (from lowest to highest): Foundational, Band 1, Band 2, and Band 3. In writing, there are two complexity bands (from lowest to highest): Emergent and Conventional. First Contact survey items used for determining complexity bands are included in Appendix C.1. Based on the educator's responses, the student's assigned complexity band is automatically calculated and stored in the system.

- For the ELA reading testlets, Kite Suite uses the responses from the Expressive Communication and Reading Skills questions to assign a student to one of four complexity bands.
- For the mathematics testlets, Kite Suite uses the responses from the Expressive Communication and Math Skills questions to assign a student to one of four complexity bands.
- For writing testlets, Kite Suite uses the responses from the Writing Skills question to assign a student to one of two complexity bands.

For reading and mathematics, if a different complexity band is indicated between the two sets of questions (Expressive Communication and the subject area questions), the system selects the lower band. The goal is to present a testlet that is approximately matched to a student's knowledge, skills, and understandings. That is, within reason, the system should recommend a testlet that is neither too easy nor too difficult and that provides a positive experience for the student entering the assessment. The correspondence among common student characteristics indicated on the First Contact survey, the corresponding First Contact complexity bands, and the recommended linkage levels are shown in Table 4.2.[24]

Table 4.2

Correspondence Among Student Characteristics Recorded on First Contact Survey, Complexity Bands, and Linkage Levels

Common First Contact survey responses about the student	First Contact complexity band	Linkage level
Does not use speech, sign, or augmentative and alternative communication; does not read any words when presented in print (reading); or does not sort objects (math)	Foundational	Initial Precursor
Uses one word, sign, or symbol to communicate; recognizes symbols (reading); or sorts symbols (math)	Band 1	Distal Precursor
Uses two words, signs, or symbols to communicate; reads at the primer to second grade level (reading); or adds/subtracts up to 80% of the time (math)	Band 2	Proximal Precursor
Regularly combines three or more spoken words to communicate for a variety of purposes; able to read print at the third-grade level or above (reading); or regularly add/subtract and form groups of objects (math)	Band 3	Target

The writing First Contact item is used to assign the two types of writing testlets: emergent and conventional. Students whose educators indicated they wrote by scribbling, copying or using word bands, or writing words corresponding to some sounds are assigned an emergent-level testlet. Students whose educator indicated they wrote words or simple phrases, sentences or complete ideas, or paragraph-length text without copying and using spelling are assigned the conventional writing testlet.

4.2.1.2. Initialization and Adaptive Routing

Each student is assigned as few as six to as many as nine testlets per subject during the spring assessment window. The number of testlets is determined by the assessment blueprint.[25] In mathematics, each testlet measures a single EE, so the number of testlets a student is assigned is equal to the number of EEs on the blueprint for the student's grade. The same is true for ELA, except that all writing EEs are measured on a single writing testlet. Thus, the number of testlets a student is assigned in ELA is equal to the number of non-writing EEs on the blueprint, plus one additional writing testlet. The system determines the linkage level for each testlet. The assignment is adaptive between testlets. Each spring testlet is

[24] For a description of linkage levels, see Chapter 2 of this manual.
[25] For a description of the assessment blueprints, see Chapter 2 of this manual.

packaged and delivered separately, and the test administrator determine when to schedule each testlet within the larger window.

The linkage level of the first testlet assigned to a student is based on First Contact survey responses. The correspondence between the First Contact complexity bands and first assigned linkage levels is shown in Table 4.3. Additionally, the level of the writing testlet for a student (i.e., emergent or conventional), is also assigned based on the First Contact survey, using the writing complexity band.

Table 4.3

Correspondence of Complexity Bands and Linkage Level

First Contact complexity band	Linkage level
Foundational	Initial Precursor
Band 1	Distal Precursor
Band 2	Proximal Precursor
Band 3	Target

The second and subsequent testlets are assigned based on the student's previous performance. That is, the linkage level associated with the next testlet a student receives is based on the student's performance on the previously administered testlet. The goal is to maximize the match of student knowledge, skills, and understandings to the appropriate linkage level content. Specific explanations of this process are as follows:

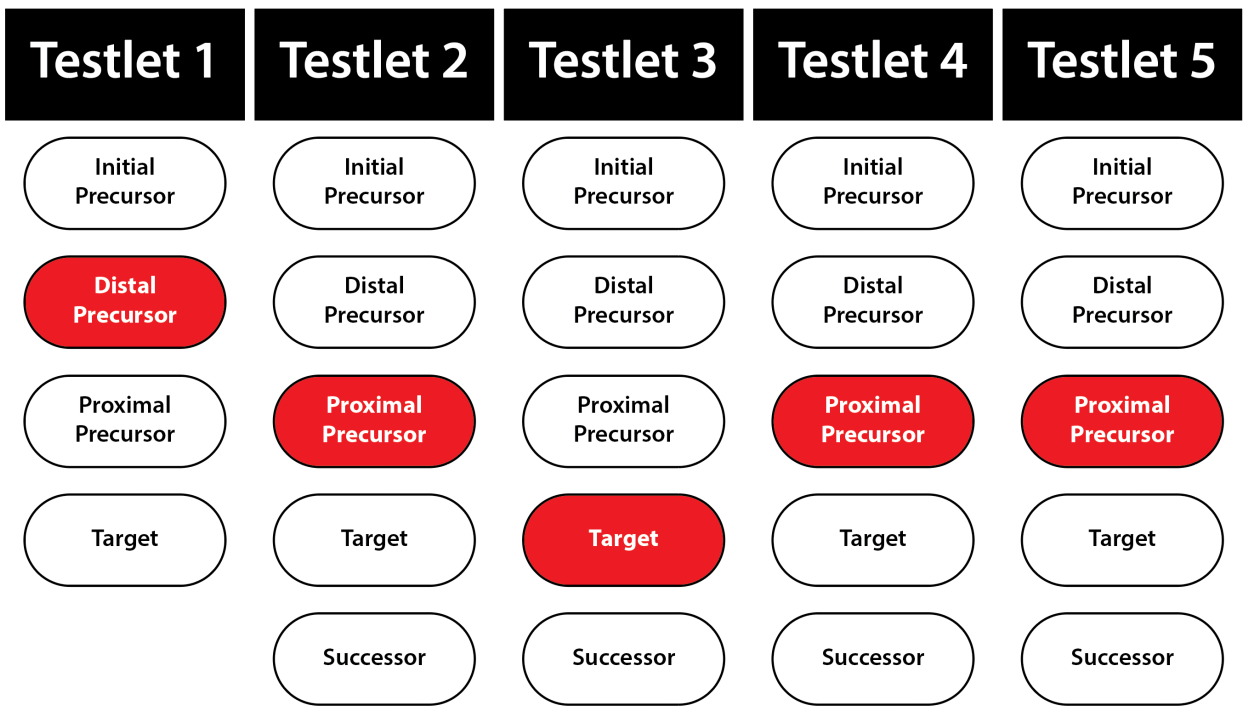
- The system adapts up one linkage level if students responded correctly to 80% or more of the items measuring the previously tested EE. If testlets were already at the highest level (i.e., Successor), they would remain there.
- The system adapts down one linkage level if students responded correctly to less than 35% of the items measuring the previously tested EE. If testlets were already at the lowest level (i.e., Initial Precursor), they would remain there.
- Testlets remain at the same linkage level if students responded correctly to between 35% and 80% of the items measuring the previously tested EE.

Threshold values for routing were selected with the number of items included in a testlet (typically three to five items) in mind. In a testlet that contains three items measuring the EE, if a student responds incorrectly to all items or correctly answers only one item (proportion correct less than .35), the linkage level of the testlet is likely too challenging. To provide a better match for the student's knowledge, skills, and understandings, the student would be routed to a lower linkage level. A single correct answer could be attributed to either a correct guess or true knowledge that did not translate to the other items measuring the EE. Similarly, if a student responds correctly to at least four items on a testlet with five items (proportion correct greater than .80) measuring the EE, the linkage level of the testlet is likely too easy. The student would be routed to a higher linkage level to allow the student the opportunity to demonstrate more advanced knowledge or skill. However, if the student responds to two of the three items correctly or three of five items correctly (proportion correct between .35 and .80), it cannot be assumed the student has completely mastered the knowledge, skills, or understanding being assessed at that linkage level. Therefore, the student is neither routed up nor down for the subsequent testlet.

Figure 4.2 provides an example of testlet adaptations for a student who completed five testlets. In the example, on the first assigned testlet at the Distal Precursor level, the student answered all of the items correctly, so the next testlet was assigned at the Proximal Precursor level. The next two testlets adapted up and down a level, respectively, whereas the fifth testlet remained at the same linkage level as the previous testlet.

Figure 4.2

Linkage Level Adaptations for a Student Who Completed Five Testlets



4.2.2. Assessment Administration Windows

Assessments are administered in the spring assessment window for operational reporting. Optional assessments are available during the instructionally embedded assessment window for educators to administer for formative information.

4.2.2.1. Instructionally Embedded Assessment Window

During the instructionally embedded assessment window, testlets are optionally available for test administrators to assign to their students. When choosing to administer the optional testlets during the instructionally embedded assessment window, educators decide which EEs and linkage levels to assess for each student. The assessment delivery system recommends a linkage level for each EE based on the educator's responses to the student's First Contact survey, but educators can choose a different linkage level based on their own professional judgment. In 2021–2022, the instructionally embedded assessment window occurred between September 13, 2021, and February 23, 2022. States were given the option of using the entire window or setting their own dates within the larger window. Across all states, the instructionally embedded assessment window ranged from 4–23 weeks.

4.2.2.2. Spring Assessment Window

During the spring assessment window, students are assessed on all of the EEs on the assessment blueprint in ELA and mathematics. The linkage level for each EE is determined by the system. In 2021–2022, the spring assessment window occurred between March 14, 2022, and June 10, 2022. States were given the option of using the entire window or setting their own dates within the larger window. Across all states, the spring assessment window ranged from 3–13 weeks.

4.3. Resources and Materials

Test administrators, school staff, district staff, and IEP teams are provided with multiple resources to support the assessment administration process.

Resources are provided on the DLM website and in the Kite Suite. Some states provide additional materials on their own customized landing page (i.e., dynamiclearningmaps.org/{statename}) of the DLM website and on their own department of education website. Test administrators are made aware of their state-specific webpage through training, manuals, webinars, and replies from Service Desk inquiries. The About DLM tab of the website includes information about topics related to the DLM System as a whole and may be of interest to a variety of audiences. To provide updates and reminders to all participating states, the DLM website also features a Test Updates section of the homepage. This is a newsfeed-style area that addresses timely topics such as assessment deadlines, resource updates, and system status. Additionally, the Test Updates page offers educators the option to subscribe to an electronic mailing list to automatically receive the same message via email without visiting the website. The DLM website also provides resources that cover assessment administration training information; student and roster data management; test delivery protocols and setup; and accessibility features, protocols, and documentation.

This section provides an overview of resources and materials available for test administrators and district-level staff.

4.3.1. Test Administrator Resources

While some resources for test administrators are available in the Kite Suite, the majority of DLM resources are available on the DLM website.

4.3.1.1. Test Administrator Resources Provided on the DLM Website

The DLM website provides specific resources designed for test administrators. These resources are available to all states (Table 4.4) to promote consistent assessment administration practices.

Table 4.4

DLM Resources for Test Administrators and States

Resource	Description
About Testlet Information Pages	Provides guidance for test administrators on the types and uses of information in the Testlet Information Pages provided for each testlet.
Accessibility Manual (PDF)	Provides guidance to state leaders, districts, educators, and Individualized Education Program (IEP) teams on the selection and use of accessibility supports available in the DLM System.
Guide to DLM Required Test Administrator Training (PDF)	Helps users access DLM Required Test Administrator Training in Moodle.
Guide to Practice Activities and Released Testlets (PDF)	Supports the test administrator in accessing practice activities and released testlets in Kite Student Portal.
Instructional Resources on the DLM Website	Provides links to additional resources for test administrators, including lists of EEs, a list of materials commonly needed for testlets, professional development modules supporting EEs, guidance on using mini-maps to plan instruction, accessing and using familiar texts, and released testlets and sample Testlet Information Pages.
Test Administration Manual (PDF)	Supports the test administrator in preparing themselves and students for testing.
Test Updates Page (webpage)	Breaking news on assessment administration activities. Users can sign up to receive alerts when new resources become available.
Training Video Transcripts (PDF)	Links to transcripts (narrator notes) for the DLM Required Test Administrator Training modules.

In addition, there are several helplet videos available on the DLM website[26] to support assessment administration:

- Accessibility in DLM Assessments
- Completing the First Contact Survey and PNP Profile
- DLM Instructionally Embedded Assessments
- DLM Writing Testlets

[26] https://dynamiclearningmaps.org/educator-resource-videos-ye

- Getting Started in Educator Portal
- Monitoring the Assessment Using Extracts
- More About Initial Precursor Items
- Overview of DLM ELA Testlets
- Overview of DLM Mathematics Testlets
- Test Tickets and TIPs in the Spring Window
- Using Kite Student Portal
- Using the DLM Instruction and Assessment Planner
- Verifying Rosters for Teachers
- Verifying Student Data for Teachers

4.3.1.2. Test Administrator Resources Provided in Kite Suite

The resources for test administrators that are provided in the Kite Suite include the TIPs as well as the practice activities and released testlets.

4.3.1.2.1. Testlet Information Pages

TIPs provide test administrators with information specific to each testlet. Test administrators receive a TIP in Educator Portal for each testlet after it is assigned to a student, and they are instructed to review the TIP before beginning the student's assessment.

Each TIP states whether a testlet is computer-delivered or educator-administered and indicates the number of items on the testlet. The TIP also provides information for each testlet regarding the materials needed, including substitute materials allowed.

The TIP also provides information on the exceptions to allowable supports. While a test administrator typically uses all appropriate PNP features and other flexibility tools described in the Allowable Practices section of the *Test Administration Manual*, the TIP indicates when it is not appropriate to use a support on a specific testlet. This may include limits on the use of definitions, translation, read aloud, calculators (for mathematics testlets), or other supports.

If there are further unique instructions for a given testlet, they are provided in the TIP. For test administrators who deliver human read aloud that includes descriptions of graphics, alternate text descriptions of images are provided.

TIPs for ELA testlets also provide the name of the text used in the testlet, identify the text as informational or literature, and label the text as familiar or unfamiliar. They also include the name of the grade-level text that the DLM text is associated with and note if assessment administration time is expected to be longer than usual because the linkage level requires a comparison between two texts. TIPs for mathematics testlets also provide information on specific mathematics terminology.

Testlets that require special setup before assessment administration begins, such as mathematics testlets designed for students with blindness or visual impairments, have additional instructions.

4.3.1.2.2. Practice Activities and Released Testlets

Practice activities and released testlets are available to support test administrators and students as they prepare for testing.

- The educator practice activity is designed to teach test administrators how to deliver educator-administered testlets, while the student practice activity is designed to teach students about the testlets and item features in the Kite Suite.
- The released testlets are similar to operational DLM testlets in content and format and are designed to be used for practice.

For more information on practice activities and released testlets, see Chapter 3 of this manual.

4.3.2. District-Level Staff Resources

Resources are available for three district-level supporting roles: Assessment Coordinator, Data Manager, and Technology Personnel. The Assessment Coordinator oversees the assessment process, which includes managing staff roles and responsibilities, developing and implementing a comprehensive training plan, developing a schedule for test implementation, monitoring and supporting test preparations and administration, and developing a plan to facilitate communication with parents or guardians and staff. The Data Manager manages educator, student, and roster data. Technology Personnel verify that network and testing devices are prepared for assessment administration.

Resources for each of these roles are made available on the state's customized DLM webpage. Each role has its own manual. A prerecorded training addressing each role and a FAQ compiled from Q&A sessions are also provided. Each role is also guided to supporting resources for other roles where responsibilities overlap. For example, Data Managers are guided to the *Test Administration Manual* to support data-related activities that are assigned to the test administrator and connect to troubleshooting data issues experienced by the test administrator. Technology Personnel are also guided to the Kite and Educator Portal webpage for information and documents connected to Kite Student Portal, Local Caching Server use, supported browsers, and bandwidth requirements. Assessment Coordinators are also guided to resources developed for the Data Manager, Technology Personnel, and test administrators for specific information and supplemental knowledge of the responsibilities of each of those roles. Some of those resources include the *Guide to DLM Required Test Administrator Training*, the *Test Administration Manual*, the Test Updates webpage, and electronic mailing lists.

Descriptions of training for district-level roles are provided in Chapter 9 of this manual.

4.4. Test Administrator Responsibilities and Procedures

The *Test Administration Manual* (DLM Consortium, 2021a) describes procedures for test administrators, which are organized into four sets of tasks for different parts of the school year: (1) before assessments, (2) during the instructionally embedded assessment window, (3) during the spring assessment window, and (4) while preparing for the next year.

4.4.1. Before Beginning Assessments

Test administrators are directed to perform multiple steps to prepare for student testing, including confirming student eligibility to participate in the DLM alternate assessment and sharing information about the assessment with parents to prepare them for their child's testing experience. Test administrators are also directed to review the *Test Administration Manual* and become familiar with available resources, including state webpages, practice activities and released testlets, and procedures for preparing to give the assessment.

1. The manual directs test administrators to prepare for the computer-delivered aspects of the assessment system. Test administrators must activate their Kite Educator Portal account, complete the Security Agreement in Kite Educator Portal, and complete the DLM Required Test Administrator Training (see Chapter 9 of this manual). Test administrators review their state's guidance on required and recommended professional development modules.
2. Test administrators are also directed to review the *Accessibility Manual* (DLM Consortium, 2021b) and work with IEP teams to determine what accessibility supports should be provided for each student taking the DLM assessments. Test administrators record the chosen supports in the PNP in Kite Educator Portal. Test administrators are also directed to review their state's requirements for documentation of DLM accessibility supports as testing accommodations and adjust the testing accommodations in the IEP as necessary.
3. Test administrators are also tasked with reviewing student data, including student demographic information and roster data in Kite Educator Portal, for accuracy. Test administrators also must ensure that the PNP and the First Contact survey are updated and complete in Kite Educator Portal. Test administrators must ensure that the Kite Student Portal is installed on testing devices. They must also make sure that they are familiar with their role as test administrator and the students are familiar with DLM testlets by utilizing the practice activities and released testlets. Finally, test administrators must check student devices for compatibility with Kite Student Portal.

4.4.2. Administration in the Instructionally Embedded and Spring Assessment Windows

In the optional instructionally embedded assessment window, test administrators choose appropriate EEs for instruction, retrieve instructional information for the EE, and select the EE and linkage level for the student in the Instruction and Assessment Planner. Test administrators deliver instruction until they determine the student is ready for assessment. Test administrators then confirm test assignment in the Instruction and Assessment Planner, retrieve the TIP, and gather necessary materials before beginning testing. They follow this step for each EE and linkage level they wish to assess during the optional instructionally embedded assessment window.

In the spring assessment window, EEs are delivered in a pre-specified order until all EEs on the blueprint have been assessed, with the linkage level determined by the system. Testlets can be administered throughout the assessment window, and test administrators are encouraged to administer the testlets when the students are ready to engage with the content.

In both windows, after the testlet has been assigned, test administrators assess the student on the testlet. While testing, users can go forward and backward within a testlet as much as needed before submitting answers. Student usernames and passwords are checked so that the students can access the assessments in Kite Student Portal.

4.4.3. Preparing for Next Year

Educators are directed to prepare for the following year by evaluating students' accessibility supports (PNP settings) with IEP teams and making decisions about supports and tools for next school year. They are also directed to review the blueprint for the next grade as a source of information to plan academic IEP goals.

4.5. Security

This section describes secure assessment administration, including test administrator training, security during administration, and the Kite Suite; secure storage and transfer of data; and plans for forensic analyses for the investigation of potential security issues. Test security procedures during item development and review are described in Chapter 3.

4.5.1. Training and Certification

Test security is promoted through the DLM Required Test Administrator Training and certification requirements for test administrators. Test administrators are expected to deliver DLM assessments with integrity and maintain the security of testlets. The training for assessment administration details test security measures. Each year, test administrators must renew their DLM Security Agreement through Kite Educator Portal (Figure 4.3). Test administrators are not granted access to Kite Educator Portal if they have not completed the Security Agreement.

Figure 4.3

Test Security Agreement Text

The Kite Suite provides opportunities for flexible assessment administration; however, all assessments delivered during the school year are secure.

Test administrators and other educational staff who support implementation are responsible for following the Kite test security standards:

1. Assessments (testlets) are not to be stored or saved on computers or personal storage devices; shared via email or other file sharing systems; or reproduced by any means.
2. Except where explicitly allowed as described in the Test Administration Manual, electronic materials used during assessment administration may not be printed.
3. Those who violate the Kite test security standards may be subject to their state's regulations or state education agency policy governing test security.
4. Educators are encouraged to use resources provided by Kite Suite, including practice activities and released testlets, to prepare themselves and their students for the assessments.
5. Users will not give out, loan or share their password with anyone. Allowing others access to an Educator Portal account may cause unauthorized access to private information. Access to educational records is governed by federal and state law.

Questions about security expectations should be directed to the local assessment coordinator.

☑ I have read this security agreement and agree to follow the standards.

Save

Although each state may have additional security expectations and security-related training requirements, all test administrators in each state are required to meet these minimum training and certification requirements.

4.5.2. Maintaining Security During Test Administration

Several aspects of the DLM System support test security and test administrator integrity during use of the system. Because TIPs are the only printed material, there is limited risk of exposure. Guidance is provided in the *Test Administration Manual* and on TIPs regarding allowable and not allowable practices. This guidance is intended to promote implementation fidelity and reduce the risk of cheating or other types of misadministration. For a description of fidelity to intended practices, see the description of test administration observations in section 4.7.1 of this chapter.

Agile Technology Solutions, the organization that develops and maintains the Kite Suite and provides DLM Service Desk support to test administrators in the field, has procedures in place to handle alleged security breaches (e.g., test content is made public). Any reported test security incident is assumed to be a breach and is handled accordingly. In the event of a test security incident, access is disabled at the appropriate level. Depending on the situation, the testing window could be suspended, or test sessions could be removed. Test forms could also be removed if exposed or if data is exposed by a form. If necessary, passwords would be changed for users at the appropriate level.

4.5.3. Security in the Kite Suite

The Kite Suite prioritizes security to ensure confidentiality, integrity, and availability for all application data. All Kite Suite data is housed within the United States, including application backups and recovery data. Kite Suite runs in Amazon Web Services (AWS) that implements a “Shared Responsibility” model as it pertains to security controls. AWS is responsible for the security of the cloud, which protects all the infrastructure and services that AWS offers. This is composed of the hardware, software, networking, and physical access to the facilities, and all of the security controls associated with those, including environmental and physical controls. Just as the responsibility to operate the IT environment is shared between AWS and its customers, so is the management, operation, and verification of the IT controls. AWS runs an extensive compliance program reflecting the depth and breadth of their security controls. AWS is NIST 800-53 and FedRAMP compliant. For the controls that are not covered by AWS, the Kite team aligns with NIST standards.

Application access and support access to Kite Suite data follows the principle of least privilege. Access to Kite Suite data is provided through role-based access control systems that limit data to be available to those individuals that require access to perform their jobs. Access is regularly audited by our documented daily, weekly, and monthly security checkout processes.

All Kite Suite network transmissions are encrypted to prevent interception, disruption of reception, communications deception, and/or derivation of intelligence by analysis of transmission characteristics such as signal parameters or message externals. All client web traffic is HTTPS encrypted, with support limited to modern, secure algorithms within the TLS 1.2 or greater protocol. This secures all communication during the session, including the authentication and authorization stages. Support sessions and data transfers are protected using Secure Shell (SSH), an encrypted protocol designed to give a secure connection over an insecure network, such as the internet. All internal network traffic is also encrypted to protect data in transit between network tiers and application components.

Intrusion prevention is a critical component of the Kite Suite security implementation. The Kite Suite implementation in AWS, Kite Suite security processes and procedures, and the Kite Suite development lifecycle all contribute to intrusion prevention.

All Kite Suite Windows Servers utilize Microsoft tools for antivirus, anti-malware, and software firewalls. All laptops and desktops for project staff are fully managed with current antivirus, anti-malware, and storage encryption.

To protect the integrity of test items and scoring materials, the Kite Test Security Agreement lists the security standards that all educators involved with administering tests must follow to protect both the student’s privacy as well as test items and scoring materials.

4.5.4. Secure Test Content

Test content is stored in Kite Content Builder. All items used for released testlets exist in a separate pool from items used for operational testing purposes, ensuring that no items are shared among secure and non-secure pools. Only authorized users of the Kite assessment system have access to view items. Testlet assignment logic prevents a student from being assigned the same testlet more than once, except in cases of manual override for test reset purposes.

4.5.5. Data Security

Project staff collect personally identifiable information (PII) protocols and usage rules from states. Project staff document any applicable state laws regarding PII, state PII handling rules, and state-specific PII breach procedures. The information is housed in the shared resources where Service Desk agents and other project staff can access the information as needed. The protocols are followed with precision due to the sensitive nature of PII and the significant consequences tied to breaches of the data.

The procedures that are implemented in the case of a security incident, privacy incident, or data breach that involve PII or sensitive personal information are implemented by an investigation team that focuses first on mitigation of immediate risk, followed by identification of solutions to identified problems and communication with the DLM Governance Board.

4.5.6. State-Specific Policies and Practices

Some states also adopt more stringent requirements for access to test content and for the handling of secure data, above and beyond those for the overall DLM System. Each DLM agreement with a state education agency (SEA) includes a Data Use Agreement. The Data Use Agreement addresses the data security responsibilities of DLM project staff in regard to the Family Educational Rights and Privacy Act (FERPA, Family Educational Rights and Privacy Act, 1974). The agreement details the role of Accessible Teaching, Learning, and Assessment Systems (ATLAS) as the holder of the data and the rights of the SEA as the owner of the data. In many cases, the standard Data Use Agreement is modified to include state-specific data security requirements. Project staff document these requirements for each state, and the Implementation and Service Desk teams implement the requirements.

The Implementation team collects state education authorities' policy guidance on a range of state policy issues such as individual student test resets, district testing window extensions, and allowable sharing of PII. In all cases, the needed policy information is collected on a state summary sheet and recorded in a software program jointly accessed by Service Desk agents and the Implementation team.

The Implementation team reviews the state testing policies during Service Desk agent training and provides updates during the state testing windows to supervisors of the Service Desk agents. As part of the training, the Service Desk agents are directed to contact the Implementation team with any questions that require state input or the state to develop or amend a policy.

4.6. Evidence from the DLM System

This section describes evidence collected by the DLM System during the 2021–2022 operational administration of the DLM alternate assessment. The categories of evidence include data relating to administration time, device usage, adaptive routing, and accessibility support selections.

4.6.1. Administration Time

Estimated administration time varies by student and subject. Testlets can be administered separately across multiple testing sessions as long as they are all completed within the testing window.

The published estimated total testing time per testlet is around 5–10 minutes in mathematics, 10–15 minutes in reading, and 10–20 minutes for writing. The estimated total testing time is 60–75 minutes per student in ELA and 35–50 minutes in mathematics in the spring assessment window. Published estimates are slightly longer than anticipated real testing times because of the assumption that test administrators need time for setup. Actual testing time per testlet varies depending on each student's unique characteristics.

Kite Student Portal captured start dates, end dates, and time stamps for every testlet. The difference between these start and end times was calculated for each completed testlet. Table 4.5 summarizes the distribution of test times per testlet. The distribution of test times in Table 4.5 is consistent with the distribution observed in prior years. Most testlets took around seven minutes or less to complete, with mathematics testlets generally taking less time than ELA testlets. Time per testlet may have been impacted by student breaks during the assessment (for more information about breaks, see the Accessibility section above). Testlets with shorter than expected administration times are included in an extract made available to each state. States can use this information to monitor assessment administration and address as necessary. For a description of the administration time monitoring extract, see section 4.7.4 of this chapter.

Table 4.5

Distribution of Response Times per Testlet in Minutes

Grade	Min	Median	Mean	Max	25Q	75Q	IQR
English language arts							
3	.033	3.92	4.84	88.80	2.58	5.95	3.37
4	.150	4.17	5.18	89.93	2.75	6.37	3.62
5	.100	4.25	5.23	88.93	2.78	6.51	3.72
6	.100	4.22	5.21	89.90	2.78	6.43	3.65
7	.117	4.95	5.99	88.62	3.17	7.50	4.33
8	.167	4.30	5.23	88.85	2.85	6.48	3.63
9	.217	4.72	5.93	88.55	2.90	7.30	4.40
10	.183	4.50	5.54	89.93	2.82	6.87	4.05
11	.183	5.03	6.41	89.50	3.12	7.90	4.78
12	.217	4.88	6.28	87.18	2.82	7.80	4.98
Mathematics							
3	.100	1.85	2.62	87.57	1.10	3.18	2.08
4	.067	1.45	2.09	85.32	0.90	2.42	1.52
5	.067	1.62	2.32	86.05	1.00	2.70	1.70
6	.050	1.68	2.36	86.30	1.07	2.77	1.70
7	.083	1.62	2.27	82.90	0.97	2.72	1.75
8	.050	1.60	2.30	88.53	0.98	2.70	1.72
9	.083	1.72	2.44	79.50	0.98	2.95	1.97
10	.067	1.60	2.29	85.87	0.97	2.70	1.73
11	.083	1.68	2.40	88.73	1.05	2.85	1.80
12	.083	1.65	2.41	48.90	0.98	2.87	1.88

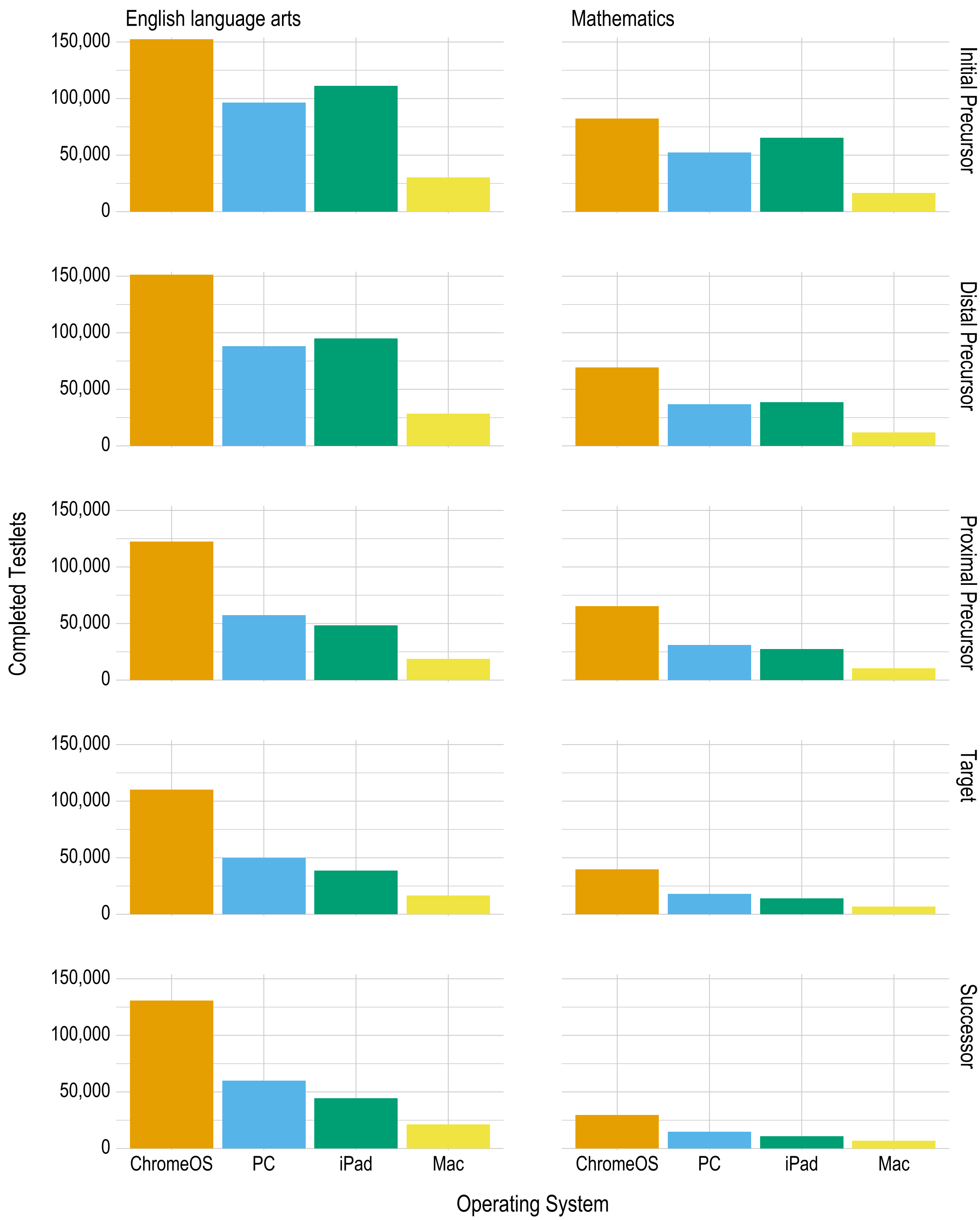
Note. Min = minimum, Max = maximum, 25Q = lower quartile, 75Q = upper quartile, IQR = interquartile range.

4.6.2. Device Usage

Testlets may be administered on a variety of devices. Kite Student Portal captured the operating system used for each testlet completed. Although these data do not capture specific devices used to complete each testlet (e.g., SMART Board, switch system, etc.), they provide high-level information about how students access assessment content. For example, we can identify how often an iPad is used relative to a Chromebook or traditional PC. Figure 4.4 shows the number of testlets completed on each operating system by subject and linkage level for 2021–2022. Overall, 45% of testlets were completed on a Chromebook, 24% were completed on a PC, 23% were completed on an iPad, and 8% were completed on a Mac.

Figure 4.4

Distribution of Devices Used for Completed Testlets



4.6.3. Blueprint Coverage

Each student is assessed on all EEs included on the assessment blueprint.[27] Table 4.6 summarizes the number of EEs required for each grade and subject.

Table 4.6

Essential Elements Required for Blueprint Coverage

Grade	English language arts (*n*)	Mathematics (*n*)
3	10	8
4	11	8
5	10	8
6	11	7
7	13	7
8	13	8
9	14	7
10	14	8
11	14	6

Across all grades, 96% of students in ELA and 97% of students in mathematics were assessed on all of the EEs and met blueprint requirements. Table 4.7 summarizes the total number of students and the percentage of students meeting blueprint requirements based on their complexity band for each subject. When comparing complexity band distributions, there was a slightly lower percentage of Foundational students not meeting requirements. However, all complexity band groups had over 93% of students meeting the coverage requirements.

Table 4.7

Student Blueprint Coverage by Complexity Band

Complexity Band	*n*	% meeting requirements
English language arts		
Foundational	12,014	93.9
Band 1	33,393	96.1
Band 2	32,876	97.1
Band 3	9,949	96.8
Mathematics		
Foundational	12,380	94.4
Band 1	33,181	96.4
Band 2	34,363	97.5
Band 3	8,128	97.9

[27] For a description of the assessment blueprints see Chapter 2 of this manual.

4.6.4. Adaptive Delivery

Following the spring 2022 administration, analyses were conducted to determine the mean percentage of testlets that adapted from the first to second testlet administered for students within a grade, subject, and complexity band. The aggregated results can be seen in Table 4.8 and Table 4.9 for ELA and mathematics, respectively.

For the majority of students across all grades who were assigned to the Foundational Complexity Band by the First Contact survey, testlets did not adapt to a higher linkage level after the first assigned testlet (ranging from 64% to 92% across both subjects). Consistent patterns were not as apparent for students who were assigned to Band 1, Band 2, or Band 3. Distributions across the three categories were more variable across grades and subjects. Results indicate that linkage levels of students assigned to higher complexity bands are more variable with respect to the direction in which students move between the first and second testlets. However, this finding of more variability in the higher complexity bands is consistent with prior years, which showed the same trend. Several factors may help explain these results, including more variability in student characteristics within this group and content-based differences across grades and subjects. Further exploration is needed in this area.

Table 4.8

*Adaptation of Linkage Levels Between First and Second English Language Arts Testlets (*N *= 88,232)*

Grade	Foundational		Band 1			Band 2			Band 3		
	Adapted up (%)	Did not adapt (%)	Adapted up (%)	Did not adapt (%)	Adapted down (%)	Adapted up (%)	Did not adapt (%)	Adapted down (%)	Adapted up (%)	Did not adapt (%)	Adapted down (%)
Grade 3	13.8	86.2	59.5	21.8	18.7	77.7	14.5	7.8	90.8	6.6	2.6
Grade 4	27.8	72.2	17.6	28.9	53.5	63.1	25.6	11.3	50.2	20.6	29.2
Grade 5	28.0	72.0	24.1	31.4	44.5	59.9	33.3	6.8	88.6	8.5	3.0
Grade 6	32.0	68.0	11.8	22.0	66.2	25.0	39.4	35.6	35.9	47.7	16.4
Grade 7	28.7	71.3	31.2	25.5	43.3	52.2	35.1	12.7	69.6	25.1	5.3
Grade 8	35.7	64.3	30.6	22.3	47.1	66.4	21.7	11.9	86.1	10.3	3.6
Grade 9	13.6	86.4	32.2	32.4	35.4	18.9	32.8	48.2	65.5	22.6	11.9
Grade 10	7.6	92.4	30.5	32.5	37.0	14.9	30.8	54.2	61.1	23.9	15.0
Grade 11	28.6	71.4	12.1	36.6	51.2	62.8	24.3	12.9	65.9	21.7	12.4
Grade 12	25.0	75.0	11.4	33.9	54.7	59.7	27.5	12.9	58.8	23.0	18.2

Note. Foundational is the lowest complexity band, so testlets could not adapt down a linkage level.

Table 4.9

*Adaptation of Linkage Levels Between First and Second Mathematics Testlets (*N *= 88,052)*

Grade	Foundational		Band 1			Band 2			Band 3		
	Adapted up (%)	Did not adapt (%)	Adapted up (%)	Did not adapt (%)	Adapted down (%)	Adapted up (%)	Did not adapt (%)	Adapted down (%)	Adapted up (%)	Did not adapt (%)	Adapted down (%)
Grade 3	12.6	87.4	11.2	31.4	57.4	20.1	52.3	27.7	74.6	15.7	9.7
Grade 4	14.5	85.5	19.6	33.1	47.3	66.1	26.2	7.7	78.2	18.1	3.7
Grade 5	16.3	83.7	13.2	30.9	55.9	40.8	25.9	33.3	66.3	22.9	10.8
Grade 6	17.3	82.7	14.6	42.9	42.5	28.5	36.2	35.3	49.1	43.1	7.8
Grade 7	17.3	82.7	12.9	27.5	59.7	19.8	20.9	59.3	72.5	19.2	8.3
Grade 8	15.3	84.7	15.7	49.4	34.9	30.5	54.8	14.7	49.7	22.5	27.8
Grade 9	14.5	85.5	21.6	48.0	30.4	55.8	36.1	8.0	57.9	34.5	7.6
Grade 10	14.1	85.9	22.6	30.4	47.0	34.2	20.6	45.1	3.8	21.4	74.8
Grade 11	27.7	72.3	11.0	28.2	60.8	25.8	37.0	37.2	16.0	12.3	71.7
Grade 12	25.5	74.5	6.9	30.7	62.5	23.5	38.3	38.3	15.2	14.5	70.3

Note. Foundational is the lowest complexity band, so testlets could not adapt down a linkage level.

After the second testlet is administered, testlets continue to adapt based on the same routing rules. Table 4.10 shows the total number and percentage of testlets that were assigned at each linkage level during the spring assessment window. Because writing testlets are not assigned at a specific linkage level, those testlets are not included in Table 4.10. Testlets were fairly evenly distributed across the five linkage levels, with slightly fewer assignments at the Target linkage level in ELA and slightly more assignments at the Initial Precursor linkage level in mathematics. There were fewer assignments at the Target and Successor levels.

Table 4.10

Distribution of Linkage Levels Assigned for Assessment

Linkage level	*n*	%
English language arts		
Initial Precursor	170,081	24.5
Distal Precursor	142,525	20.5
Proximal Precursor	134,753	19.4
Target	103,132	14.9
Successor	143,936	20.7
Mathematics		
Initial Precursor	215,969	33.5
Distal Precursor	155,915	24.2
Proximal Precursor	133,261	20.7
Target	78,259	12.1
Successor	61,274	9.5

4.6.5. Administration Incidents

DLM staff annually evaluates testlet assignment to ensure students are correctly assigned to testlets. Administration incidents that have the potential to affect scoring are reported to state education agencies in a supplemental Incident File. No incidents were observed during the 2021–2022 operational assessment windows. Assignment of testlets will continue to be monitored in subsequent years to track any potential incidents and report them to state education agencies.

4.6.6. Accessibility Support Selections

Table 4.11 shows selection rates for the three categories of accessibility supports. Each of the support categories are discussed in detail above in the Accessibility section. Overall, 85,429 students (88%) had at least one support selected. The most commonly selected supports in 2021–2022 were human read aloud, test administrator enters responses for student, and spoken audio. Additionally, educators reported in the First Contact survey (see section 4.2.1.1 of this chapter) that 40% of students were able to access a computer independently, with or without assistive technology.

Table 4.11

Accessibility Supports Selected for Students (N = 96,576)

Support	*n*	%
Supports provided in Kite Student Portal		
Spoken audio	51,301	53.1
Magnification	13,471	13.9
Color contrast	8,059	8.3
Overlay color	3,647	3.8
Invert color choice	2,473	2.6
Supports requiring additional tools/materials		
Individualized manipulatives	40,406	41.8
Calculator	28,280	29.3
Single-switch system	3,767	3.9
Alternate form - visual impairment	2,068	2.1
Two-switch system	1,165	1.2
Uncontracted braille	109	0.1
Supports provided outside the system		
Human read aloud	75,846	78.5
Test administrator enters responses for student	52,439	54.3
Partner-assisted scanning	8,490	8.8
Language translation of text	1,569	1.6
Sign interpretation of text	1,218	1.3

4.7. Evidence From Monitoring Assessment Administration

Monitoring of assessment administration was conducted using various materials and strategies. DLM project staff developed an assessment administration monitoring protocol for use by DLM staff, state education agency staff, and local education agency staff. Project staff also reviewed Service Desk contacts and hosted regular check-in calls to monitor common issues and concerns during the assessment window. This section provides an overview of all resources and supports as well as more detail regarding the assessment administration observation protocol and its use, check-in calls with states, and methods for monitoring testlet delivery.

4.7.1. Test Administration Observations

DLM project staff developed an assessment administration observation protocol to standardize data collection across observers and locations. This assessment administration protocol is available for use by state and local education agencies; however, participation in the test administration observations is not required. The majority of items in the protocol are based on direct recording of what is observed and require little inference or background knowledge. Information from the protocol is used to evaluate several assumptions in the validity argument, addressed in the Test Administration Observation Results section of this chapter.

One observation form is completed per testlet administered. Some items are differentiated for computer-delivered and educator-administered testlets. The four main sections include Preparation/Set Up, Administration, Accessibility, and Observer Evaluation. The Preparation/Set Up section includes documentation of the testing location, testing conditions, the testing device used for the testing session, and documentation of the test administrator's preparation for the session. The Administration section is provided for the documentation of the student's response mode, general test administrator behaviors during the session, subject-specific test administrator behaviors, any technical problems experienced with the Kite Suite, and documentation of student completion of the testlet. The Accessibility section focuses on the use of accessibility features, any difficulty the student encountered with the accessibility features, and any additional devices the student uses during the testing session. Finally, Observer Evaluation requires that the observer rate overall student engagement during the session and provide any additional relevant comments.

The protocol is available as an online survey (optimized for mobile devices and with branching logic) administered through Kite Survey Solutions, a survey platform within the Kite Suite.

Training resources are provided to state education agency staff to support fidelity of use of the assessment administration protocol and increase the reliability of data collected (see Table 4.12). State education agency staff have access to the *Test Administration Observation Training* video on the use of the *Test Administration Observation Protocol*. The links to this video, the *Guidance for Local Observers*, and the *Test Administrator Observation Protocol* are provided on the state side of the DLM website, and state education agencies are encouraged to use this information in their state monitoring efforts. State education agencies are able to use these training resources to encourage use of the protocol among local education agency staff. States are also cautioned that the protocol is only to be used to document observations for the purpose of describing the administration process. It is not to be used for evaluating or coaching test administrators or gauging student academic performance. This caution, as well as general instructions for completing and submitting the protocol, are provided in the form itself.

Table 4.12

DLM Resources for Test Administration Monitoring Efforts

Resource	Description
DLM Test Administration Observation Research Protocol (PDF)	Provides observers with a standardized way to describe the assessment administration.
Guide to Test Administration Observations: Guidance for Local Observers (PDF)	Provides observers with the purpose and use of the observation protocol as well as general instructions for use.
Test Administration Observation Training Video (Vimeo video)	Provides training on the use of the *Test Administration Observation Protocol*.

During 2021–2022, there were 157 assessment administration observations collected in six states. Table 4.13 shows the number of observations collected by state. Of the observations, 115 (73%) were of computer-delivered assessments and 42 (27%) were of educator-administered testlets. The observations consisted of 84 (54%) ELA reading testlets, 8 (5%) ELA writing testlets, and 65 (41%) mathematics testlets.

Table 4.13

*Educator Observations by State (*N *= 157)*

State	*n*	%
Arkansas	70	44.6
Iowa	15	9.6
Kansas	4	2.5
Missouri	13	8.3
North Dakota	1	0.6
West Virginia	54	34.4

To investigate the assumptions that underlie the claims of the validity argument, several parts of the test administration observation protocol were designed to provide information corresponding to the assumptions. One assumption addressed is that educators allow students to engage with the system as independently as they are able. For computer-delivered testlets, related evidence is summarized in Table 4.14; behaviors were identified as supporting, neutral, or nonsupporting. For example, clarifying directions (51.3% of observations) removes student confusion about the task demands as a source of construct-irrelevant variance and supports the student's meaningful, construct-related engagement with the item. In contrast, using physical prompts (e.g., hand-over-hand guidance) indicates that the test administrator directly influenced the student's answer choice. Overall, 60% of observed behaviors were classified as supporting, with 2% of observed behaviors reflecting nonsupporting actions.

Table 4.14

Test Administrator Actions During Computer-Delivered Testlets (n = 115)

Action	*n*	%
Supporting		
Read one or more screens aloud to the student	76	66.1
Navigated one or more screens for the student	60	52.2
Clarified directions or expectations for the student	59	51.3
Repeated question(s) before student responded	33	28.7
Neutral		
Used pointing or gestures to direct student attention or engagement	42	36.5
Used verbal prompts to direct the student's attention or engagement (e.g., "look at this.")	39	33.9
Entered one or more responses for the student	21	18.3
Used materials or manipulatives during the administration process	17	14.8
Asked the student to clarify or confirm one or more responses	11	9.6
Repeated question(s) after student responded (gave a second trial at the same item)	10	8.7
Allowed student to take a break during the testlet	6	5.2
Nonsupporting		
Physically guided the student to a response	3	2.6
Reduced the number of answer choices available to the student	3	2.6

Note. Respondents could select multiple responses to this question.

For DLM assessments, interaction with the system includes interaction with the assessment content as well as physical access to the testing device and platform. The fact that educators navigated one or more screens in 52% of the observations does not necessarily indicate the student was prevented from engaging with the assessment content as independently as possible. Depending on the student, test administrator navigation may either support or minimize students' independent, physical interaction with the assessment system. While not the same as interfering with students' interaction with the content of the assessment, navigating for students who are able to do so independently conflicts with the assumption that students are able to interact with the system as intended. The observation protocol did not capture why the test administrator chose to navigate, and the reason was not always obvious.

A related assumption is that students are able to interact with the system as intended. Evidence for this assumption was gathered by observing students taking computer-delivered testlets, as shown in Table 4.15. Independent response selection was observed in 48% of the cases. Non-independent response selection may include allowable practices, such as test administrators entering responses for the student. The use of materials outside of Kite Student Portal was seen in 10% of the observations. Verbal prompts for navigation and response selection are strategies within the realm of allowable flexibility during test administration. These strategies, which are commonly used during direct instruction for students with the most significant cognitive disabilities, are used to maximize student engagement with the system and promote the type of student-item interaction needed for a construct-relevant response. However, they also indicate that students were not able to sustain independent interaction with the system throughout the entire testlet.

Table 4.15

*Student Actions During Computer-Delivered Testlets (*n *= 115)*

Action	*n*	%
Selected answers independently	55	47.8
Navigated screens independently	43	37.4
Selected answers after verbal prompts	30	26.1
Navigated screens after test administrator pointed or gestured	27	23.5
Navigated screens after verbal prompts	25	21.7
Used materials outside of Kite Student Portal to indicate responses to testlet items	12	10.4
Asked the test administrator a question	6	5.2
Revisited one or more questions after verbal prompt(s)	6	5.2
Independently revisited a question after answering it	3	2.6
Skipped one or more items	1	0.9

Note. Respondents could select multiple responses to this question.

Another assumption in the validity argument is that students are able to respond to tasks irrespective of sensory, mobility, health, communication, or behavioral constraints. This assumption was evaluated by having observers note whether there was difficulty with accessibility supports (including lack of appropriate available supports) during observations of educator-administered testlets. Of the 42 observations of educator-administered testlets, observers noted difficulty in 1 case (2%). For computer-delivered testlets, evidence to evaluate the assumption was collected by noting students who indicated responses to items using varied response modes such as gesturing (25%) and using manipulatives or materials outside of Kite (10%). Additional evidence for this assumption was gathered by observing whether students were able to complete testlets. Of the 157 test administration observations collected, students completed the testlet in 113 cases (72%).[28]

Finally, the test administration observations allow for an evaluation of the assumption that test administrators enter student responses with fidelity. To record student responses with fidelity, test administrators needed to observe multiple modes of communication, such as verbal, gesture, and eye gaze. Table 4.16 summarizes students' response modes for educator-administered testlets. The most frequently observed behavior was *verbally indicated response to test administrator who selected answers*.

[28] In all instances where the testlet was not completed, no reason was provided by the observer.

Table 4.16

Primary Response Mode for Educator-Administered Testlets (n = 42)

Response mode	*n*	%
Verbally indicated response to test administrator who selected answers	24	57.1
Gestured to indicate response to test administrator who selected answers	20	47.6
Eye gaze system indication to test administrator who selected answers	3	7.1
No observable response mode	2	4.8

Note. Respondents could select multiple responses to this question.

Computer-delivered testlets provided another opportunity to confirm fidelity of response entry when test administrators entered responses on behalf of students. This support is recorded on the PNP Profile and is recommended for a variety of situations (e.g., students who have limited motor skills and cannot interact directly with the testing device even though they can cognitively interact with the onscreen content). Observers recorded whether the response entered by the test administrator matched the student's response. In 21 of 115 (18%) observations of computer-delivered testlets, the test administrator entered responses on the student's behalf. In 18 (86%) of those cases, observers indicated that the entered response matched the student's response, while the remaining 3 observers either responded that they could not tell if the entered response matched the student's response, or they left the item blank.

4.7.2. Formative Monitoring Techniques

Several techniques for formative monitoring purposes are available for the DLM System. First, because DLM assessments are delivered as a series of testlets, an assessment administration monitoring extract was available on demand in Kite Educator Portal. This extract allowed state and local staff to check each student's progress toward completion of all required testlets. For each student, the extract listed the number of testlets completed and expected for each subject. To support local capacity for monitoring, webinars were delivered before the testing window opened. These webinars targeted district and school personnel who monitor assessments and had not yet been involved in DLM assessments.

Formative monitoring also occurred through regular calls with DLM staff and state education agencies. Throughout most of the year, these calls were scheduled twice per month. Topics related to monitoring that appeared on agendas for partner calls included assessment window preparation, anticipated high-frequency questions from the field, and an opportunity for state education agency-driven discussion. Particular attention was paid to questions from the field concerning sources of confusion among test administrators that could compromise assessment results. During the spring assessment window, check-in calls were hosted on the weeks between the regularly scheduled partner calls. The purpose of the check-in calls is to keep the DLM Governance Board apprised of any issues or concerns that arise during the testing window, which allows them to provide timely information to districts. States are provided with a description of the issues as well as actions that are in place to remedy the situation. During these meetings, partner states are encouraged to share any concerns that have arisen during the week from the field and to provide feedback on implemented fixes, if any were necessary.

4.7.3. Monitoring Testlet Delivery

Prior to the opening of a testing window, Agile Technology Solutions staff initiated an automated enrollment process that works in conjunction with test administrator EE and linkage level selection to assign the first testlet. Students who had missing or incorrect information in Kite Educator Portal were included in error logs that detail which information was missing (e.g., First Contact survey is not submitted) or incorrect (e.g., student is enrolled in a grade that is not tested). These error logs were accessed and evaluated by Agile Technology Solutions staff. When testlets could not be assigned for large numbers of students in a state due to missing or incorrect data, DLM staff worked with relevant state education agencies to either communicate general reminders to the field or solve problems regarding specific students.

Once the student completed the first testlet, adaptive delivery drove the remaining testlet assignments. During the spring assessment window, the DLM psychometric team monitored test delivery to ensure students received testlets according to auto-enrollment specifications. This included running basic frequency statistics to verify counts appeared as expected by grade, state, and testing model and verifying correct assignment to initial testlet-based rules that govern that process.

4.7.4. Data Forensics Monitoring

Two data forensics monitoring reports are available in Educator Portal. The first report includes information about testlets completed outside of normal business hours. The second report includes information about testlets that were completed within a short period of time.

The Testing Outside of Hours report allows state education agencies to specify days and hours within a day that testlets are expected to be completed. Each state can select its own days and hours for setting expectations. For example, a state could elect to flag any testlet completed outside of Monday through Friday from 6:00 a.m. to 5:00 p.m. local time. The Testing Outside of Hours report then identifies students who completed assessments outside of the defined expected hours. Overall, 10,366 (1%) ELA and mathematics testlets were completed outside of the expected hours by 8,315 (9%) students.

The Testing Completed in a Short Period of Time report identifies students who completed a testlet within an unexpectedly short period of time. The threshold for inclusion in the report was testlet completion time of less than 30 seconds in mathematics and 60 seconds in ELA. The report is intended for state users to identify potentially aberrant response patterns; however there are many legitimate reasons a testlet may be submitted in a short time period. Overall, 52,824 (4%) testlets were completed in a short period of time by 19,437 (22%) students.

4.8. Evidence From Test Administrators

This section first describes evidence collected from the spring 2022 test administrator survey. Data on user experience with the DLM System as well as student opportunity to learn is evaluated annually through a survey that test administrators are invited to complete after administration of the spring assessment. Test administrators receive one survey per rostered DLM student, which collects information about that student's assessment experience. As in previous years, the survey was distributed to test administrators in Kite Student Portal, where students completed assessments. The survey consisted of four blocks. Blocks 1 and 4 were administered in every survey. Block 1 included questions about the test administrator's perceptions of the assessments and the student's interaction with the content, and Block 4 included questions about the test administrator's background. Block 2 was spiraled, so test administrators received one randomly assigned section. In these sections, test administrators were asked about one of the following topics per survey: relationship to ELA instruction, relationship to mathematics instruction, or relationship to science instruction. Block 3 was added in 2021 and remained in the survey in 2022 to gather information about educational experiences during the COVID-19 pandemic. After evidence from the spring 2022 test administrator survey is presented, this section also presents evidence collected from First Contact survey responses and educator cognitive labs.

4.8.1. User Experience With the DLM System

A total of 16,836 test administrators responded to the survey (69%) about 53,684 students' experiences. Test administrators are instructed to respond to the survey separately for each of their students. Participating test administrators responded to surveys for a median of two students. Test administrators reported having an average of 11 years of experience in ELA, 11 years in mathematics, and 10 years with students with significant cognitive disabilities.

The following sections summarize responses regarding both educator and student experience with the system.

4.8.1.1. Educator Experience

Test administrators were asked to reflect on their own experience with the assessments as well as their comfort level and knowledge administering them. Most of the questions required test administrators to respond on a 4-point scale: *strongly disagree, disagree, agree,* or *strongly agree*. Responses are summarized in Table 4.17.

Nearly all test administrators (96%) agreed or strongly agreed that they were confident administering DLM testlets. Most respondents (89%) agreed or strongly agreed that the required test administrator training prepared them for their responsibilities as test administrators. Most test administrators also responded that they had access to curriculum aligned with the content that was measured by the assessments (86%) and that they used the manuals and the Educator Resources page (89%).

Table 4.17

Test Administrator Responses Regarding Test Administration

Statement	SD		D		A		SA		A+SA	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
I was confident in my ability to deliver DLM testlets.	218	1.4	437	2.9	6,156	40.9	8,237	54.7	14,393	95.6
Required test administrator training prepared me for the responsibilities of a test administrator.	408	2.7	1,180	7.9	7,293	48.6	6,120	40.8	13,413	89.4
I have access to curriculum aligned with the content measured by DLM assessments.	497	3.3	1,574	10.5	7,503	50.0	5,423	36.2	12,926	86.2
I used manuals and/or the DLM Educator Resource Page materials.	382	2.5	1,215	8.1	8,082	53.8	5,336	35.5	13,418	89.3

Note. SD = strongly disagree; D = disagree; A = agree; SA = strongly agree; A+SA = agree and strongly agree.

4.8.1.2. Student Experience

The spring 2022 test administrator survey included three items about how students responded to test items. Test administrators were asked to rate statements from *strongly disagree* to *strongly agree*. Results are presented in Table 4.18. The majority of test administrators agreed or strongly agreed that their students responded to items to the best of their knowledge, skills, and understandings; were able to respond regardless of disability, behavior, or health concerns; and had access to all necessary supports to participate.

Table 4.18

Test Administrator Perceptions of Student Experience with Testlets

	SD		D		A		SA		A+SA	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Student responded to items to the best of his/her knowledge, skills, and understanding	1,808	3.7	3,503	7.2	25,442	52.1	18,075	37.0	43,517	89.1
Student was able to respond regardless of his/her disability, behavior, or health concerns	2,806	5.7	4,123	8.4	24,656	50.4	17,360	35.5	42,016	85.9
Student had access to all necessary supports to participate	1,568	3.2	2,328	4.8	25,282	51.9	19,562	40.1	44,844	92.0

Note. SD = strongly disagree; D = disagree; A = agree; SA = strongly agree; A+SA = agree and strongly agree.

Annual survey results show that a small percentage of test administrators disagree that their student was able to respond regardless of disability, behavior, or health concerns; had access to all necessary supports; and was able to effectively use supports. In spring 2020, DLM staff conducted educator focus groups with educators who disagreed with one or more of these survey items to learn about potential accessibility gaps in the DLM System (Kobrin et al., 2022). A total of 18 educators from 11 states participated in six focus groups. The findings revealed that many of the challenges educators described were documented in existing materials (e.g., wanting clarification about allowable practices that are described in the *Test Administration Manual*, such as substituting materials; desired use of not-allowed practices like hand-over-hand that are used during instruction). DLM staff are using the focus group findings to review existing materials and develop new resources that better communicate information about allowable practices to educators.

4.8.2. Opportunity to Learn

Table 4.19 reports the opportunity to learn results. Approximately 71% of responses (*n* = 34,877) reported that most or all ELA testlets matched instruction, compared to 62% (*n* = 30,386) for mathematics. More specific measures of instructional alignment are planned to better understand the extent that content measured by DLM assessments matches students' academic instruction.

Table 4.19

Educator Ratings of Portion of Testlets That Matched Instruction

	None		Some (< half)		Most (> half)		All		Not applicable	
Subject	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
English language arts	2,869	5.8	10,636	21.7	19,619	39.9	15,258	31.1	733	1.5
Mathematics	3,472	7.1	14,010	28.7	18,210	37.4	12,176	25.0	865	1.8

A subset of test administrators was asked to indicate the approximate number of hours spent instructing students on each of the conceptual areas by subject (i.e., ELA, mathematics). Test administrators responded using a 6-point scale: 0 hours, 0–5 hours, 6–10 hours, 11–15 hours, 16–20 hours, or more than 20 hours. Table 4.20 and Table 4.21 indicate the amount of instructional time spent on conceptual areas for ELA and mathematics, respectively. Using 11 or more hours per conceptual area as a criterion for instruction, 51% of the test administrators provided this amount of instruction to their students in ELA, and 42% did so in mathematics.

Table 4.20

Instructional Time Spent on ELA Conceptual Areas

		Number of hours											
		0		0–5		6–10		11–15		16–20		>20	
Conceptual area	Median	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Determine critical elements of text	6–10	3,042	15.6	3,859	19.8	2,861	14.7	2,350	12.1	2,862	14.7	4,513	23.2
Construct understandings of text	11–15	2,151	11.1	3,505	18.1	2,829	14.6	2,453	12.7	3,032	15.7	5,399	27.9
Integrate ideas and information from text	11–15	2,637	13.7	3,643	18.9	2,916	15.1	2,568	13.3	2,965	15.4	4,567	23.7
Use writing to communicate	6–10	2,994	15.5	4,001	20.7	2,942	15.2	2,345	12.1	2,762	14.3	4,300	22.2
Integrate ideas and information in writing	6–10	3,840	19.9	3,942	20.5	2,857	14.8	2,346	12.2	2,686	13.9	3,599	18.7
Use language to communicate with others	16–20	1,212	6.3	2,394	12.4	2,398	12.4	2,230	11.5	3,177	16.4	7,949	41.1
Clarify and contribute in discussion	11–15	2,601	13.4	3,386	17.5	2,766	14.3	2,466	12.8	3,091	16.0	5,030	26.0
Use sources and information	6–10	4,472	23.1	4,106	21.2	2,951	15.3	2,285	11.8	2,447	12.7	3,066	15.9
Collaborate and present ideas	6–10	4,268	22.0	4,164	21.5	2,952	15.3	2,304	11.9	2,533	13.1	3,136	16.2

Table 4.21

Instructional Time Spent on Mathematics Conceptual Areas

		Number of hours											
		0		0–5		6–10		11–15		16–20		>20	
Conceptual area	Median	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Understand number structures (counting, place value, fraction)	16–20	1,414	6.5	3,560	16.3	3,081	14.1	2,637	12.1	3,639	16.7	7,452	34.2
Compare, compose, and decompose numbers and steps	6–10	3,206	14.8	4,281	19.8	3,406	15.8	2,910	13.5	3,332	15.4	4,458	20.6
Calculate accurately and efficiently using simple arithmetic operations	11–15	3,245	15.0	3,391	15.7	2,888	13.4	2,441	11.3	3,350	15.5	6,308	29.2
Understand and use geometric properties of two- and three-dimensional shapes	6–10	4,156	19.2	5,606	25.9	3,911	18.1	3,021	14.0	2,632	12.2	2,316	10.7
Solve problems involving area, perimeter, and volume	1–5	8,864	41.0	4,648	21.5	2,909	13.5	2,033	9.4	1,707	7.9	1,457	6.7
Understand and use measurement principles and units of measure	1–5	5,610	26.0	5,608	26.0	3,703	17.2	2,567	11.9	2,221	10.3	1,870	8.7
Represent and interpret data displays	6–10	5,509	25.6	5,105	23.7	3,775	17.5	2,695	12.5	2,386	11.1	2,088	9.7
Use operations and models to solve problems	6–10	4,689	21.7	4,099	19.0	3,262	15.1	2,782	12.9	2,989	13.9	3,738	17.3
Understand patterns and functional thinking	6–10	3,269	15.1	5,058	23.4	3,848	17.8	3,179	14.7	3,069	14.2	3,212	14.8

Results from the test administrator survey were also correlated with total linkage levels mastered by conceptual area, as reported on individual student score reports.[29] While a direct relationship between amount of instructional time and number of linkage levels mastered in the area is not expected, as some students may spend a large amount of time on an area and demonstrate mastery at the lowest linkage level for each EE, we generally expect that students who mastered more linkage levels in the area would also have spent more instructional time in the area. More evidence is needed to evaluate this assumption.

Table 4.22 summarizes the Spearman rank-order correlations between ELA conceptual area instructional time and linkage levels mastered in the conceptual area as well as between mathematics conceptual area instructional time and linkage levels mastered in the conceptual area. Correlations ranged from 0.14 to 0.36, with the strongest correlations observed for writing conceptual areas (ELA.C2.1 and ELA.C2.2) in ELA, and measurement, data, and analytic procedures conceptual areas (M.C4.1 and M.C1.3) in mathematics.

Table 4.22

Correlation Between Instructional Time and Linkage Levels Mastered by Conceptual Area

Conceptual area	Correlation with instruction time
English language arts	
ELA.C1.1: Determine critical elements of text	.22
ELA.C1.2: Construct understandings of text	.30
ELA.C1.3: Integrate ideas and information from text	.29
ELA.C2.1: Use writing to communicate	.36
ELA.C2.2: Integrate ideas and information in writing	.35
Mathematics	
M.C1.1: Understand number structures (counting, place value, fraction)	.14
M.C1.2: Compare, compose, and decompose numbers and steps	.29
M.C1.3: Calculate accurately and efficiently using simple arithmetic operations	.31
M.C2.1: Understand and use geometric properties of two- and three-dimensional shapes	.17
M.C2.2: Solve problems involving area, perimeter, and volume	.27
M.C3.1: Understand and use measurement principles and units of measure	.24
M.C3.2: Represent and interpret data displays	.26
M.C4.1: Use operations and models to solve problems	.32
M.C4.2: Understand patterns and functional thinking	.21

[29] See Chapter 7 of this manual for a description of results and reporting.

Another dimension of opportunity to learn is student engagement with instruction. The First Contact survey (see section 4.2.1.1 of this chapter) contains two questions about student engagement during computer- and educator-directed instruction. Table 4.23 shows the percentage of students who demonstrated different levels of attention by instruction type. Overall, 87% of students demonstrated fleeting or sustained attention to computer-directed instruction and 86% of students demonstrated fleeting or sustained attention to educator-directed instruction.

Table 4.23

Student Attention Levels During Instruction

	Demonstrates little or no attention		Demonstrates fleeting attention		Generally sustains attention	
Type of instruction	*n*	%	*n*	%	*n*	%
Computer-directed (*n* = 83,824)	10,848	12.9	46,675	55.7	26,301	31.4
Educator-directed (*n* = 90,552)	12,651	14.0	56,079	61.9	21,822	24.1

4.8.3. Educator Ratings on First Contact Survey

Before administering testlets, educators complete the First Contact survey, which is a survey of learner characteristics (see section 4.2.1.1 of this chapter for more details). Because ratings on the First Contact survey are distinct from the DLM assessment (which uses only a subset of items to calculate the student complexity band for each subject), they can serve as one source of external evidence regarding the construct being measured. The First Contact survey includes academic skill items: nine in the reading section and 13 in the mathematics section.

For each academic item on the First Contact survey, test development teams reviewed the learning maps to identify tested nodes that measured the same skill. Not all First Contact items directly corresponded to nodes in the map. Tested nodes were identified for two of the reading items and nine of the mathematics items. A summary of the First Contact academic items and the number of nodes identified in the learning maps is provided in Table 4.24.

Table 4.24

First Contact Items With Nodes Identified in the Learning Maps

First Contact item	Number of assessed nodes	Number of linkage levels measuring the nodes
Reading		
Recognizes single symbols presented visually or tactually	1	1
Identifies individual words without symbol support	1	7
Mathematics		
Creates or matches patterns of objects or images	2	3
Identifies simple shapes in 2 or 3 dimensions	8	4
Sorts objects by common properties (e.g., color, size, shape)	1	11
Adds or subtracts by joining or separating groups of objects	2	6
Adds and/or subtracts using numerals	13	7
Forms groups of objects for multiplication or division	2	6
Multiplies and/or divides using numerals	15	4
Tells time using an analog or digital clock	4	2
Uses common measuring tools (e.g., ruler or measuring cup)	2	2

For each tested node identified by the test development teams, all EEs and linkage levels measuring the node were identified. A dataset was created that included student mastery of the EE and linkage level measuring the node, as well as First Contact survey responses.[30] The First Contact items asked educators to use a 4-point scale to indicate how consistently students demonstrated each skill: *almost never* (0%–20% of the time), *occasionally* (21%–50% of the time), *frequently* (51%–80% of the time), or *consistently* (81%–100% of the time).

Polychoric correlations for reading and mathematics were calculated to determine the relationship between

[30] See Chapter 7 of this manual for a description of linkage level mastery and scoring rules.

the educator's First Contact rating and the student's reported mastery of the linkage level measuring nodes associated with the First Contact items.

Moderate but positive correlations are expected between First Contact ratings and student mastery of the linkage level for several reasons. The First Contact items were not originally designed to align with assessment items or linkage level statements. Also, educators are required to complete the First Contact survey before testlet administration; some educators complete it at the beginning of the school year. Educators may choose to update survey responses during the year but do not have to. Therefore, First Contact ratings may reflect student knowledge or understandings before instruction, while linkage level mastery represents end-of-year performance. However, in general, higher First Contact ratings are expected to be associated with student mastery of the linkage level measuring the same skill.

Correlations for First Contact items with linkage level mastery are summarized in Table 4.25.

Table 4.25

Correlations of First Contact Item Responses to Linkage Level Mastery

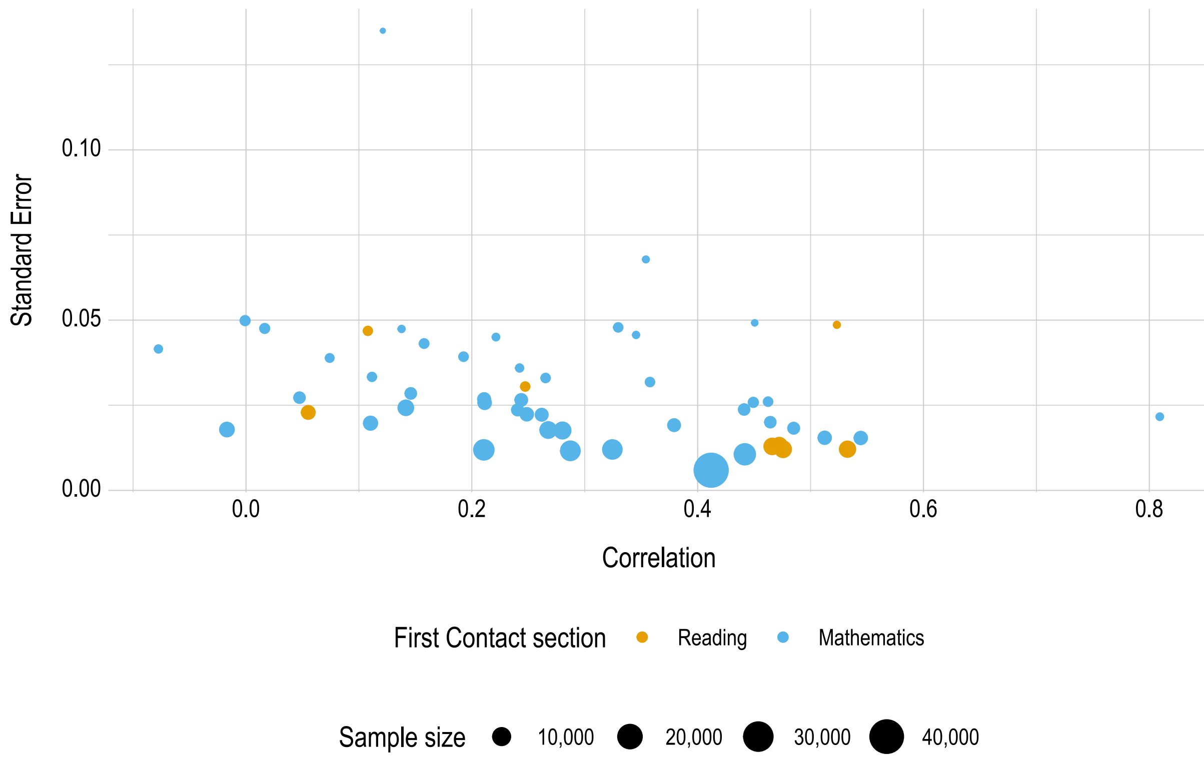
		Correlation			Standard Error		
First Contact section	Linkage levels (*n*)	Min	Max	Median	Min	Max	Median
Reading	8	.06	.53	.47	0.01	0.05	0.02
Mathematics	45	-.08	.81	.26	0.01	0.13	0.03

Mathematics First Contact items varied most in their relationship to linkage level mastery. Because mathematics nodes represent finer-grained skills, and test development teams identified more nodes in mathematics, more correlations were calculated (n = 45) than for reading (n = 8). Mathematics results were also likely affected by sample size. As few as 297 student data points were available for some linkage levels, compared to at least 915 in reading. The decreased sample size is likely attributable to fewer students testing at the Target and Successor linkage levels (see section 4.6.4 of this chapter). Furthermore, a negative relationship between mathematics First Contact rating and linkage level mastery was observed in three instances. An example is seen in the relationship between the Successor level of the grade 3 EE M.EE.3.OA.9 and the First Contact item "Creates or matches patterns of objects or images." The linkage level statement for this EE and level is "Determine the pattern rule; extend a pattern by applying the pattern rule." Although the linkage level measures the node "Extend a symbolic pattern by applying a rule," it also measures other nodes that are not aligned to any First Contact item; this combination likely contributed to the negative relationship observed. However, small sample size is associated with increased standard errors (Moinester & Gottfried, 2014), and therefore these negative correlations should be interpreted with caution.

Overall, 94% (n = 50) of the correlations were positive, indicating generally positive associations between linkage level mastery and First Contact ratings. Results for all correlations are summarized in Figure 4.5.

Figure 4.5

Relationship of First Contact Responses to Linkage Level Mastery



4.8.4. Educator Cognitive Labs

Educator cognitive labs have been recommended as a potential source of response process evidence for alternate assessments based on alternate achievement standards, in which educator ratings are the items (Goldstein & Behuniak, 2010). This approach was used for DLM educator-administered testlets because educators interpret student behavior and respond to items about the student's response. Most of these testlets involve test administrator interpretation of the responses of students who are working on consistent, intentional communication and who are working on foundational skills that promote their access to grade-level content. Writing testlets are also educator-administered at all linkage levels.

Cognitive labs were conducted in spring 2015 with 15 educators in five schools across two states. Educators completed think-aloud procedures while preparing for and administering educator-administered testlets in reading, writing, and math. They were first presented with the TIP, which is a short document that provides background information needed to prepare to administer the testlet (see section 4.3.1.2.1 of this chapter).

Educators were asked to think out loud as they read through the TIP. Next, the educator gathered the materials needed for the assessment and administered the testlet. Probes were sometimes used during the process to ask about educator interpretation of the on-screen instructions and the rationale behind decisions they made during administration. When the testlet was finished, educators also completed post-hoc interviews about the contents of test-administration instructions, use of materials, clarity of procedures, and interpretation of student behaviors. All labs were video recorded and an observer took notes during the administration. The initial phase of analysis involved recording evidence of intended administration and sources of challenge to intended administration at each of the following stages: (1) preparation for administration, (2) interpretation of educator directions within the testlet, (3) testlet administration, (4) interpretation of student behaviors, and (5) recording student responses. Through this lens, we were able to look for evidence related to fidelity (1, 2, 3, and 5) as well as response process (4). These 15 labs were the first phase of data collection using this protocol. Preliminary evidence on interpretation of student behaviors indicates that the ease of determining student intent depended in part on the student's response mode.

- Educators were easily able to understand student intent when the student indicated a response by picking up objects and handing them to the educator.
- In a case where the student touched the object rather than handing it to the educator, the educator accepted that response and entered it, but speculated as to whether the student was just choosing the closest object.
- When a student briefly touched one object and then another, the educator entered the response associated with the second object but commented that she was not certain if the student intended that choice.
- When a student used eye gaze, the educator held objects within the student's field of vision and put the correct response away from the current gaze point so that a correct response required intentional eye movement to the correct object.
- When a student's gesture did not exactly match one of the response options, the educator was able to verbalize the process of deciding how to select the option that most closely matched the student's behavior. Her process was consistent with the expectations in the *Test Administration Manual*.
- In one case, the educator moved objects to prepare for the next item, which took her attention away from the student and caused her to miss his eye gaze that indicated a response. She recorded *no response*. However, this was observed for a student whose communication and academic skills were far beyond what was being assessed. The testlet was not appropriate for this student and his typical response mode for DLM testlets was verbal.

4.9. Conclusion

Delivery of the DLM System was designed to align with instructional practice and be responsive to individual student needs. Assessment delivery options allow for necessary flexibility to reflect student needs while also including constraints to maximize comparability and support valid interpretation of results. The dynamic nature of DLM assessment administration is reflected in the initial input through the First Contact survey, as well as adaptive routing between testlets. Evidence collected from the DLM System, test administration monitoring, and test administrators indicates that students are able to successfully interact with the system to demonstrate their knowledge, skills, and understandings.

5. Modeling

To provide feedback about student performance, the Dynamic Learning Maps® (DLM®) Alternate Assessment System draws on a well-established research base in cognition and learning theory and relatively uncommon operational psychometric methods. The approach uses innovative, operational psychometric methods to provide feedback about student mastery of skills. This chapter describes the psychometric model that underlies the DLM System and describes the process used to estimate item and student parameters from student assessment data.

5.1. Psychometric Background

Learning maps, which are networks of sequenced learning targets, are at the core of the DLM assessments in English language arts and mathematics. In general, a learning map is a collection of skills to be mastered that are linked together by connections between the skills. The connections between skills indicate what should be mastered prior to learning additional skills. Together, the skills and their prerequisite connections map out the progression of learning within a given subject. Stated in the vocabulary of traditional psychometric methods, a learning map defines a large set of discrete latent variables indicating students' learning status on key skills and concepts relevant to a large content domain, as well as a series of pathways indicating which topics (represented by latent variables) are prerequisites for learning other topics.

Because of the underlying map structure and the goal of providing more fine-grained information beyond a single raw or scale score value, student results are reported as a profile of skill mastery. This profile is created using diagnostic classification modeling, which draws on research in cognition and learning theory to provide information about student mastery of multiple skills measured by the assessment. Diagnostic classification models (DCMs) are confirmatory latent class models that characterize the relationship of observed responses to a set of categorical latent variables (e.g., Bradshaw, 2016; Rupp et al., 2010). DCMs are also known as cognitive diagnosis models (e.g., Leighton & Gierl, 2007) or multiple classification latent class models (Maris, 1999) and are mathematically equivalent to Bayesian networks (e.g., Almond et al., 2015; Mislevy & Gitomer, 1995; Pearl, 1988). This is the main difference from more traditional psychometric models, such as item response theory, which model a single, continuous latent variable. DCMs provide information about student mastery on multiple latent variables or skills of interest.

DCMs have primarily been used in educational measurement settings in which more detailed information about test-takers' skills is of interest, such as in assessing individual mathematics skills (e.g., Bradshaw et al., 2014), different levels of reading complexity (e.g., Templin & Bradshaw, 2014), and the temporal acquisition of science skills (e.g., Templin & Henson, 2008). To provide detailed profiles of student mastery of the skills, or attributes, measured by the assessment, DCMs require the specification of an item-by-attribute Q-matrix, indicating the attributes measured by each item. In general, for a given item, i, the Q-matrix vector would be represented as $q_i = [q_{i1}, q_{i2}, ..., q_{i\mathrm{A}}]$, where A is the total number of attributes. Similar to a factor pattern matrix in a confirmatory factor model, Q-matrix indicators are binary: either the item measures an attribute ($q_{ia} = 1$) or it does not ($q_{ia} = 0$).

For each item, there is a set of conditional item-response probabilities that corresponds to the student's possible mastery patterns. Although DCMs can be defined using any number of latent categories for each attribute, it is most common to use binary attributes, which provide more interpretable results to stakeholders (Bradshaw & Levy, 2019). When an item measures a single binary attribute, only two statuses are possible for any examinee: a master of the attribute or a nonmaster of the attribute.

In general, the modeling approach involves specifying the Q-matrix, determining the probability of being classified into each category of mastery (master or nonmaster), and relating those probabilities to students' response data to determine a posterior probability of being classified as a master or nonmaster for each attribute. For DLM assessments, the attributes for which probabilities of mastery are calculated are the Essential Element (EE) linkage levels.

5.2. Essential Elements and Linkage Levels

Because the primary goal of the DLM assessments is to measure what students with the most significant cognitive disabilities know and can do, alternate grade-level expectations called EEs were created to provide students in the population access to the general education grade-level academic content. See Chapter 2 of this manual for a complete description. Each EE has an associated set of linkage levels that are ordered by increasing complexity. There are five linkage levels for each EE in English language arts and mathematics: Initial Precursor, Distal Precursor, Proximal Precursor, Target, and Successor.

5.3. Overview of the DLM Modeling Approach

Many statistical models are available for estimating the probability of mastery for attributes in a DCM. The statistical model used to determine the probability of mastery for each linkage level for DLM assessments is the log-linear cognitive diagnosis model (LCDM). The LCDM is a DCM model that provides a general statistical framework for obtaining probabilities of class membership for each measured attribute (Henson et al., 2009). Student mastery statuses for each linkage level are obtained from a Bayesian estimation procedure, which contributes to an overall profile of mastery.

5.3.1. Model Specification

Each linkage level was calibrated separately for each EE using separate LCDMs. Each linkage level within an EE is estimated separately because of the administration design in which overlapping data from students taking testlets at multiple levels within an EE is uncommon. Also, because items were developed to meet a precise cognitive specification, all master and nonmaster probability parameters for items measuring a linkage level were assumed to be equal. That is, all items were assumed to be fungible, or exchangeable, within a linkage level. As such, each class (i.e., masters or nonmasters) has a single probability of responding correctly to all items measuring the linkage level, as depicted in Table 5.1. Therefore, for each item measuring the same linkage level, the probability of providing a correct response is held constant for all students in each mastery class. Chapter 3 of this manual details item review procedures intended to support the fungibility assumption, and section 5.4.1 of this chapter describes empirical evidence to support this constraint.

Table 5.1

Depiction of Fungible Item Parameters for Items Measuring a Single Linkage Level

Item	Class 1 (Nonmasters)	Class 2 (Masters)
1	π_1	π_2
2	π_1	π_2
3	π_1	π_2
4	π_1	π_2
5	π_1	π_2

Note. π represents the probability of providing a correct response.

The DLM scoring model for the 2021–2022 administration was as follows. Each linkage level within each EE was considered the latent variable to be measured (the attribute). Using DCMs, a probability of mastery on a scale of 0 to 1 was calculated for each linkage level within each EE. Students were then classified into one of two classes for each linkage level of each EE: either master or nonmaster. As described in Chapter 6 and Chapter 7 of this manual, a posterior probability of at least .8 was required for mastery classification.

The general form of DCMs is shown in Equation 5.1. In Equation 5.1, π_{ic} is the conditional probability of a student in class c providing a correct response to item i, and x_{ij} is the observed response (i.e., 0 or 1) of student j to item i. Thus, $\pi_{ic}^{x_{ij}}(1-\pi_{ic})^{1-x_{ij}}$ represents the probability of a respondent in class c providing the observed response to item i. Finally, ν_c represents the base rate probability that any given respondent belongs to class c.

$$P(X_j = x_j) = \sum_{c=1}^{C} \nu_c \prod_{i=1}^{I} \pi_{ic}^{x_{ij}}(1-\pi_{ic})^{1-x_{ij}} \tag{5.1}$$

Different types of DCMs use different measurement models to define π_{ic} in Equation 5.1. For DLM assessments, item responses are modeled using the LCDM, as described by Henson et al. (2009). The LCDM defines the conditional probabilities using a generalized linear model with a logit link function. Specifically, using the LCDM, π_{ic} is defined as seen in Equation 5.2, where α_c is a binary indicator of mastery status for a student in class c for that attribute.

$$\pi_{ic} = P(X_{ic} = 1|\alpha_c) = \frac{\exp(\lambda_{i,0} + \lambda_{i,1}\alpha_c)}{1+\exp(\lambda_{i,0} + \lambda_{i,1}\alpha_c)} \tag{5.2}$$

Equation 5.2 utilizes the LCDM notation described by Rupp et al. (2010), where the λ subscripts follow structure of “item,effect”, where the effect is defined as 0 = intercept, 1 = main effect. All items in a linkage level were assumed to measure that linkage level, meaning the Q-matrix for the linkage level was a column of ones. As such, each item measured one latent variable, resulting in two parameters per item: (a) an intercept ($\lambda_{i,0}$) that corresponds to the probability of answering the item correctly for examinees who have not mastered the linkage level and (b) a main effect ($\lambda_{i,1}$) that corresponds to the increase in the probability of answering the item correctly for examinees who have mastered the linkage level. Because students who have mastered the linkage level should also have a higher probability of providing a correct response than students who have not, $\lambda_{i,1}$ is constrained to be positive to ensure monotonicity (Henson et al., 2009). As per the assumption of item fungibility, a single set of probabilities was estimated for all items within a linkage level. Therefore, Equation 5.2 can be simplified to remove the item-level λ parameters. Equation 5.3 removes the item-level effects, showing that for each linkage level we now estimate only one intercept shared by all items (λ_0) and one main effect shared by all items (λ_1).

$$\pi_{ic} = P(X_{ic} = 1|\alpha_c) = \frac{\exp(\lambda_0 + \lambda_1\alpha_c)}{1 + \exp(\lambda_0 + \lambda_1\alpha_c)} \quad (5.3)$$

Finally, because each linkage level is estimated separately as a single attribute LCDM, there are only two possible mastery classes (i.e., nonmasters and masters). Therefore, only a single structural parameter was needed (ν), which is the probability that a randomly selected student who is assessed on the linkage level is a master (i.e., the analogous map parameter). The base rate of the other class (i.e., nonmastery) is deterministically calculated as $1 - \nu$. In total, three parameters per linkage level are specified in the DLM scoring model: a fungible intercept, a fungible main effect, and the proportion of masters.

5.3.2. Model Calibration

A Bayesian approach was used to calibrate the DCMs. A Bayesian approach was preferred over a simpler maximum likelihood approach because the posterior distributions derived from Bayesian methods offer more robust methods for evaluating model fit (see section 5.4.1). We specifically selected an empirical Bayes procedure for several reasons. In any Bayesian approach, prior distributions must be specified for each parameter in the model. An Empirical Bayes procedure uses the data to estimate a prior distribution, whereas a standard Bayesian procedure would fix the prior distribution *a priori* (Carlin & Louis, 2001). The empirical priors offer several advantages. First, due to the number of models that are estimated (i.e., 815 linkage levels), an *a priori* specification of prior distributions would require the same priors for each model. By using empirical priors, we can select prior distributions specific to each linkage level, rather than using a single general prior that may be more or less appropriate for any given linkage level, thus increasing the information available in the estimation process (Nabi et al., 2022). Second, if *a priori* priors are used, there are many decisions that a practitioner must make when eliciting the fixed priors. Different decisions lead to different prior distributions, which would then affect the resulting posterior distributions (Falconer et al., 2022; Stefan et al., 2020). Using empirical priors removes the practitioner degrees of freedom that could result in different priors resulting from practitioner decisions. Finally, empirical prior distributions are often more informative than a general prior fixed *a priori*. More informative priors make the estimation more efficient by focusing the sampling more closely on the highest density area of the posterior distribution without biasing the final parameter estimates (Petrone et al., 2014).

Across all grades and subjects, there were 163 EEs, each with five linkage levels, resulting in a total of 163 $\times$ 5 $=$ 815 separate calibration models. Each separate calibration included all operational items for the EE and linkage level. Each model was estimated using a two-step Empirical Bayes procedure (Casella, 1985; Efron, 2014) using the software package rstan (Stan Development Team, 2022). The rstan package is an interface to the *Stan* probabilistic programming language (Carpenter et al., 2017). The first step of the process used for calibrating the DLM model consists of fitting a number of bootstrapped models with an optimization algorithm to estimate the standard error of each parameter. The second step then used the standard errors from step 1 as prior distributions in a fully Bayesian estimation of the model using a Markov Chain Monte Carlo procedure. Each step is described in detail in the following sections.

5.3.2.1. Step 1: Estimation of Bootstrapped Models

In the first step, the data for each attribute were bootstrap resampled 100 times (Babu, 2011). For each bootstrap resample, the LCDM was fit using the low-memory Broyden-Fletcher-Goldfarb-Shanno optimization algorithm (Liu & Nocedal, 1989; Nocedal & Wright, 2006). The low-memory Broyden-Fletcher-Goldfarb-Shanno algorithm is a widely used maximum likelihood optimization algorithm that can efficiently estimate many types of models, including the LCDM. After estimating the LCDM on each of the bootstrapped samples, there are 100 estimates of each of the three model parameters. We denote these parameters as:

$$\lambda_0^* = [\lambda_{0_1}, \lambda_{0_2}, \lambda_{0_3}, ..., \lambda_{0_{100}}]$$
$$\lambda_1^* = [\lambda_{1_1}, \lambda_{1_2}, \lambda_{1_3}, ..., \lambda_{1_{100}}]$$
$$\nu^* = [\nu_1, \nu_2, \nu_3, ..., \nu_{100}]$$

For each parameter, we calculated the mean value and standard deviation across the 100 bootstrap samples. These values were then used to define the prior distributions in the second step.

5.3.2.2. Step 2: Estimation of Final Bayesian Model

In the second step, the full data set was used to estimate the LCDM for each linkage level using Markov Chain Monte Carlo and the Hamiltonian Monte Carlo algorithm (Betancourt, 2018; Neal, 2011). The prior distribution for each parameter is defined using the values from the first step. The prior for the intercept (λ_0) is defined as a normal distribution with a mean and standard deviation equal to the corresponding values from the first step. We define the mean and standard deviation of λ_0^* as $\mu_{\lambda_0^*}$ and $\sigma_{\lambda_0^*}$, respectively. The prior for the intercept in the LCDM is then given as:

$$\lambda_0 \sim \mathcal{N}(\mu_{\lambda_0^*},\ \sigma_{\lambda_0^*}) \tag{5.4}$$

Similarly, the main effect parameter (λ_1) is also defined with a normal distribution. However, the prior distribution of the main effect is truncated at 0, forcing the main effect to be positive to ensure monotonicity in the LCDM.

$$\lambda_1 \sim \begin{cases} 0, & \text{if } \lambda_1 \leq 0 \\ \mathcal{N}(\mu_{\lambda_1^*},\ \sigma_{\lambda_1^*}), & \text{otherwise} \end{cases} \tag{5.5}$$

Finally, because the base rate of linkage level class membership (ν) is a probability that must be between 0 and 1, a beta distribution is used for the prior. The beta distribution is governed by two shape parameters, α and β. Given these two shape parameters, the mean of the beta distribution is given by:

$$\mu = \frac{\alpha}{\alpha + \beta} \tag{5.6}$$

and the variance as:

$$\sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{5.7}$$

With some algebra, we calculate the values of the shape parameters given a mean (μ_{ν^*}) and standard deviation (σ_{ν^*}):

$$\alpha_{\nu^*} = \left(\frac{1 - \mu_{\nu^*}}{\sigma^2_{\nu^*}} - \frac{1}{\mu_{\nu^*}}\right)\mu^2_{\nu^*} \tag{5.8}$$

$$\beta_{\nu^*} = \alpha_{\nu^*}\left(\frac{1}{\mu_{\nu^*}} - 1\right) \tag{5.9}$$

The prior distribution for ν is then defined as:

$$\nu \sim \mathcal{B}(\alpha_{\nu^*},\ \beta_{\nu^*}) \tag{5.10}$$

After the prior distributions have been defined, the LCDM was estimated. To ensure the posterior was adequately explored, four chains were estimated. For each chain, we specified 2,000 warm-up iterations and retained 1,000 post-warm-up iterations. This resulted in a posterior distribution of 4,000 draws (i.e., 1,000 from each of the four chains). After estimation, we ensured that the model had converged and adequately explored the posterior space by evaluating the $\widehat{R}$ and effective sample size metrics described by Vehtari et al. (2021). Using the cutoffs recommended by Vehtari et al. (2021), we ensured that all $\widehat{R}$ values were below 1.01 and that all effective sample sizes were greater than 400. After model evaluation, the mean of the posterior distribution for each of the three model parameters was taken. These parameter estimates were then used for scoring the DLM assessments.

5.3.3. Estimation of Student Mastery Probabilities

Once the LCDM parameters have been calibrated, student mastery probabilities are then obtained for each assessed linkage level. For DLM scoring, student mastery probabilities are *expected a posteriori*, or EAP, estimates. This is also the method most commonly used in scale score assessments (e.g., item response theory).[31] For each student j and linkage level l, EAP estimates of mastery probability, $\hat{\alpha}_{jl}$, are obtained using the following formula:

$$\hat{\alpha}_{jl} = \frac{\prod_{i=1}^{I_j}\left[\pi_{i1}^{X_{ji}}(1 - \pi_{i1})^{(1-X_{ji})}\right]\nu_1}{\sum_{c=0}^{1}\prod_{i=1}^{I_j}\left[\pi_{ic}^{X_{ji}}(1 - \pi_{ic})^{(1-X_{ji})}\right]\nu_c} \tag{5.11}$$

31 For a thorough discussion of the EAP estimates in scale score and diagnostic settings, see Chapter 10 of Rupp et al. (2010).

In Equation 5.11, $X_j i$ is the dichotomous response of student j to item i and π_{ic} is the model-based probability of answering item i correctly, conditional on student j having mastery status c for the linkage level, as defined in Equation 5.3. The mastery status can take two values: masters (c = 1) and nonmasters (c = 0). Finally, ν_c is the base rate probability of membership in each mastery status (see Equation 5.1). Thus, the numerator represents the likelihood of the student being in class $c = 1$ (i.e., the master class), and the denominator is the total likelihood across both classes. The EAP estimate is then the proportion of the total likelihood that comes from the master class.

5.4. Model Evaluation

There are many ways to evaluate DCMs. Ravand and Baghaei (2020) suggest four main areas for evaluation: (1) fit, (2) classification consistency and accuracy, (3) item discrimination, and (4) congruence of attribute difficulty with substantive expectations. Fit can be further broken down into different types of fit (e.g., model fit and item fit).

Many of these aspects are described in other sections of this manual and published research. Item fit is described with other measures of item quality in Chapter 3 of this manual, and classification consistency is discussed with other measures of consistency and reliability in Chapter 8 of this manual. The congruence of difficulty and expectations is discussed in Chapter 2 of this manual and the work of W. J. Thompson and Nash (2019) and W. J. Thompson and Nash (2022). Finally, item discrimination is described later in this chapter in section 5.5.3, in the context of estimated model parameters.

In this section, we focus on two aspects that are critical to inferences of student mastery: model fit and classification accuracy. Model fit has important implications for the validity of inferences that can be made from assessment results. If the model used to calibrate and score the assessment does not fit the data well, results from the assessment may not accurately reflect what students know and can do. Also called absolute model fit (e.g., Chen et al., 2013), this aspect involves an evaluation of the alignment between the three parameters estimated for each linkage level and the observed item responses. The second aspect is classification accuracy. This refers to how well the classifications represent the true underlying latent class. The accuracy of the assessment results (i.e., the classifications) is a prerequisite for any inferences that would be made from the results. Thus, the accuracy of the classifications is perhaps the most crucial aspect of model evaluation from a practical and operational standpoint. These aspects are discussed in the following sections.

5.4.1. Model Fit

Absolute model fit is evaluated through posterior predictive model checks, as described by W. J. Thompson (2019). Using parameter posterior distributions, we create a distribution for the expected number of students at each raw score point (i.e., the number of correct item responses). We then compare the observed number of students at each score point to the expected distribution using a χ^2-like statistic. Finally, we can compare our χ^2-like statistic to a distribution of what would be expected, given the expected distributions of students at each score point. This results in a posterior predictive *p*-value (*ppp*), which represents how extreme our observed statistic is compared to the model-implied expectation. Very low values indicate poor model fit, whereas very high values may indicate overfitting. For details on the calculation of this statistic, see W. J. Thompson (2019).

Due to the large number of models being evaluated (i.e., 815 linkage levels), the *ppp* values were adjusted using the Holm correction, which is uniformly more powerful than the popular Bonferroni method (Holm, 1979). Linkage levels were flagged for misfit if the adjusted *ppp* value was less than .05. Table 5.2 shows the percentage of models with acceptable model fit by linkage level. Across all linkage levels, 631 (77%) of the estimated models showed acceptable model fit. Misfit was not evenly distributed across the linkage levels. The lower linkage levels were flagged at a higher rate than the higher linkage levels. This is likely due to the greater diversity in the student population at the lower linkage levels (e.g., required supports, expressive communication behaviors, etc.), which may affect item response behavior. To address the model misfit flags, we are prioritizing test development for linkage levels flagged for misfit so that testlets contributing to misfit can be retired.[32] We also plan to incorporate additional item quality statistics to the review of field test data to ensure that only items and testlets that conform to the model expectations are promoted to the operational assessment. Overall, however, the fungible LCDM models appear to largely reflect the observed data. Additionally, model fit is evaluated on an annual basis and continues to improve over time as a result of adjustments to the pool of available content (i.e., improved item writing practices, retirement of testlets contributing to misfit). Finally, it should be noted that a linkage level flagged for model misfit may still have high classification accuracy, indicating that student mastery classifications can be trusted, even in the presence of misfit.

Table 5.2

*Percentage of Models With Acceptable Model Fit (*ppp *> .05)*

Linkage Level	English Language Arts (%)	Mathematics (%)
Initial Precursor	64.6	23.9
Distal Precursor	70.8	82.1
Proximal Precursor	84.4	74.6
Target	90.6	83.6
Successor	94.8	97.0

5.4.2. Classification Accuracy

The most practically important aspect of model fit for DCMs is classification accuracy. Classification accuracy is a measure of how accurate or uncertain classification decisions are for a given attribute in a DCM (the linkage level for DLM assessments). This measure of model fit is conceptualized by a summary of a 2×2 contingency table of the true and model-estimated mastery statuses (Sinharay & Johnson, 2019).[33]

For an operational assessment, we do not know students' true mastery status. However, we can still estimate the classification accuracy for each linkage level, as shown by Wang et al. (2015) and Johnson and Sinharay (2018) with

[32] For a description of item development practices, see Chapter 3 of this manual.

[33] For a discussion of the closely related classification consistency, see Chapter 8 of this manual.

$$\hat{\mathrm{P}}_{\mathrm{A}} = \frac{1}{\mathrm{N}}\sum_{n=1}^{\mathrm{N}}\tilde{\alpha}\mathrm{P}(\alpha = 1|\mathbf{X} = \mathbf{x}_n) + \frac{1}{\mathrm{N}}\sum_{n=1}^{\mathrm{N}}(1-\tilde{\alpha})\mathrm{P}(\alpha = 0|\mathbf{X} = \mathbf{x}_n). \quad (5.12)$$

In Equation 5.12, N is the total number of students, $\tilde{\alpha}$ is the model-estimated mastery status, and $\mathrm{P}(\alpha = 1|\mathbf{X} = \mathbf{x}_n)$ is the model-estimated probability that the linkage level was mastered (or not mastered for $\alpha = 0$). Johnson and Sinharay (2018) recommended interpretive guidelines for the classification accuracy, $\hat{\mathrm{P}}_{\mathrm{A}}$: ≥.99 = Excellent, .95–.98 = Very Good, .89–.94 = Good, .83–.88 = Fair, .55–.82 = Poor, and <.55 = Weak.

Across all estimated models, 615 linkage levels (75%) demonstrated at least fair classification accuracy. Table 5.3 shows the number and percentage of models within each linkage level that demonstrated each category of classification accuracy. Results are fairly consistent across linkage levels, with no one level showing systematically higher or lower accuracy. As was the case for model misfit, linkage levels flagged for low classification accuracy are prioritized for test development.

Table 5.3

Estimated Classification Accuracy by Linkage Level

	Weak (%)	Poor (%)	Fair (%)	Good (%)	Very Good (%)	Excellent (%)
Linkage Level	0.00–.55	.55–.82	.83–.88	.89–.94	.95–.98	.99–1.00
English Language Arts						
Initial Precursor	0 (0.0)	2 (2.1)	22 (22.9)	45 (46.9)	16 (16.7)	11 (11.5)
Distal Precursor	0 (0.0)	23 (24.0)	30 (31.2)	21 (21.9)	10 (10.4)	12 (12.5)
Proximal Precursor	0 (0.0)	41 (42.7)	34 (35.4)	11 (11.5)	6 (6.2)	4 (4.2)
Target	0 (0.0)	35 (36.5)	32 (33.3)	16 (16.7)	7 (7.3)	6 (6.2)
Successor	0 (0.0)	27 (28.1)	24 (25.0)	30 (31.2)	10 (10.4)	5 (5.2)
Mathematics						
Initial Precursor	0 (0.0)	1 (1.5)	8 (11.9)	50 (74.6)	8 (11.9)	0 (0.0)
Distal Precursor	0 (0.0)	22 (32.8)	29 (43.3)	13 (19.4)	2 (3.0)	1 (1.5)
Proximal Precursor	0 (0.0)	22 (32.8)	19 (28.4)	20 (29.9)	5 (7.5)	1 (1.5)
Target	0 (0.0)	10 (14.9)	26 (38.8)	22 (32.8)	8 (11.9)	1 (1.5)
Successor	0 (0.0)	17 (25.4)	22 (32.8)	18 (26.9)	10 (14.9)	0 (0.0)

When looking at absolute model fit and classification accuracy in combination, linkage levels flagged for absolute model misfit often have high classification accuracy. Of the 184 linkage levels that were flagged for absolute model misfit, 169 (92%) showed fair or better classification accuracy. Thus, even when misfit is present, we can be confident in the accuracy of the mastery classifications. In total, 98% of linkage levels (n = 800) had acceptable absolute model fit and/or acceptable classification accuracy.

5.5. Calibrated Parameters

As stated in the previous section, the item parameters for diagnostic assessments are the conditional probability of nonmasters providing a correct response (i.e., the inverse logit of λ_0) and the conditional probability of masters providing a correct response (i.e., the inverse logit of $\lambda_0 + \lambda_1$). Because of the assumption of fungibility, parameters are calculated for each of the 815 linkage levels across English language arts and mathematics. Across all linkage levels, the conditional probability that masters provide a correct response is generally expected to be high, while it is expected to be low for nonmasters. In addition to the item parameters, the psychometric model also includes a structural parameter, which defines the base rate of class membership for each linkage level. A summary of the operational parameters used to score the 2021–2022 assessment is provided in the following sections.

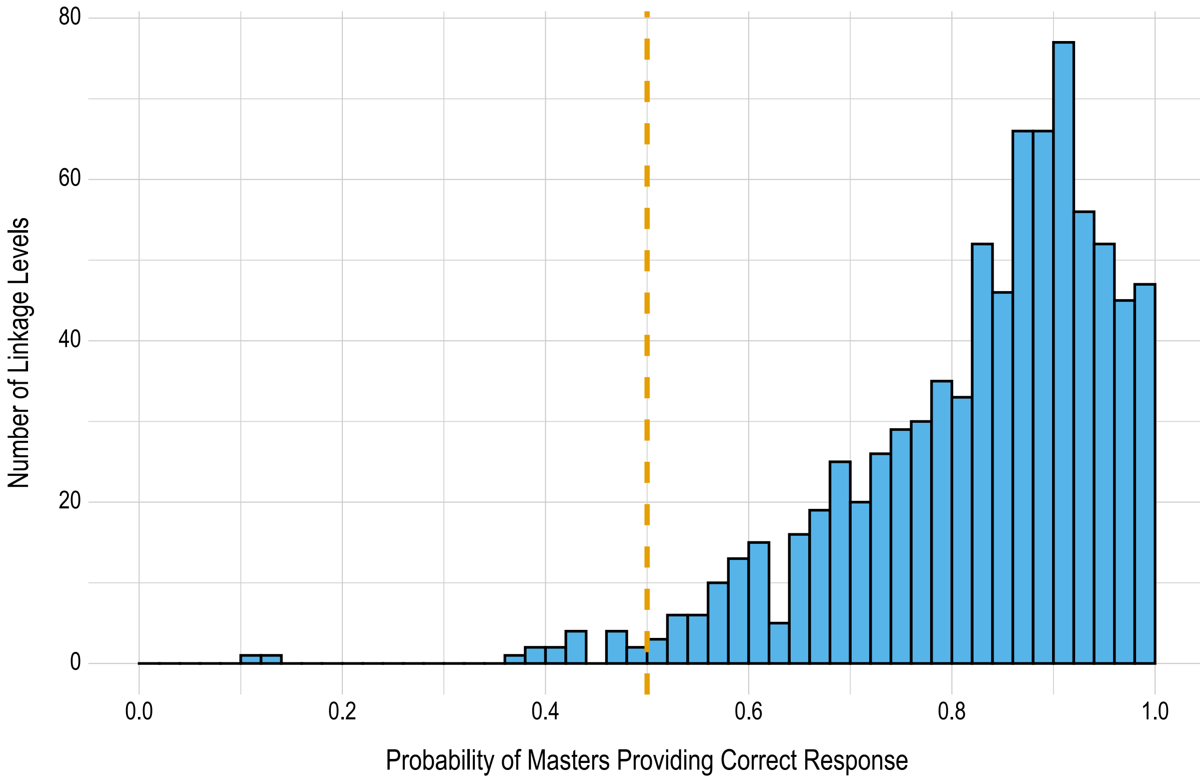
5.5.1. Probability of Masters Providing Correct Response

When items measuring each linkage level function as expected, students who have mastered the linkage level have a high probability of providing a correct response to items measuring the linkage level. Instances where masters have a low probability of providing correct responses may indicate that the linkage level does not measure what it is intended to measure, or that students who have mastered the content select a response other than the key. These instances may result in students who have mastered the content providing incorrect responses and being incorrectly classified as nonmasters. This outcome has implications for the validity of inferences that can be made from results, including educators using results to inform instructional planning in the subsequent year.

Using the 2021–2022 operational calibration, Figure 5.1 depicts the conditional probability of masters providing a correct response to items measuring each of the 815 linkage levels. Because the point of maximum uncertainty is .50 (i.e., equal likelihood of mastery or nonmastery), masters should have a greater than 50% chance of providing a correct response. The results in Figure 5.1 demonstrate that the vast majority of linkage levels (*n* = 798, 98%) performed as expected. Additionally, 93% of linkage levels (*n* = 760) had a conditional probability of masters providing a correct response over .60. Only a few linkage levels (*n* = 5, 1%) had a conditional probability of masters providing a correct response less than .40. Of these five linkage levels with a conditional probability of masters providing a correct response less than .40, the Successor linkage level was the most prevalent, with three linkage levels (60%). Thus, the vast majority of linkage levels performed consistently with expectations for masters of the linkage levels.

Figure 5.1

Probability of Masters Providing a Correct Response to Items Measuring Each Linkage Level



Note. Histogram bins are shown in increments of .02. Reference line indicates .50.

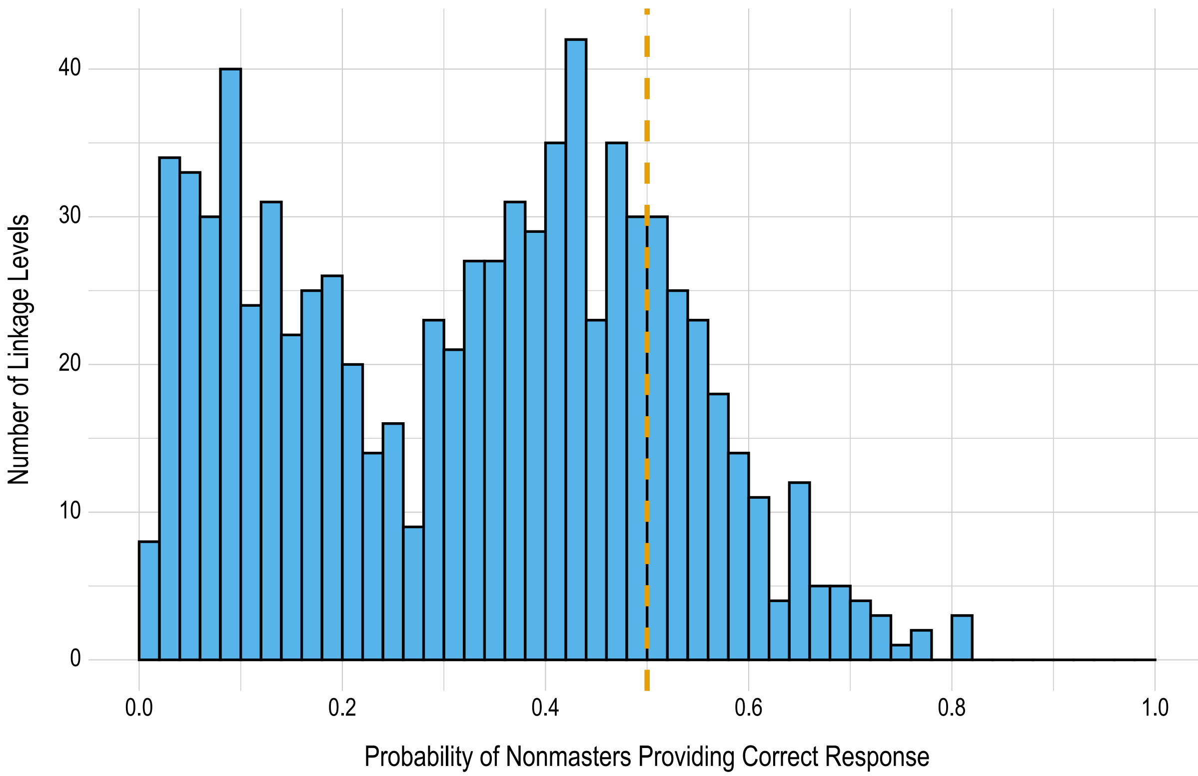
5.5.2. Probability of Nonmasters Providing Correct Response

When items measuring each linkage level function as expected, nonmasters of the linkage level have a low probability of providing a correct response to items measuring the linkage level. Instances where nonmasters have a high probability of providing correct responses may indicate that the linkage level does not measure what it is intended to measure, or that the correct answers to items measuring the level are easily guessed. These instances may result in students who have not mastered the content providing correct responses and being incorrectly classified as masters. This outcome has implications for the validity of inferences that can be made from results and for educators using results to inform instructional planning in the subsequent year.

Figure 5.2 summarizes the probability of nonmasters providing correct responses to items measuring each of the 815 linkage levels. There is greater variation in the probability of nonmasters providing a correct response to items measuring each linkage level than was observed for masters, as shown in Figure 5.2. While the majority of linkage levels (n = 655, 80%) performed as expected, nonmasters sometimes had a greater than .50 chance of providing a correct response to items measuring the linkage level. Although most linkage levels (n = 490, 60%) have a conditional probability of nonmasters providing a correct response less than .40, 50 (6%) have a conditional probability for nonmasters providing a correct response greater than .60, indicating there are many linkage levels where nonmasters are more likely than not to provide a correct response. This may indicate the items (and linkage level as a whole, since the item parameters are shared) were easily guessable or did not discriminate well between the two groups of students. Of these 50 linkage levels with a conditional probability for nonmasters providing a correct response greater than .60, the Successor linkage level was the most prevalent with 28 linkage levels (56%).

Figure 5.2

Probability of Nonmasters Providing a Correct Response to Items Measuring Each Linkage Level



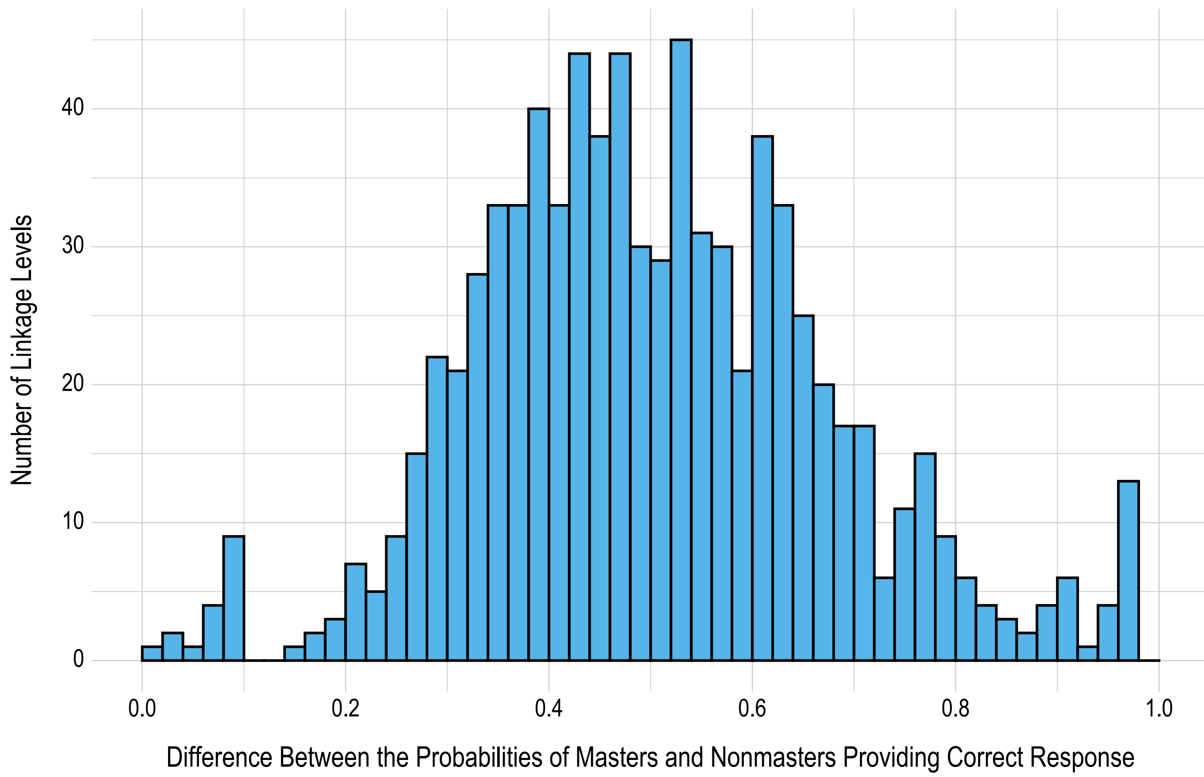
Note. Histogram bins are in increments of .02. Reference line indicates .50.

5.5.3. Item Discrimination

The discrimination of a linkage level represents how well the items are able to differentiate masters and nonmasters. For diagnostic models, this is assessed by comparing the conditional probabilities of masters and nonmasters providing a correct response. Linkage levels that are highly discriminating will have a large difference between the conditional probabilities, with a maximum value of 1.00 (i.e., masters have a 100% chance of providing a correct response and nonmasters a 0% chance). Figure 5.3 shows the distribution of linkage level discrimination values. Overall, 71% of linkage levels (*n* = 579) have a discrimination greater than .40, indicating a large difference between the conditional probabilities (e.g., .75 to .35, .90 to .50, etc.). However, there were 17 linkage levels (2%) with a discrimination of less than .10, indicating that masters and nonmasters tend to perform similarly on items measuring these linkage levels. Of these 17 linkage levels with a discrimination of less than .10, the Target linkage level was the most prevalent, with 7 linkage levels (41%).

Figure 5.3

Difference Between Masters' and Nonmasters' Probability of Providing a Correct Response to Items Measuring Each Linkage Level



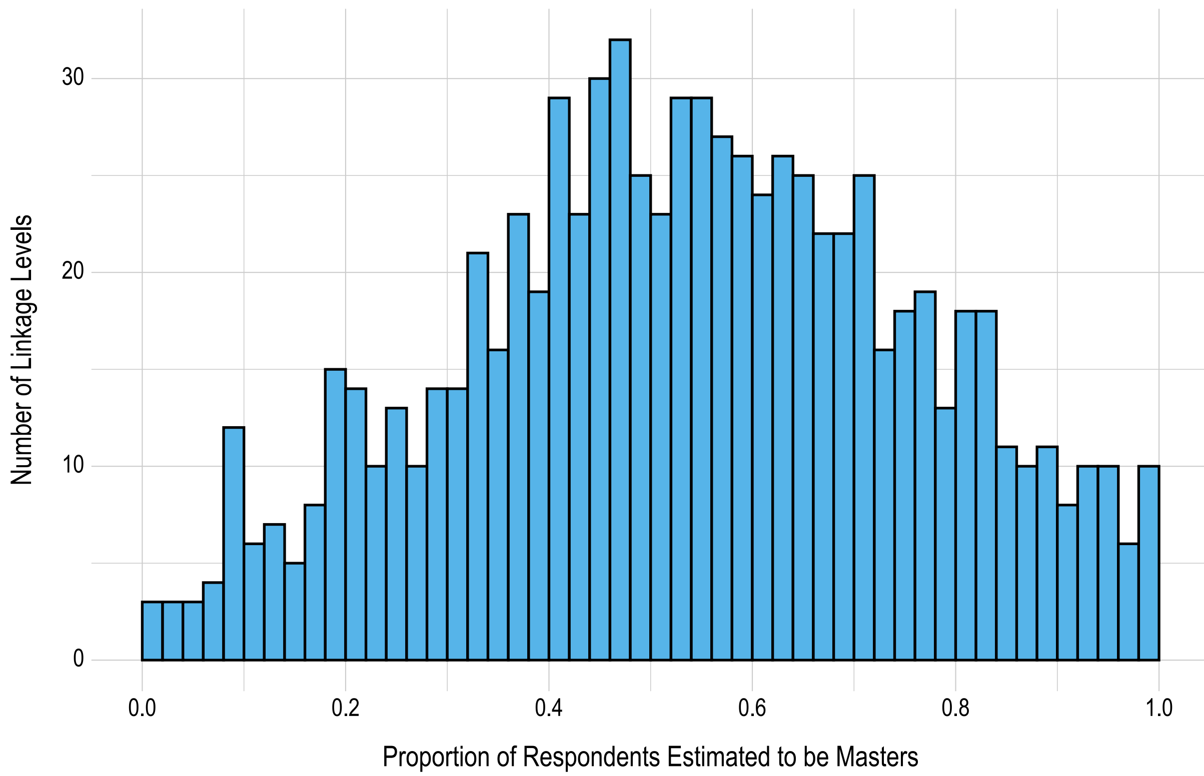
Note. Histogram bins are in increments of .02.

5.5.4. Base Rate Probability of Class Membership

The base rate of class membership is the DCM structural parameter and represents the estimated proportion of students in each class for each EE and linkage level. A base rate close to .50 indicates that students assessed on a given linkage level are, *a priori*, equally likely to be a master or nonmaster. Conversely, a high base rate would indicate that students testing on a linkage level are, *a priori*, more likely to be masters. Figure 5.4 depicts the distribution of the base rate probabilities. Overall, 69% of linkage levels (*n* = 565) had a base rate of mastery between .25 and .75. On the edges of the distribution, 97 linkage levels (12%) had a base rate of mastery less than .25, and 153 linkage levels (19%) had a base rate of mastery higher than .75. This indicates that students are more likely to be assessed on linkage levels they have mastered than those they have not mastered.

Figure 5.4

Base Rate of Linkage Level Mastery



Note. Histogram bins are shown in increments of .02.

5.6. Conclusion

In summary, the DLM modeling approach uses well-established research in Bayesian inference networks and diagnostic classification modeling to determine student mastery of skills measured by the assessment. A DCM is estimated for each linkage level of each EE to determine the probability of student mastery. Items within the linkage level are assumed to be fungible, with equivalent item probability parameters for masters and nonmasters, owing to the conceptual approach used to construct DLM testlets. An analysis of the estimated models indicates that the estimated models have acceptable levels of absolute model fit and classification accuracy. Additionally, the estimated parameters from each DCM are generally within the optimal ranges.

6. Standard Setting

The standard setting process for the Dynamic Learning Maps® (DLM®) Alternate Assessment System in English language arts (ELA) and mathematics was originally conducted following the 2014–2015 administration. Four performance level descriptors (PLDs) were developed to describe performance on the assessment. A 4-day standard setting meeting specified cut points and included a follow-up evaluation of impact data and cut points. Because of blueprint revisions adopted beginning in 2019–2020 (see Chapter 2), the original cut points were adjusted in spring 2022.[34]

This chapter provides a brief description of the development of the rationale for the standard setting approach; the policy PLDs; methods, preparation, procedures, and results of the original standard setting meeting and follow-up evaluation of the impact data and cut points; specification of grade- and subject-specific PLDs, which were developed after approval of the consortium cuts; and methods, preparation, procedures, and results from the adjustment to the cut points in 2022, including updated grade- and subject-specific PLDs. A more detailed description of the original DLM standard setting activities and results can be found in the *2015 Year-End Model Standard Setting: English Language Arts and Mathematics* Technical Report (Karvonen, Clark, & Nash, 2015) and in the corresponding peer-reviewed academic journal article (Clark et al., 2017).

6.1. Original Standard Setting Process

The 2014–2015 school year was the first fully operational testing year for the DLM assessments in ELA and mathematics. The operational testing window ended on June 12, 2015, and DLM staff conducted standard setting during June 15–18, 2015, in Kansas City, Missouri. The standard setting event included all states administering DLM assessments in 2014–2015 with the purpose of establishing a set of cut points for each of the two testing models. The DLM Technical Advisory Committee (TAC) advised on the standard setting methodology from early design through to the development of grade- and subject-specific PLDs and review of impact data after the event. Although the DLM Governance Board voted on acceptance of final cut points, individual states had the option to adopt the consortium cut points or develop their own independent cut points.

6.1.1. Standard Setting Approach: Rationale and Overview

The approach to standard setting was developed to be consistent with the DLM Alternate Assessment System's design and to rely on established methods, such as recommended practices for developing, implementing, evaluating, and documenting standard settings (Cizek, 1996; Hambleton et al., 2012) and the *Standards on Educational and Psychological Testing* (AERA et al., 2014). The DLM standard setting approach used DLM mastery classifications and drew from several established methods, including generalized holistic (Cizek & Bunch, 2006) and body of work (Kingston & Tiemann, 2012).

Because the DLM assessments are based on large, fine-grained learning maps and make use of diagnostic classification modeling rather than traditional psychometric methods, the standard setting approach relied on the aggregation of dichotomous classifications of linkage level mastery for each Essential Element (EE) in the blueprint. Drawing from the generalized holistic and body of work methods, the standard setting method used a profile approach to classify student mastery of linkage levels into performance levels (see

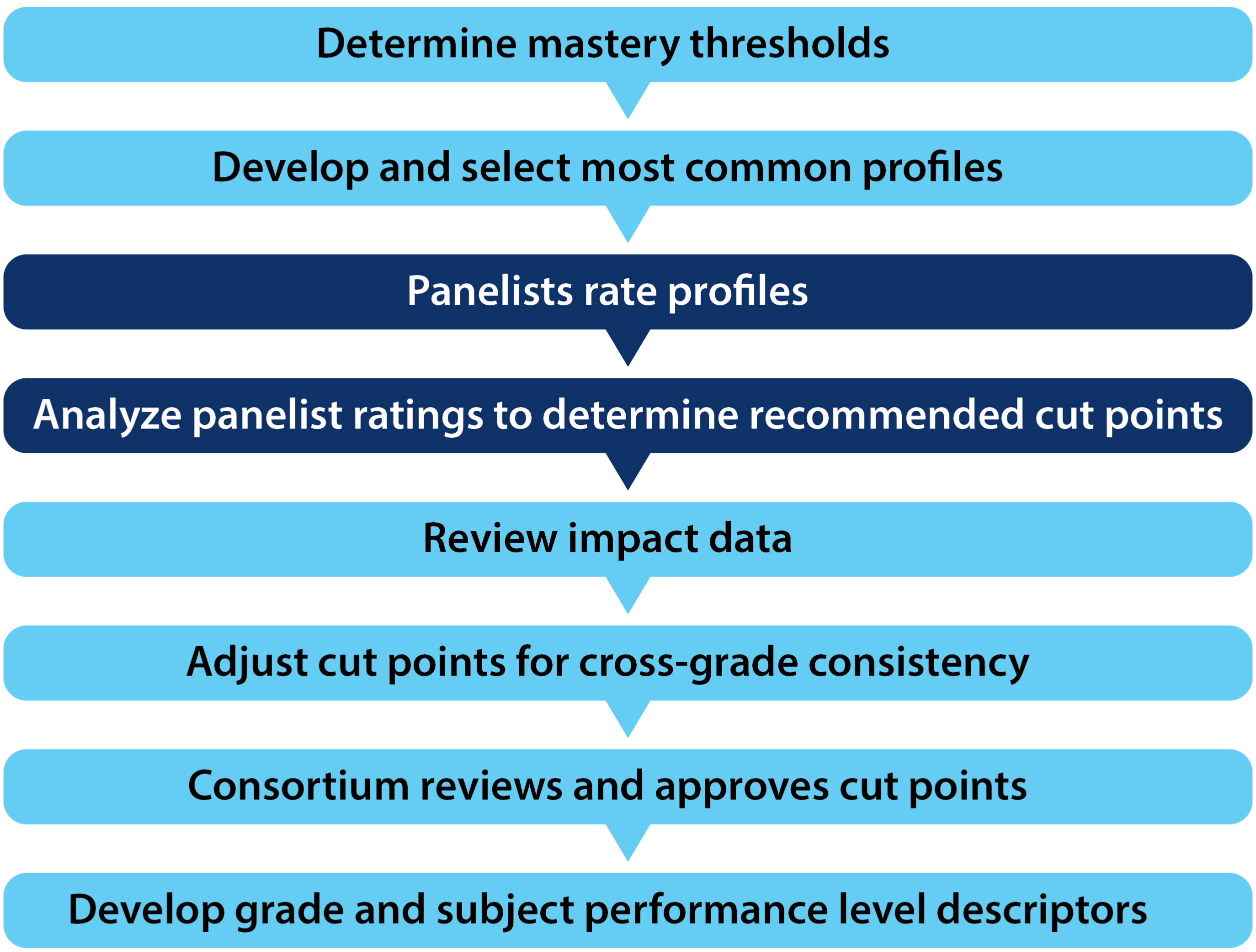
[34] Adjustment occurred in spring 2022 rather than 2020 due to COVID-19 affecting both the 2020 and 2021 administrations.

Clark et al., 2017). Profiles provided a holistic view of student performance by summarizing across the EEs and linkage levels. Cut points were determined by evaluating the total number of mastered linkage levels. Although the number of mastered linkage levels is not an interval scale, the process for identifying the DLM cut points is roughly analogous to assigning a cut point along a scale score continuum.

Before making a final decision whether to use the profile approach, the DLM TAC reviewed a preliminary description of the proposed methods. At the TAC's suggestion, DLM staff conducted a mock panel process using this profile-based approach to evaluate the feasibility of the rating task and the likelihood of obtaining sound judgments using this method. Figure 6.1 summarizes the complete set of sequential steps included in the DLM standard setting process. This includes steps conducted before, during, and after the on-site meeting during June 2015.

Figure 6.1

Steps of the DLM Standard Setting Process



Note. Dark shading represents steps conducted at the standard setting meeting in June 2015.

6.1.2. Policy Performance Level Descriptors

Student results are reported as performance levels, and PLDs are used to inform the interpretation of those scores. The DLM Governance Board developed PLDs through a series of discussions and draft PLD reviews between July and December 2014. Discussion began at the July 2014 governance meeting with governance board members in attendance who had special education and assessment backgrounds. As part of the discussion, the group reviewed the language used in the general education consortia and in the Common Core State Standards for key features describing performance. Following the meeting, governance board members took draft PLDs back to their states and were responsible for collecting feedback at the state and local level according to their own state policies and practices for stakeholder involvement. Table 6.1 presents the final version of policy PLDs. The consortium-level definition of proficiency was *at target*. Policy PLDs served as anchors for panelists during the standard setting process.

Table 6.1

Final Performance Level Descriptors for the DLM Consortium

Performance level descriptors
The student demonstrates *emerging* understanding of and ability to apply content knowledge and skills represented by the Essential Elements.
The student's understanding of and ability to apply targeted content knowledge and skills represented by the Essential Elements is *approaching the target*.
The student's understanding of and ability to apply content knowledge and skills represented by the Essential Elements is *at target*.
The student demonstrates *advanced* understanding of and ability to apply targeted content knowledge and skills represented by the Essential Elements.

6.1.3. Profile Development

Prior to the standard setting meeting, DLM staff generated profiles of student learning that summarized linkage level mastery for each assessed EE. First, for each EE and linkage level that was assessed, we calculated the students' probability of mastery using the diagnostic classification model (see Chapter 5 of this manual). For each linkage level, all students with a probability greater than or equal to .8 received a linkage level mastery status of 1, or mastered. All students with a probability lower than .8 received a linkage level mastery status of 0, or not mastered.[35]

The threshold value was applied to student assessment data to create profiles of student mastery, which summarize linkage level mastery by EE. Profiles were created using data for each subject and grade. Each profile listed all the linkage levels for all the EEs from the blueprint, along with the conceptual area for each EE, with shaded boxes indicating the mastered linkage levels. Figure 6.2 provides an example profile for a hypothetical student.

[35] Maximum uncertainty occurs when the probability is .5 and maximum certainty when the probability approaches 0 or 1. Considering the risk of false positives and negatives, the threshold used to determine mastery classification was set at .8.

Figure 6.2

Example Standard Setting Profile for a Hypothetical Student

		Level Mastery				
Area	Essential Element	Initial Precursor	Distal Precursor	Proximal Precursor	Target	Successor
ELA.C1.1	ELA.RL.4.1	Identify familiar people, objects, places, or events	Identify character actions in a familiar story	Identify character actions	Recount events in a story using details	Recount the key details of a story
ELA.C1.2	ELA.RL.4.2	Identify familiar people, objects, places, or events	Identify major events in a familiar story	Identify a character's actions and corresponding consequences	Identify the theme of a familiar story	Identify the specific theme of a story
ELA.C1.1	ELA.RL.4.3	Understand object names	Identify concrete details in a familiar story	Identify characters, setting, and major events	Describe characters in a narrative	Describe characters, setting, and events
ELA.C1.2	ELA.RL.4.4	Understand object names	Identify the meaning of words	Identify words or phrases to complete a literal sentence	Identify the meaning of an unambiguous word	Identify multiple meanings of a word
ELA.C1.1	ELA.RL.4.5	Identify familiar people, objects, places, or events	Name or identify objects in pictures	Identify the beginning, middle, and end of a familiar story	Identify story characteristics	Identify story elements that change
ELA.C1.2	ELA.RL.4.6	Understand object names	Identify character actions in a familiar story	Identify character actions	Identify the narrator of a story	Identify narrator point of view
ELA.C1.1	ELA.RI.4.1	Understand object names	Name or identify objects in pictures	Identify concrete details in an informational text	Identify explicit details in informational texts	Identify words related to explicit information
ELA.C1.1	ELA.RI.4.2	Understand object names	Name or identify objects in pictures	Identify concrete details in informational texts	Identify the overall topic of a familiar text	Identify topic-related words in an informational text
ELA.C1.1	ELA.RI.4.3	Understand object names	Use category knowledge to draw conclusions	Identify concrete details in an informational text	Identify concrete details related to people, events, or ideas	Compare key details
ELA.C1.2	ELA.RI.4.4	Understand object names	Identify the meaning of words	Identify words or phrases to complete a literal sentence	Identify the meaning of an unambiguous word	Identify the multiple meanings of a word

Note. Green shading represents linkage level mastery.

Profiles were available for all students who participated in the spring assessment window by May 15, 2015 (*n* = 49,958). The frequency with which each precise profile (i.e., pattern of linkage level mastery) occurred in this population was computed. Based on these results, the three most common profiles were selected for each possible total linkage level mastery value (i.e., total number of linkage levels mastered) for each grade and subject. In instances in which data were not available at a specific linkage level value (e.g., no students mastered exactly 47 linkage levels for a grade and subject), profiles were based on simulated data. To simulate profiles, the DLM test development teams used adjacent profiles for reference and created simulated profiles that represented likely patterns of mastery. Fewer than 10% of all the profiles developed were simulated.[36]

6.1.4. Panelists

DLM staff worked with participating states in March 2015 to recruit standard setting panelists. States were responsible for communicating within their state to recruit potential panelists. Panelists sought were those with both content knowledge and expertise in the education and outcomes of students with the most significant cognitive disabilities, including educators and school and district administrators. Other subject

[36] Further detail on specific procedures for preparing standard setting profiles may be found in Chapter 1 of the *2015 Year-End Model Standard Setting: English Language Arts and Mathematics* Technical Report (Karvonen, Clark, & Nash, 2015).

matter experts, such as higher education institution faculty or state and regional educational staff, were also suggested for consideration. Employers were considered at the high school level only, specifically targeting companies that employ individuals with disabilities.

The 46 panelists who participated in standard setting represented varying backgrounds. Table 6.2 and Table 6.3 summarize their demographic information. Most of the selected panelists were classroom educators. Panelists had a range of years of experience with ELA, mathematics, and working with students with the most significant cognitive disabilities.

Half of the participants had experience with setting standards for other assessments (*n* = 23). Some panelists already had experience with the DLM assessment, either from writing items (*n* = 4) or externally reviewing items and testlets (*n* = 18). Only two panelists reported having less than 1 year or no experience with alternate assessments; both were classroom educators with at least 13 years of experience working with students with the most significant cognitive disabilities.[37]

Table 6.2

Panelist Demographic Characteristics

Demographic subgroup	*n*
Gender	
Female	42
Male	4
Race	
White	35
African American	5
Asian	2
Not disclosed	2
American Indian/Alaska Native	1
Hispanic/Latino	1
Native Hawaiian/Pacific Islander	0
Professional role	
Classroom educator	32
District staff	6
Other	6
Building administrator	2
University faculty/staff	0

[37] Further detail on standard setting volunteers, selection process, and panel composition may be found in Chapter 3 of the *2015 Year-End Model Standard Setting: English Language Arts and Mathematics* Technical Report (Karvonen, Clark, & Nash, 2015).

Table 6.3

Panelist Years of Experience

Experience type	Average	Minimum	Maximum
English language arts	15.8	0	50
Mathematics	15.9	0	49
Students with significant cognitive disabilities	15.1	1	36

6.1.5. Meeting Procedures

Panelists participated in a profile-based standard setting procedure to make decisions about cut points. The panelists participated in four rounds of activities in which they moved from general to precise recommendations about cut points.

The primary tools of this procedure were range-finding folders and pinpointing folders. The range-finding folders contained profiles of student work that represented the scale range. Pinpointing folders contained profiles for specific areas of the range.

Throughout the procedure, DLM staff instructed panelists to use their best professional judgment and consider all students with the most significant cognitive disabilities to determine which performance level best described each profile. Each panel had at least two, and up to three, grade-level cut points to set.

The subsequent sections provide details of the final procedures, including quality assurance used for determining cut points.[38]

6.1.5.1. Training

Panelists were provided with training both before and during the standard setting workshop. Advance training was available online, on demand, in the 10 days prior to the standard setting workshop. The advance training addressed the following topics:

1. Students who take the DLM assessments
2. Content of the assessment system, including DLM learning maps, EEs, claims and conceptual areas, linkage levels, and alignment
3. Accessibility by design, including the framework for the DLM System's cognitive taxonomy and strategies for maximizing accessibility of the content; the use of the Personal Needs and Preferences Profile to provide accessibility supports during the assessment; and the use of the First Contact survey to determine linkage level assignment
4. Assessment design, including item types, testlet design, and sample items from various linkage levels in both subjects
5. An overview of the assessment model, including test blueprints and the timing and selection of testlets administered

[38] Further information regarding all meeting procedures and fidelity of the final procedures to the planned procedures may be found in Chapter 4 and the appendix of the *2015 Year-End Model Standard Setting: English Language Arts and Mathematics* Technical Report (Karvonen, Clark, & Nash, 2015).

6. A high-level introduction to two topics that would be covered in more detail during on-site training: the DLM approach to scoring and reporting and the steps in the standard setting process

Additional panelist training was conducted at the standard setting workshop. The purposes of on-site training were twofold: (1) to review advance training concepts that panelists had indicated less comfort with and (2) to complete a practice activity to prepare panelists for their responsibilities during the panel meeting. The practice activity consisted of range finding using training profiles for just a few total linkage levels mastered (e.g., 5, 10, 15, 20). Overall, panelists participated in approximately 8 hours of standard setting training before beginning the practice activity.

6.1.5.2. Range Finding

During the range-finding process, panelists reviewed a limited set of profiles to assign general divisions between the performance levels using a two-round process. The goal of range finding was to locate ranges (in terms of number of linkage levels mastered) in which panelists agreed that approximate cut points should exist.

First, panelists independently evaluated profiles and identified the performance level that best described each profile. Once all panelists completed their ratings, the facilitator obtained the performance level recommendations for each profile by a raise of hands.

After a table discussion of how panelists chose their ratings, the panelists were given the opportunity to adjust their independent ratings if they chose. A second round of ratings were recorded and shared with the group.

Using the second round of ratings, built-in logistic regression functions were used to calculate the probability of a profile being categorized in each performance level, conditioned on the number of linkage levels mastered, and the most likely cut points for each performance level were identified. In instances in which the logistic regression function could not identify a value (i.e., the group unanimously agreed on the categorization of profiles to performance levels, so there was no variance in the ratings to fit a logistic regression), the approximate cut point was determined as the midway point between the unanimous profiles. For example, if all profiles with 10 linkage levels mastered were unanimously rated as the Emerging performance level, and all profiles with 15 linkage levels mastered were unanimously rated as the Approaching the Target performance level, the approximate cut point was set at 13.[39]

6.1.5.3. Pinpointing

Pinpointing rounds followed range finding. During pinpointing, panelists reviewed additional profiles to refine the cut points. The goal of pinpointing was to pare down to specific cut points in terms of the number of linkage levels mastered within the general ranges determined in range finding, while not relying on conjunctive or compensatory judgments.

First, panelists reviewed profiles for the seven linkage levels including and around the cut point value identified during range finding. Next, panelists independently evaluated the leveled profiles and assigned each a performance level—those in the higher level and those in the lower level. Once all panelists

[39] Chapter 4 of the *2015 Year-End Model Standard Setting: English Language Arts and Mathematics* Technical Report (Karvonen, Clark, & Nash, 2015) provides greater detail on range finding and pinpointing and includes the number of linkage levels per grade and subject.

completed their ratings, the facilitator obtained the recommendations for each profile by a raise of hands.

After discussion of the ratings, a second round of ratings commenced. Panelists were given the opportunity to adjust their independent ratings if they chose. Using the second round of ratings, built-in logistic regression functions were used to calculate the probability of a profile being categorized in each performance level, conditioned on the number of linkage levels mastered, and the most likely cut points for each performance level were identified. In instances in which the logistic regression function could not identify a value (e.g., the group unanimously agreed on the categorization of profiles to performance levels), psychometricians evaluated the results to determine the final recommended cut point based on panelist recommendations.[40]

6.1.5.4. Panelist Evaluations and Panel-Recommended Cut Points

Across all cut points (*N* = 414), panelists indicated they were comfortable with the group-recommended cut point in 94.0% of cases. Table 6.4 provides the panelist comfort rating of group-recommended cut points. Only 6.0% of responses (*n* = 25) indicated a discomfort with a group-recommended cut point. For 9 out of 18 cut point panels (50.0%; i.e., one panel for each grade and subject), panelists indicated comfort with all three recommended cut points. Most recommendations for a change to the cut point were for only one of the three cut points for a given panel, and most often, the recommended changes differed from the initial recommendation by only a single linkage level.[41]

Table 6.4

Panelist Comfort With Group-Recommended Cut Points

Subject	*N* Panelists	*N* Ratings (*n* panelists × *n* cut points evaluated)	"Yes" ratings	% Agreement
English language arts	24	216	206	95
Mathematics	22	198	183	92

6.1.6. Smoothing the Cut Points

To mitigate the effect of sampling error and issues related to a system of cut points across a series of grade levels, adjustments were made to the panel-recommended cut points in an effort to systematically smooth distributions within the system of cut points being considered.[42] The goal of the smoothing process was to have a more consistent percentage of students in each performance level and across grade levels within each subject. The smoothing process followed these steps:

---

40 Chapter 4 of the *2015 Year-End Model Standard Setting: English Language Arts and Mathematics* Technical Report (Karvonen, Clark, & Nash, 2015) provides greater detail on range finding and pinpointing and includes the number of linkage levels per grade and subject.

41 Chapter 5 of the *2015 Year-End Model Standard Setting: English Language Arts and Mathematics* Technical Report (Karvonen, Clark, & Nash, 2015) provides greater detail on final independent evaluations of panel-recommended cut points.

42 The specific steps applied to each subject within each grade level can be found in the *2015 Year-End Model Standard Setting: English Language Arts and Mathematics* Technical Report (Karvonen, Clark, & Nash, 2015).

1. For each grade and subject, calculate the cumulative percentage of students at each number of total linkage levels mastered.
2. Perform a probit transformation to convert each cumulative percentage to a *z*-score.
3. Find the *z*-score associated with each of the panel-recommended cut points.
4. For each *z*-score identified in Step 3, calculate a new weighted *z*-score by assigning 0.5 weight to the current *z*-score, and 0.25 weight to each adjacent grade.[43] For example, when calculating the weighted *z*-score for the Grade 4 cut point between the Emerging and Approaching performance levels, 0.5 weight would be given to the *z*-score for the Grade 4 Emerging/Approaching cut point, 0.25 weight would be given to the *z*-score for the Grade 3 Emerging/Approaching cut point, and 0.25 weight would be given to the *z*-score for the Grade 5 Emerging/Approaching cut point.
5. For each grade and subject, the total linkage levels mastered associated with the *z*-score closest to the weighted average for each cut point is the smoothed cut point.

For a complete description of the smoothing process, see the *2015 Year-End Model Standard Setting: English Language Arts and Mathematics* Technical Report (Karvonen, Clark, & Nash, 2015).

6.1.7. Results

This section summarizes the panel-recommended and smoothed cut points and presents impact data for the final cut points.[44]

6.1.7.1. Panel Recommended and Smoothed Cut Points

Table 6.5 displays the cut point recommendations reached by the panelists following the range-finding and pinpointing process.

[43] For Grades 3 and 11, which had only one adjacent grade, 0.667 weight was given to the current grade, and 0.333 weight was given to the adjacent grade.

[44] Additional detailed results are provided in Chapter 5 of the *2015 Year-End Model Standard Setting: English Language Arts and Mathematics* Technical Report (Karvonen, Clark, & Nash, 2015).

Table 6.5

Final English Language Arts and Mathematics Panel Cut Point Recommendations

Grade	Emerging/ Approaching	Approaching/ Target	Target/ Advanced	Total linkage levels possible
English language arts				
3	40	55	73	80
4	35	55	74	85
5	43	59	79	85
6	19	41	63	80
7	23	48	67	90
8	26	51	69	85
9	19	50	72	85
10	15	47	73	85
11	23	48	69	85
Mathematics				
3	15	24	42	55
4	19	29	50	80
5	13	30	39	75
6	16	26	42	55
7	18	41	51	70
8	22	37	53	70
9	9	26	34	40
10	6	16	37	45
11	13	24	39	45

As described in section 6.1.6, a smoothing procedure was applied to the panel-recommended cut points to mitigate the effect of sampling error and issues related to a system of cut points across a series of grade levels. Table 6.6 shows the smoothed cut points that were derived from the methods described above.

Table 6.6

Smoothed Cut Point Recommendations

Grade	Emerging/ Approaching	Approaching/ Target	Target/ Advanced	Total linkage levels possible
English language arts				
3	36	50	71	80
4	38	57	75	85
5	35	53	76	85
6	27	46	65	80
7	27	52	73	90
8	23	48	72	85
9	20	48	68	85
10	17	47	72	85
11	18	47	70	85
Mathematics				
3	12	21	37	55
4	20	30	56	80
5	15	32	48	75
6	13	28	38	55
7	19	37	53	70
8	17	40	53	70
9	10	21	33	40
10	8	21	36	45
11	8	18	38	45

6.1.7.2. Final Impact Data

Figure 6.3 and Figure 6.4 display the results of the smoothed cut points in terms of impact for ELA and mathematics, respectively.[45] Table 6.7 includes the demographic data for students included in the impact data.

[45] Chapter 5 of the *2015 Year-End Model Standard Setting: English Language Arts and Mathematics* Technical Report (Karvonen, Clark, & Nash, 2015) reports the frequency distributions for the panel-recommended cut points.

Figure 6.3

English Language Arts Impact Data Using Smoothed Cut Points

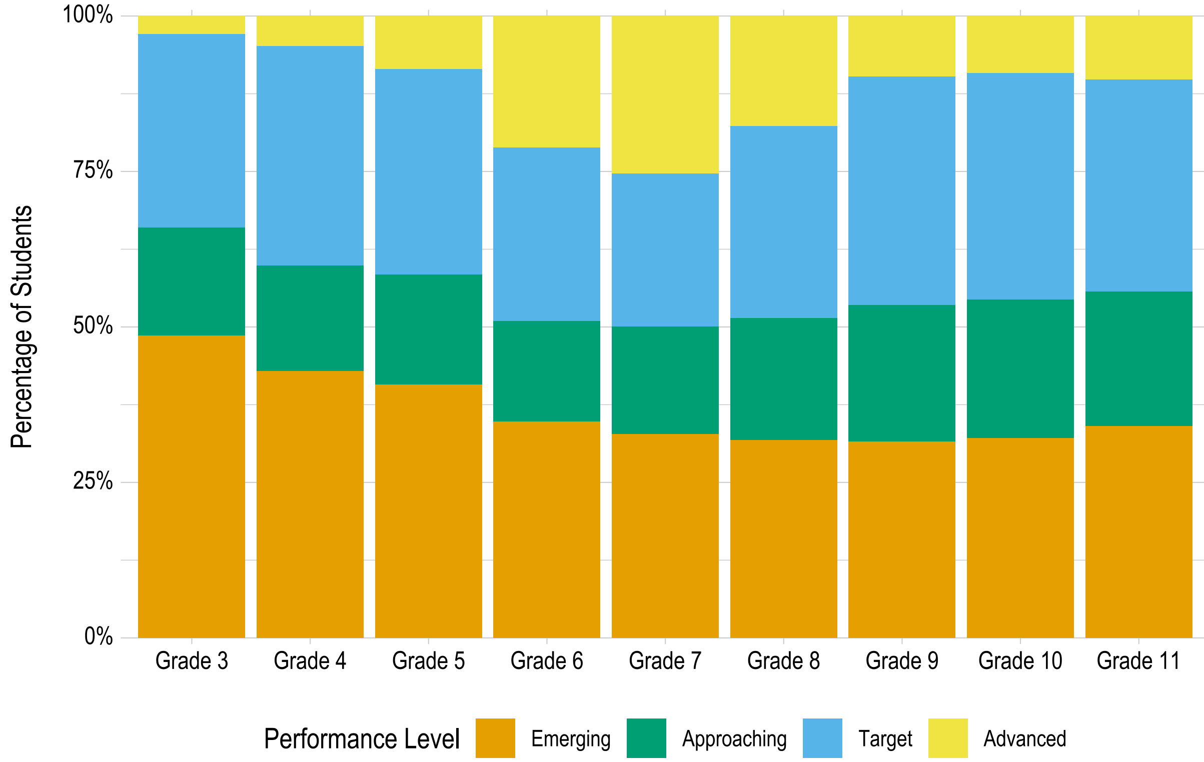


Figure 6.4

Mathematics Impact Data Using Smoothed Cut Points

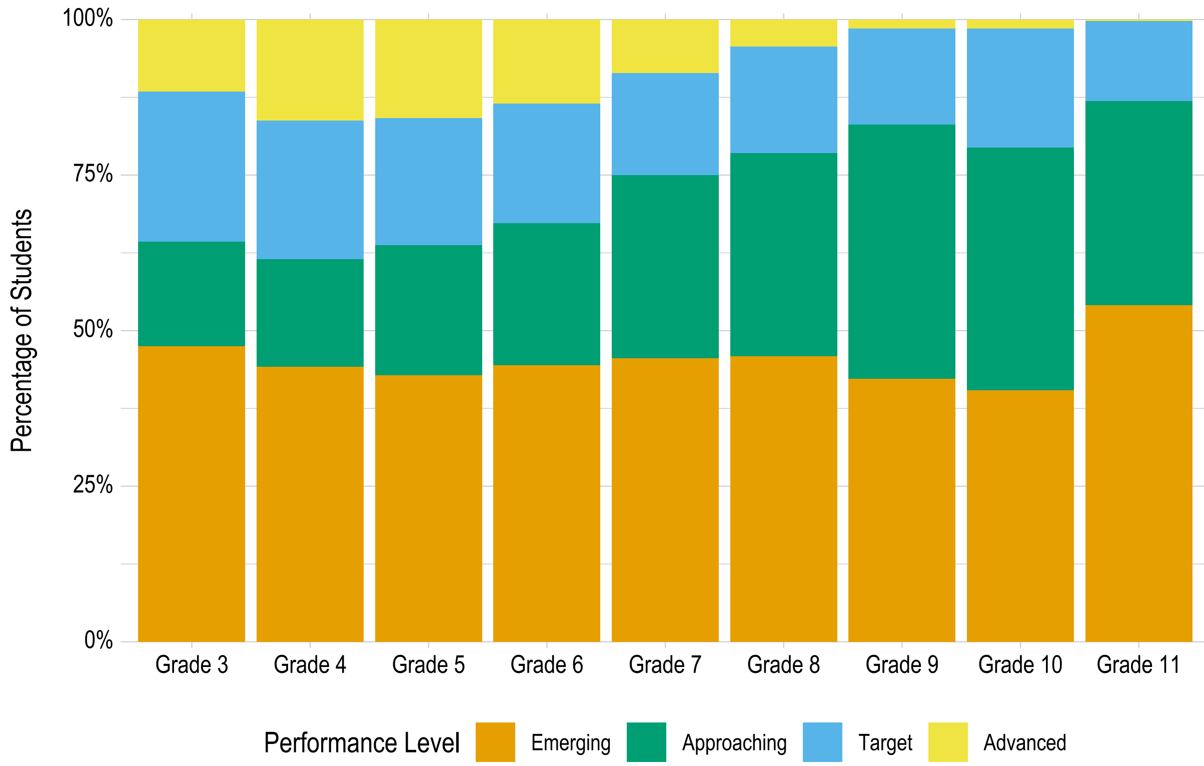


Table 6.7

Demographic Information for Students Included in Impact Data (N = 54,204)

Subgroup	*n*	%
Gender		
Male	35,132	64.8
Female	18,068	33.3
Missing†	1,004	1.9
Primary disability		
Intellectual disability	6,250	11.5
Autism	5,097	9.4
Other health impairment	4,526	8.3
Multiple disabilities	1,455	2.7
Specific learning disability	637	1.2
Other	2,429	4.5
Missing†	33,810	62.4
Race		
White	28,459	52.5

Table 6.7

*Demographic Information for Students Included in Impact Data (*N *= 54,204) (continued)*

Subgroup	*n*	%
African American	9,309	17.2
Two or more races	6,741	12.4
American Indian	3,723	6.9
Asian	1,645	3.0
Native Hawaiian/Pacific Islander	388	0.7
Alaska Native	244	0.5
Missing†	3,695	6.8
Hispanic ethnicity		
Non-Hispanic	22,776	42.0
Hispanic	4,790	8.8
Missing†	26,638	49.1
English learning (EL) participation		
Not EL eligible or monitored	51,547	95.1
EL eligible or monitored	2,630	4.9
Missing†	27	<0.1
English language arts complexity band		
Foundational	9,306	17.2
Band 1	15,324	28.3
Band 2	19,636	36.2
Band 3	9,937	18.3
Missing	1	<0.1
Mathematics complexity band		
Foundational	10,282	19.0
Band 1	16,433	30.3
Band 2	20,398	37.6
Band 3	7,089	13.1
Missing	2	<0.1

† Demographic variables were not required in 2014–2015.

6.1.8. External Evaluation of Standard Setting Process and Results

The DLM TAC chair was on-site for the duration of the standard setting event and reported that the standard setting meeting was well planned and implemented, the staff were helpful to the panelists, and the panelists worked hard to set standards. The full TAC accepted a resolution about the adequacy, quality of judgments, and extent to which the process met professional standards.[46]

[46] The TAC chair memorandum and TAC resolution are provided in Appendix L of the *2015 Year-End Model Standard Setting: English Language Arts and Mathematics* Technical Report (Karvonen, Clark, & Nash, 2015).

The panel-recommended cut points, adjusted cut points, and associated impact data for both sets of cut points were presented to the TAC and governance board for review. The TAC supported the DLM smoothing method and resulting adjusted cut points. Following the states' review process and discussion with DLM staff, the DLM Governance Board voted to accept the DLM-recommended smoothed cut points as the final consortium cut points with no further adjustment.

6.1.9. Grade Level and Subject Performance Level Descriptors

Based on the general approach to standard setting, which relied on mastery profiles to anchor panelists' content-based judgments, grade- and subject-specific PLDs were not used during standard setting. Instead, they emerged based on the final cut points and example profiles, and they were syntheses of content from the more fine-grained linkage level descriptors. Grade- and subject-specific PLDs were completed after standard setting in 2015.

Standard setting panelists began the process by drafting lists of skills and understandings that they determined were characteristic of specific performance levels after cut points had been established. In general, these draft lists of skills and understandings were based on linkage levels described in the mastery profiles used for standard setting—either separate linkage level statements or syntheses of multiple statements.

These draft lists of important skills were collected and used as a starting point for DLM test development teams as they developed language for grade- and subject-specific descriptions for each performance level in every grade for both ELA and mathematics. The purpose of these content descriptions was to provide information about the knowledge and skills that are typical for each performance level.

Test development teams prepared to draft PLDs by consulting published research related to PLD development (e.g., Perie, 2008) and reviewing PLDs developed for other assessment systems to consider grain size of descriptive language and formats for publication. In addition to the draft lists generated by standard setting panelists, test development teams used the following materials as they drafted specific language for each grade- and subject-specific PLD:

- DLM assessment blueprints
- Cut points set at standard setting for each grade and subject
- Sample mastery profiles from the standard setting event
- Essential Element Concept Maps for each EE included on the blueprint for each grade level
- Linkage level descriptions and associated sections of the DLM learning maps for every EE
- *The Standards of Mathematical Practice*

Test development teams reviewed the EEs, Essential Element Concept Maps, and linkage level descriptors on the profiles to determine skills and understandings assessed at the grade level. These skills and understandings come from each conceptual area assessed at the specific grade level and vary from one grade to the next. Then, the teams reviewed the draft skill lists created by standard setting panelists and final cut points approved by the consortium. Test development teams then used the sample mastery profiles to consider the types and ranges of student performances that could lead to placement into specific performance levels. Using these multiple sources of information, the test development teams evaluated the placement of skills into each of the four performance levels.

While not an exhaustive list of all the content related to each EE from the DLM learning maps, the synthesis of standard setting panelist judgments and test development team judgments provided the basis for descriptions of the typical performance of students showing mastery at each performance level. As test development teams drafted PLDs for each grade, they reviewed the descriptors in relation to each other and the underlying DLM learning map to ensure that there was differentiation in skills from one grade to the next. In very few cases, in which panelists recommended skill placement that was inconsistent with development of content knowledge as represented in the DLM maps, test development teams adjusted the placement of skills. This was only done in cases in which the original judgment of the panelists was inconsistent with a logical ordering of skill development from one level to the next in a particular grade.

DLM staff prepared initial drafts of the grade- and subject-specific descriptions for Grade 3. Project staff reviewed these drafts internally. Additional drafts were prepared for Grades 4 and 5. The DLM Governance Board reviewed a combination of Grades 3, 4, and 5 at the December 2015 consortium governance meeting. Project staff asked the governance board to review the progression of descriptors from grade to grade within the four performance levels in Grades 3, 4, and 5 and to provide general feedback to the initial drafts. Feedback from the governance board focused on utility for educators and parents and structuring the descriptions to make them more user-friendly. The primary responses to governance board feedback were to:

- Review technical language in existing drafts and simplify wherever possible.
- Organize each grade and subject-specific description so that a broad conceptual statement about what students at a performance level typically knew and were able to do was followed by specific skills and understandings shown in bulleted lists.
- Organize descriptions consistently within and across grades so that related skills were described in the same sequence within each level in a grade.

DLM staff delivered drafts of all grade- and subject-specific descriptions to the governance board for review in February 2016. After the review period ended, test development teams responded to feedback received by adjusting technical descriptions, removing any content that exceeded the requirements of EEs in the grade level, simplifying language, and clarifying descriptions of skills and understandings. These adjustments were followed by a full editorial review.

6.2. Standards Adjustment

In 2019–2020, Year-End model states adopted revised assessment blueprints in ELA and mathematics (see Chapter 2 of this manual for a description of the blueprint revisions). In short, the blueprint revisions reduced the total number of EEs that were assessed in each grade, which also reduced the total number of linkage levels available to be mastered. Because the original standard setting process resulted in cut points that were applied to the total number of linkage levels mastered in the subject, changing the number of linkage levels available required an adjustment to the original cut points set in 2015.

6.2.1. Adjustment Procedures

When the blueprint revisions were adopted, the governance board, in consultation with the TAC, opted for a statistical adjustment to the existing cut points rather than conducting a full standard setting. The statistical adjustment consisted of two steps:

1. Reduce each cut point in proportion to the number of EEs that were removed from the blueprint in each grade and subject.
2. Adjust the proportionally reduced cut points ±2 using impact data from spring 2020 (after implementing the revised blueprints) to preserve the reported performance level distributions that were reported in 2018–2019.

As described in Chapter 2, the blueprint revision process was designed to keep the construct constant. Thus, performance level distributions should also be constant. The ±2 adjustment accounts for potential differences in EEs that were removed from or retained on the blueprint (e.g., difficulty, reliability) to preserve the performance level distributions.

Due to the cancellation of assessment in spring 2020 and continued disruptions in 2021 (Accessible Teaching, Learning, and Assessment Systems [ATLAS], 2021) resulting from the COVID-19 pandemic, on advice of the TAC, the second step of the adjustment process was postponed until spring 2022. In May 2022, DLM staff collected all data for students who completed their assessment by May 6, 2022. Using the May 6 sample, we calculated the total linkage levels mastered for each student in each subject. The DLM staff then applied the proportionally reduced cut points from step 1 to calculate the percentage of students within each performance level. These percentages were compared to the corresponding percentages that were reported in 2018–2019, and the cut points were adjusted ±2 to make the 2022 percentages from the May 6 sample be more consistent with the 2018–2019 percentages.

6.2.2. Sample Representativeness

In consultation with the TAC, thresholds for proceeding with the adjustment process were put in place to ensure that the sample of students who completed their assessments by May 6, 2022, was representative of the full DLM student population. First, DLM staff set minimum sample size thresholds to ensure a sufficient number of students in each grade and subject and to get a reliable estimate of the percentage of students in each performance level. Specifically, DLM staff determined that the number of students with a completed assessment in each grade and subject by May 6, 2022, must be at least 20% of the total number of students who completed assessments during spring 2019 (the most recent year in which participation was not affected by the COVID-19 pandemic).

Additionally, DLM staff required that at least 11 of the 14 Year-End model states be represented in each grade or high school grade band. Table 6.8 shows the minimum sample size threshold and the observed number of completed assessments by May 6, 2022. Overall, the number of students who completed assessments by May 6, 2022, was well above the minimum thresholds. All 14 participating states were represented in all grades or grade bands, and in total, the students with completed assessments represented approximately 70% of all students from Year-End states who were enrolled in the DLM assessment system for 2022.

Table 6.8

Sample Size Thresholds and Observed Completions by May 6, 2022

Grade	*N* (2019)	Minimum *N* (2022)	Observed *N* (2022)	Observed states
English language arts				
3	8,777	1,756	7,754	14
4	9,518	1,904	8,070	14
5	9,525	1,905	8,103	14
6	9,559	1,912	8,169	14
7	9,555	1,911	8,341	14
8	9,432	1,887	8,535	14
9	6,739	1,348	4,081	14*
10	3,454	691	2,613	14*
11	6,137	1,228	6,529	14*
Mathematics				
3	8,763	1,753	7,757	14
4	9,503	1,901	8,082	14
5	9,526	1,906	8,165	14
6	9,551	1,911	8,639	14
7	9,544	1,909	8,796	14
8	9,426	1,886	8,702	14
9	6,736	1,348	4,303	14*
10	3,451	691	2,646	14*
11	6,143	1,229	7,008	14*

* State participation in grades 9–12 was grade banded to account for differences in high school testing requirements across states.

In addition to the overall size of the sample, we also put in place a representativeness threshold for proceeding. The representativeness threshold was designed to ensure that the students included in the May 6 sample were not systematically different from those students who completed their assessments later in the testing window. Using the May 6 sample, DLM staff calculated the percentage of students in each demographic subgroup for each grade and subject (i.e., gender, race, Hispanic ethnicity, English language learner status, and complexity band).

The DLM staff then calculated the percentage of students from the full sample of students enrolled in the system, regardless of completion status, in each of the same subgroups and calculated the Cramér's (1946) V effect size for the differences in distribution of each demographic group across samples (e.g., the distribution of race should be consistent across the May 6 and enrolled student population samples). Cramér's V is based on the chi-squared test of association and can be interpreted using the guidelines suggested by Cohen (1988).

To proceed, we required that all effect sizes must be in the negligible range (i.e., less than 0.2 using Cohen's guidelines). Table 6.9 shows the effect sizes for each grade and demographic variable for ELA and mathematics. Across both subjects, the largest effect size was 0.06, which was well below the threshold of 0.2. Thus, the sample of students who completed their assessment by May 6, 2022, was demographically consistent with the full sample of students enrolled in the system.

Table 6.9

Cramér's V Effect Sizes for Difference Between the May 6 Sample and Full Population

Grade	Gender	Race	Hispanic ethnicity	English language learner status	Complexity band
English language arts					
3	.008	.020	.014	.016	.029
4	.008	.025	.012	.031	.036
5	.006	.027	.002	.023	.040
6	.003	.020	.014	.027	.041
7	.008	.026	.001	.026	.042
8	.005	.022	.011	.026	.036
9	.016	.039	.019	.057	.034
10	.005	.013	<.001	.010	.031
11	.008	.020	.006	.008	.027
Mathematics					
3	.007	.021	.010	.017	.046
4	.007	.024	.010	.034	.044
5	.003	.027	<.001	.023	.050
6	.003	.017	.009	.023	.043
7	.007	.022	.006	.027	.039
8	.008	.021	.013	.025	.044
9	.016	.034	.005	.061	.047
10	.006	.014	<.001	.012	.030
11	.003	.017	.009	.007	.031

6.2.3. Results

In this section, the proportionally reduced and adjusted cut points, as well as impact data, are summarized.

6.2.3.1. Proportionally Reduced and Adjusted Cut Points

Table 6.10 shows the proportionally reduced cut points that result from reducing each cut point in Table 6.6 in proportion to the number of EEs removed from the assessment blueprint.

Table 6.10

Proportionally Reduced Cut Points

Grade	Emerging/ Approaching	Approaching/ Target	Target/ Advanced	Total linkage levels possible
English language arts				
3	22	31	44	50
4	25	37	49	55
5	21	31	45	50
6	19	32	45	55
7	20	38	53	65
8	18	37	55	65
9	16	40	56	70
10	14	39	59	70
11	15	39	58	70
Mathematics				
3	9	15	27	40
4	10	15	28	40
5	8	17	26	40
6	8	18	24	35
7	10	18	26	35
8	10	23	30	40
9	9	18	29	35
10	7	19	32	40
11	5	12	25	30

As described in section 6.2.1, an adjustment procedure was applied to the proportionally reduced cut points to preserve the reported performance level distributions that were reported in 2018–2019. Table 6.11 shows the final cut points derived from the adjustment process.

Table 6.11

2022 Adjusted Cut Point Recommendations

Grade	Emerging/ Approaching	Approaching/ Target	Target/ Advanced	Total linkage levels possible
English language arts				
3	22	31	44	50
4	23	37	49	55
5	21	33	45	50
6	21	33	45	55
7	18	37	53	65
8	18	35	54	65
9	17	39	55	70
10	15	38	59	70
11	15	39	58	70
Mathematics				
3	9	15	27	40
4	10	17	30	40
5	10	17	25	40
6	8	18	24	35
7	9	19	26	35
8	9	23	30	40
9	7	20	27	35
10	8	19	32	40
11	6	14	25	30

6.2.3.2. Final Impact Data

Figure 6.5 and Figure 6.6 display the results of the adjusted cut points in terms of impact for ELA and mathematics, respectively. Table 6.12 includes the demographic data for students included in the impact data (i.e., the May 6 sample).

Figure 6.5

English Language Arts Impact Data Using Adjusted Cut Points

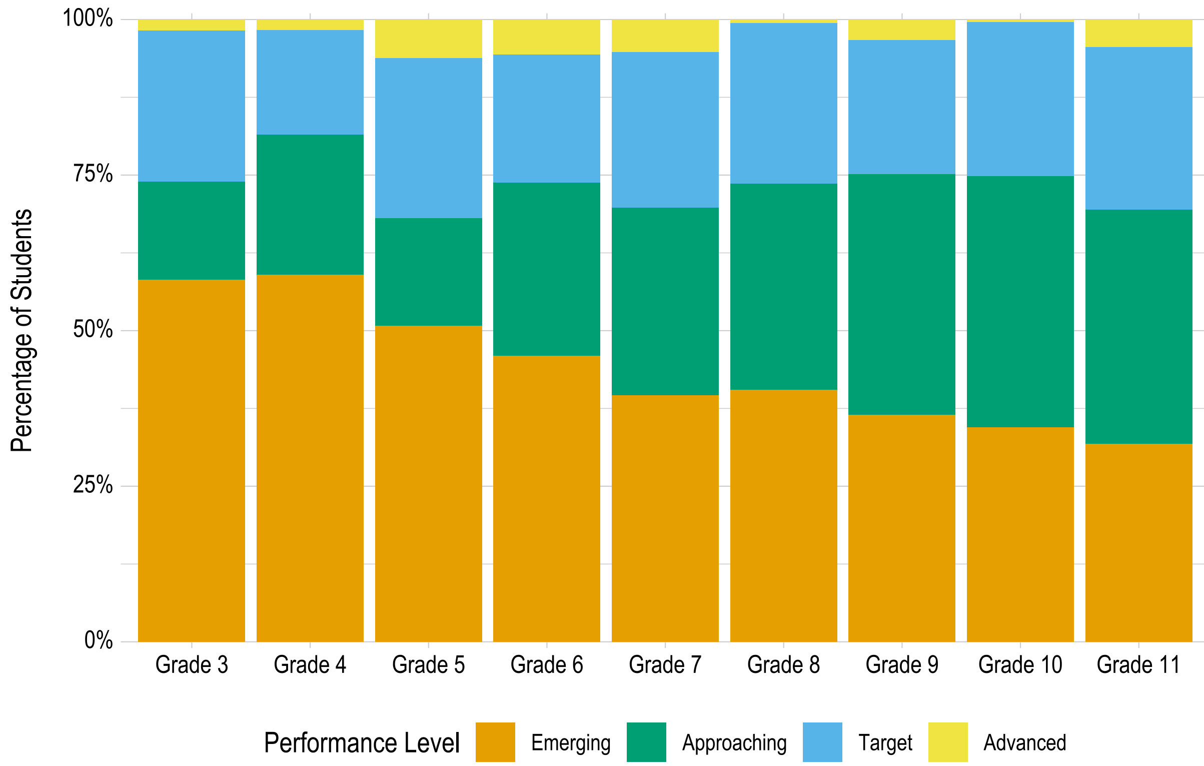


Figure 6.6

Mathematics Impact Data Using Adjusted Cut Points

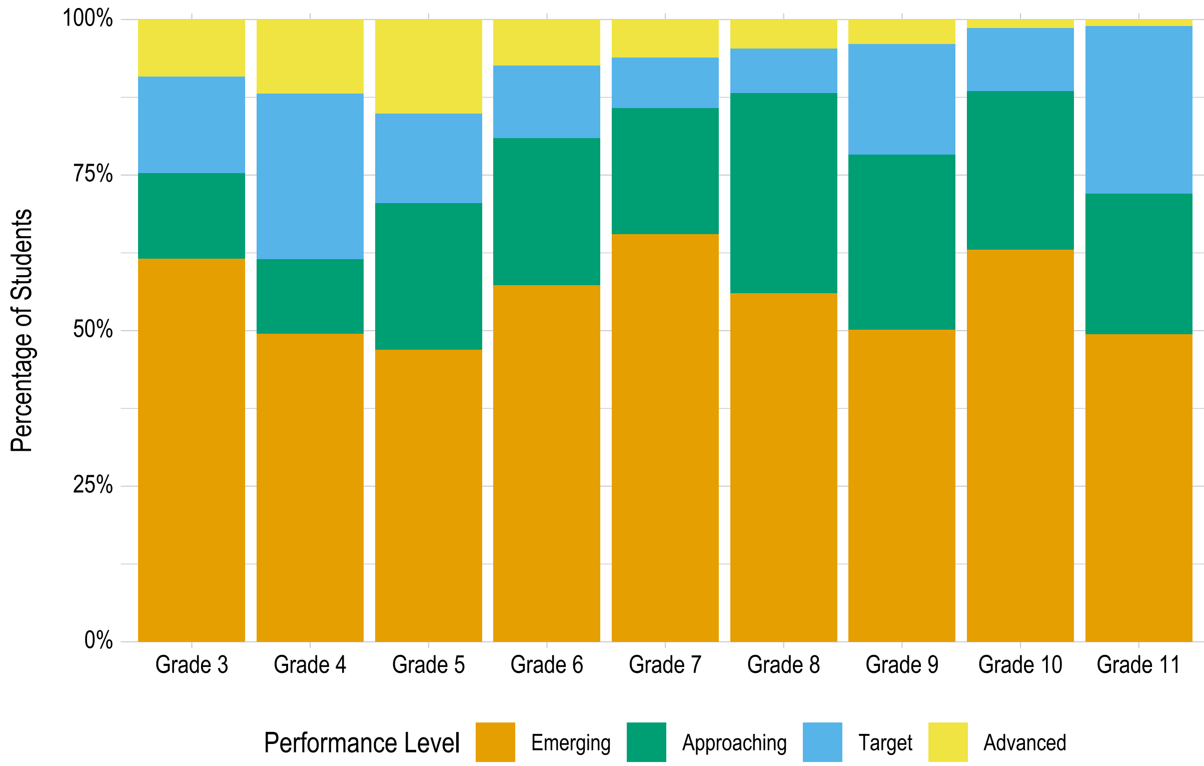


Table 6.12

*Demographic Information for Students Included in 2022 Impact Data (*N *= 65,294)*

Subgroup	*n*	%
Gender		
Male	44,029	67.4
Female	21,201	32.5
Nonbinary/undesignated	64	0.1
Race		
White	37,104	56.8
African American	12,958	19.8
Two or more races	9,399	14.4
Asian	3,569	5.5
American Indian	1,799	2.8
Native Hawaiian or Pacific Islander	301	0.5
Alaska Native	164	0.3
Hispanic ethnicity		
Non-Hispanic	49,246	75.4

Table 6.12

Demographic Information for Students Included in 2022 Impact Data (N = 65,294) (continued)

Subgroup	*n*	%
Hispanic	16,048	24.6
English learning (EL) participation		
Not EL eligible or monitored	61,440	94.1
EL eligible or monitored	3,854	5.9
English language arts complexity band		
Foundational	7,934	12.2
Band 1	23,354	35.8
Band 2	23,620	36.2
Band 3	7,287	11.2
Missing*	3,099	4.7
Mathematics complexity band		
Foundational	8,328	12.8
Band 1	23,446	35.9
Band 2	26,103	40.0
Band 3	6,221	9.5
Missing*	1,196	1.8

* Students with missing complexity band information only completed assignments in one subject by May 6, 2022 (i.e., students with a missing English language arts complexity band only completed their mathematics assessment).

6.2.3.3. Mastery Profile Development

To preserve the original content-based standard setting method and to aid in the evaluation of adjusted cut points, example mastery profiles were created for each of the cut points in Table 6.11. Because the original standard setting method and the adjustment process were based on mastery profiles, there were many different ways a student could master a specified number of linkage levels in a subject. DLM staff identified the most common profiles using data from spring 2018, spring 2019, and spring 2021. When evaluating data from 2018 and 2019, before the blueprint revisions were implemented, DLM staff included only results from EEs retained in the blueprint revision when calculating the total linkage levels mastered and building the mastery profiles. After building the profiles for each year, DLM staff combined the data and identified the three most common mastery profiles across years for the total linkage levels mastered at each cut point. DLM staff then generated a visual representation of each selected profile for each of the adjusted cut point values (see Figure 6.7).

Figure 6.7

Example Mastery Profile for a Hypothetical Student

		Estimated Mastery Level				
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.6.2	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
ELA.C1.2	ELA.EE.RL.6.4	Identify descriptive features and words	Identify words with opposite meanings	Construct word meaning using context clues	Associate word choice with textual meaning	Use semantic clues to identify word meaning
ELA.C1.2	ELA.EE.RI.6.1	Differentiate between text and pictures	Identify illustrations for a familiar text	Identify explicit details in informational texts	Understand explicit and implicit information	Identify explicit information and the need to make implicit associations
ELA.C1.2	ELA.EE.RI.6.6	Identify familiar objects through property word descriptors	Identify concrete details in an informational text	Identify relationships between concrete details	Identify words or phrases that reveal the author's point of view	Identify the author's point of view and reason for writing the text
ELA.C1.2	ELA.EE.RI.6.8	Determine similar or different	Identify details related to a topic in a text	Identify details that defend a claim	Differentiate between evidenced and non-evidenced claims	Identify points that defend a claim

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

6.2.4. *External Evaluation of Standards Adjustment Process and Results*

The TAC provided feedback throughout 2019, 2020, 2021, and 2022 on the methodology, procedures, and results of the standards adjustment process. The adjusted cut points were presented to the TAC on May 11, 2022, along with an evaluation of the impact data and a comparison of the impact data to the reported impact data from spring 2019. Following this presentation, the TAC prepared and accepted a resolution endorsing the adequacy, quality of judgments, and extent to which the process met professional standards and recommended the adoption of the adjusted cut points. The complete TAC recommendation can be found in Appendix D.1.

Following the May 11, 2022, meeting of the TAC, a set of materials was provided to the DLM Governance Board for review. Materials included the adjusted cut points, the set of example mastery profiles previously described, a summary of the adjustment methodology, and the TAC resolution. Following the distribution of materials on May 11, 2022, state education staff conducted their own internal reviews of the adjusted cut points and example mastery profiles. This review varied by state, but included the opportunity for states to collect additional feedback from subject matter experts in their department, educators, and other stakeholders. Additionally, a governance board check-in call was hosted on May 17, 2022, for state education agency staff to ask questions about the adjustment process and results. Discussion during this call primarily focused on the representativeness of the May 6 sample and the implementation of the ±2 adjustment to match the 2019 impact data.

Following the check-in call and states' internal review of materials, the governance board conducted a vote on the adjusted cut points on May 24, 2022. Of the 14 states participating in the Year-End model, 4 were absent during the voting, resulting in 10 voting members who were present. The voting process consisted of three rounds of voting. In the first round of voting, the governance board had the option to proceed or delay the final vote to allow for more time to review the standards adjustment materials. All participating states (n = 10) voted to proceed with the final vote. In the second round of voting, governance board members had a nonbinding vote to adopt the adjusted cut points. In this vote, two states abstained, and all other participating states (n = 8) voted to adopt the adjusted cut points. Following the second round, governance board members had a final opportunity for discussion and to recommend any changes to the proposed cut points. No recommendations were made, so the governance board proceeded to the third and final round of voting, which was to officially adopt the adjusted cut points. In this vote, three states abstained and the remaining states (n = 7) voted unanimously to adopt the adjusted cut points as the final consortium cut points with no further adjustment.

6.2.5. Post-Hoc Analysis

Following the close of the 2022 spring assessment window on June 10, 2022, DLM staff conducted post-hoc analyses to verify the adjustments made using the sample of students who completed their assessment by May 6, 2022. DLM staff first compared the May 6 sample used for the adjustment to the full sample of students who completed DLM assessments in spring 2022 to verify that the May 6 sample was representative of the DLM student population as the initial analyses indicated. DLM staff then conducted the same ±2 adjustment to compare the cut points that were achieved from the May 6 sample to those that would have been achieved if the full sample were utilized.

Table 6.13 shows the number of students who completed their assessment by May 6 (from Table 6.8 above) and the number of students who completed their assessment during the entire 2022 spring assessment window. Across all grades and subjects, the May 6 sample represented 75% of all completed assessments, ranging from 72% in Grade 7 ELA to 96% in Grade 10 mathematics. Thus, the May 6 sample represented an even larger percentage of the full sample than the 70% that was initially estimated.

Table 6.13

Comparison of Sample Sizes for May 6 and Full Samples

Grade	Completed by May 6	Total completions	% of total
English language arts			
3	7,754	10,712	72.4
4	8,070	11,051	73.0
5	8,103	11,090	73.1
6	8,169	11,279	72.4
7	8,341	11,640	71.7
8	8,535	11,694	73.0
9	4,081	5,630	72.5
10	2,613	2,729	95.7
11	6,529	8,296	78.7
Mathematics			
3	7,757	10,684	72.6
4	8,082	11,027	73.3
5	8,165	11,080	73.7
6	8,639	11,327	76.3
7	8,796	11,711	75.1
8	8,702	11,713	74.3
9	4,303	5,676	75.8
10	2,646	2,744	96.4
11	7,008	8,399	83.4

DLM staff then compared the demographic distribution of the May 6 sample to the full sample from the complete spring assessment window. As in the initial analyses, DLM staff calculated Cramér's V effect sizes for the differences in each subgroup between samples. Table 6.14 shows the effect sizes when comparing the May 6 sample to the sample of all students who completed DLM assessments in spring 2022. As indicated in the initial analyses, the May 6 sample was highly consistent with the full sample. All effect sizes were in the negligible range, and the largest effect size was only .079. Thus, the evidence from the full spring assessment window supports the claim that the May 6 sample was representative of the full DLM student population in 2022.

Table 6.14

Cramér's V Effect Sizes for Demographic Comparisons Between the May 6 and Full Samples

Grade	Gender	Race	Hispanic ethnicity	English language learner status	Complexity band
English language arts					
3	.007	.023	.005	.027	.012
4	.010	.027	.010	.045	.013
5	.008	.030	.005	.034	.013
6	.006	.024	.013	.033	.018
7	.009	.027	.010	.042	.019
8	.006	.024	.013	.040	.013
9	.011	.036	.020	.079	.023
10	.003	.009	.003	.002	.004
11	.009	.019	.014	.002	.011
Mathematics					
3	.007	.024	.002	.027	.029
4	.008	.026	.010	.048	.021
5	.004	.029	.004	.034	.021
6	.005	.021	.010	.029	.020
7	.008	.022	.013	.042	.013
8	.006	.021	.016	.038	.022
9	.013	.031	.007	.079	.017
10	.004	.007	.001	.001	.003
11	.004	.015	.013	.002	.014

6.2.5.1. Cut Point Comparison

Following the close of the spring assessment window, DLM staff conducted the same ±2 adjustment process that was enacted with the May 6 sample to investigate whether the use of the May 6 sample resulted in different cut points than would have been achieved with the full DLM student population. When using the data from the full spring assessment window, the adjustment process resulted in the same cut points that were achieved with the May 6 sample (see Table 6.11). Thus, the limited nature of the May 6 sample had no effect on the adjustment process, as the cut points achieved with the May 6 sample are identical to what would have been achieved with the full set of student data. Thus, in totality, the post-hoc analyses support the process and use of the May 6 sample to conduct the final step of the standards adjustment process.

6.2.6. Updated Grade and Subject Performance Level Descriptors

Following the cut point adjustments, the test development team updated the grade and subject PLDs so that they matched the revised blueprint and reflected the new cut points.

The process was conducted in two phases. Phase 1 analyzed the original grade- and subject-specific PLDs against the revised blueprint to bring the PLDs into alignment with the content in the revised blueprint. Specifically, statements related to content for EEs that were not retained on the blueprint, and therefore were no longer relevant, were removed. In Phase 2, the adjusted cut points and example mastery profiles were used to verify the content of the updated PLDs. Test development specialists analyzed the example mastery profiles around the cut scores for each grade, subject, and performance level. When necessary, statements were moved into the skill ranges reflected in the example profiles. The newly revised PLDs include the knowledge, skills, and understandings that students typically demonstrate at performance levels defined by the adjusted cut points.

The revised PLDs were reviewed internally for content accuracy and were shared with the DLM Governance Board on June 14, 2022. Appendix D.2 contains examples of grade level and content PLDs, and all PLDs for ELA[47] and mathematics[48] are available on the DLM website.

6.3. Conclusion

In summary, the performance levels for DLM assessments are determined by applying cut points to the total number of linkage levels mastered within each subject. The cut points were originally developed by experienced panelists evaluating mastery profiles that summarize the skills and understandings that a student mastered in each subject. Thus, the resulting performance levels are based on the most common profiles of skill mastery that align to the policy PLDs adopted by the DLM Governance Board in 2015. In 2022, the cut points were adjusted to reflect the revised assessment blueprints. This adjustment process resulted in cut points that accounted for the reduced number of EEs on the blueprint, while recognizing that the overall construct was stable. Finally, grade- and subject-specific PLDs that describe the skills most commonly mastered by students who achieve at each performance level were developed based on the content of the EEs and the cut points derived from the standard setting and adjustment process.

[47] https://dynamiclearningmaps.org/sites/default/files/documents/ELA_PLDs_YE.pdf
[48] https://dynamiclearningmaps.org/sites/default/files/documents/Math_PLDs_YE.pdf

7. Reporting and Results

Scoring of Dynamic Learning Maps® (DLM®) assessments is based on linkage levels mastered across all assessed Essential Elements (EEs). Performance levels describe overall achievement in each subject based on the total linkage levels mastered, with students achieving At Target or Advanced demonstrating proficiency. This chapter summarizes assessment results for the DLM Alternate Assessment System and presents evidence of their appropriateness.

The chapter begins by presenting the number of students who participated in the assessment in 2021–2022, followed by student performance data. Performance data are described in terms of the percentage of students achieving at each performance level both overall and by student group. The chapter then describes how students demonstrated linkage level mastery from among three scoring rules as well as the distribution of the highest linkage level mastered by each student and preliminary evidence about how educators define skill mastery. Because writing testlets are scored by the test administrator, evidence of interrater reliability of writing sample scoring is evaluated. The chapter then summarizes evidence that students who meet the DLM alternate achievement standards are prepared to pursue postsecondary opportunities. Finally, the chapter describes the data files, score reports, guides for interpreting score reports, and quality-control procedures for data files and score reports.

7.1. Student Participation

During spring 2022, assessments were administered to 88,404 students in 14 states. Counts of students tested in each state are displayed in Table 7.1. The assessments were administered by 23,210 educators in 12,150 schools and 3,844 school districts. A total of 1,423,995 test sessions were administered during the spring assessment window. One test session is one testlet taken by one student. Only test sessions that were complete at the close of the spring assessment window counted toward the total sessions.

Table 7.1

*Student Participation by State (*N *= 88,404)*

State	Students (*n*)
Alaska	429
Colorado	4,182
Illinois	13,917
Maryland	5,030
New Hampshire	630
New Jersey	11,266
New Mexico	2,035
New York	18,219
Oklahoma	5,292
Pennsylvania	16,650
Rhode Island	905
Utah	3,944
West Virginia	1,445
Wisconsin	4,460

Table 7.2 summarizes the number of students assessed in each grade. In grades 3–8, over 11,140 students participated in each grade. In high school, the largest number of students participated in grade 11, and the smallest number participated in grade 12. The differences in high school grade-level participation can be traced to differing state-level policies about the grade(s) in which students are assessed.

Table 7.2

*Student Participation by Grade (*N *= 88,404)*

Grade	Students (*n*)
3	11,145
4	11,435
5	11,528
6	11,699
7	12,101
8	12,120
9	5,921
10	2,907
11	8,720
12	828

Table 7.3 summarizes the demographic characteristics of the students who participated in the spring 2022 administration. The distribution of students across the different subgroups was fairly consistent with prior

years' distributions. The majority of participants were male (68%) and white (56%). About 7% of students were monitored or eligible for English learning services.

Table 7.3

*Demographic Characteristics of Participants (*N *= 88,404)*

Subgroup	*n*	%
Gender		
Male	59,978	67.8
Female	28,349	32.1
Nonbinary/undesignated	77	0.1
Race		
White	49,713	56.2
African American	18,967	21.5
Two or more races	11,719	13.3
Asian	5,055	5.7
American Indian	2,351	2.7
Native Hawaiian or Pacific Islander	421	0.5
Alaska Native	178	0.2
Hispanic ethnicity		
Non-Hispanic	66,401	75.1
Hispanic	22,003	24.9
English learning (EL) participation		
Not EL eligible or monitored	82,497	93.3
EL eligible or monitored	5,907	6.7

In addition to the spring assessment window, instructionally embedded assessments are also made available for educators to optionally administer to students during the year. Results from the instructionally embedded assessments do not contribute to final summative scoring but can be used to guide instructional decision-making. Table 7.4 summarizes the number of students who completed at least one instructionally embedded assessment by state. A total of 2,418 students in 8 states took at least one instructionally embedded testlet during the 2021–2022 academic year.

Table 7.4

Students Completing Instructionally Embedded Testlets by State (N = 2,418)

State	*n*
Colorado	88
Maryland	1,358
New Jersey	583
New Mexico	7
New York	74
Oklahoma	295
Utah	12
West Virginia	1

Note. Maryland required administration of instructionally embedded assessments during fall 2021.

Table 7.5 summarizes the number of instructionally embedded testlets taken in ELA and mathematics. Across all states, students took 19,171 ELA testlets and 19,069 mathematics testlets during the instructionally embedded window.

Table 7.5

Number of Instructionally Embedded Testlets by Grade and Subject (N = 38,240)

Grade	English language arts	Mathematics
3	951	791
4	1,346	1,308
5	1,589	1,512
6	1,577	1,236
7	2,287	1,970
8	3,775	3,341
9	308	435
10	62	75
11	5,175	6,231
12	2,101	2,170
Total	*19,171*	*19,069*

7.2. Student Performance

Student performance on DLM assessments is interpreted using cut points,[49] which describe student achievement using four performance levels. A student's performance level is determined based on the

[49] For a description of the standard setting process used to determine the cut points, see Chapter 6 of this manual.

total number of linkage levels mastered across the assessed EEs.

For the spring 2022 administration, student performance was reported using four performance levels:

- The student demonstrates *Emerging* understanding of and ability to apply content knowledge and skills represented by the EEs.
- The student's understanding of and ability to apply targeted content knowledge and skills represented by the EEs is *Approaching the Target*.
- The student's understanding of and ability to apply content knowledge and skills represented by the EEs is *At Target*. This performance level is considered to be meeting achievement expectations.
- The student demonstrates *Advanced* understanding of and ability to apply targeted content knowledge and skills represented by the EEs.

7.2.1. Overall Performance

Table 7.6 reports the percentage of students achieving at each performance level from the spring 2022 administration for ELA and mathematics. In ELA, the percentage of students who achieved at the At Target or Advanced levels (i.e., proficient) ranged from approximately 18% to 30%. In mathematics, the percentage of students meeting or exceeding At Target expectations ranged from approximately 11% to 37%.

Table 7.6

Percentage of Students by Grade and Performance Level

Grade	Emerging (%)	Approaching (%)	At Target (%)	Advanced (%)	At Target + Advanced (%)
English language arts					
3 (*n* = 11,134)	60.6	15.2	22.5	1.8	24.2
4 (*n* = 11,425)	60.7	21.5	16.1	1.7	17.8
5 (*n* = 11,510)	52.8	17.1	24.4	5.7	30.1
6 (*n* = 11,687)	48.8	26.4	19.7	5.2	24.8
7 (*n* = 12,080)	41.9	29.2	24.0	4.9	28.9
8 (*n* = 12,083)	43.1	31.4	25.0	0.5	25.5
9 (*n* = 5,907)	37.7	36.5	22.4	3.4	25.8
10 (*n* = 2,903)	36.9	38.9	23.7	0.4	24.1
11 (*n* = 8,681)	34.4	35.7	25.4	4.4	29.9
12 (*n* = 822)	42.1	34.2	20.3	3.4	23.7
Mathematics					
3 (*n* = 11,103)	63.7	12.8	14.5	9.0	23.5
4 (*n* = 11,400)	51.5	11.8	25.1	11.6	36.7
5 (*n* = 11,481)	49.3	22.6	14.0	14.2	28.2
6 (*n* = 11,653)	58.9	22.6	11.2	7.3	18.5
7 (*n* = 12,057)	66.6	19.5	7.9	6.0	14.0
8 (*n* = 12,086)	57.7	30.6	6.9	4.8	11.7
9 (*n* = 5,894)	50.4	27.6	17.8	4.2	22.0
10 (*n* = 2,899)	63.9	25.1	9.7	1.3	11.0
11 (*n* = 8,668)	50.9	22.2	26.0	0.9	27.0
12 (*n* = 811)	59.3	18.1	22.1	0.5	22.6

7.2.2. Subgroup Performance

Data collection for DLM assessments includes demographic data on gender, race, ethnicity, and English learning status. Table 7.7 and Table 7.8 summarize the disaggregated frequency distributions for ELA and mathematics, respectively, collapsed across all assessed grade levels. Although state education agencies each have their own rules for minimum student counts needed to support public reporting of results, small counts are not suppressed here because results are aggregated across states and individual students cannot be identified.

Table 7.7

ELA Performance Level Distributions by Demographic Subgroup (N = 88,232)

	Emerging		Approaching		At Target		Advanced		At Target + Advanced	
Subgroup	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Gender										
Male	28,843	48.2	15,590	26.0	13,417	22.4	2,011	3.4	15,428	25.8
Female	13,480	47.6	7,583	26.8	6,316	22.3	915	3.2	7,231	25.6
Nonbinary/undesignated	40	51.9	17	22.1	17	22.1	3	3.9	20	26.0
Race										
White	23,433	47.2	12,859	25.9	11,549	23.3	1,773	3.6	13,322	26.9
African American	8,856	46.8	5,278	27.9	4,195	22.2	598	3.2	4,793	25.3
Two or more races	5,890	50.3	3,098	26.5	2,372	20.3	339	2.9	2,711	23.2
Asian	2,891	57.3	1,134	22.5	886	17.5	138	2.7	1,024	20.3
American Indian	966	41.1	686	29.2	626	26.6	71	3.0	697	29.7
Native Hawaiian or Pacific Islander	209	50.1	99	23.7	100	24.0	9	2.2	109	26.1
Alaska Native	118	66.7	36	20.3	22	12.4	1	0.6	23	13.0
Hispanic ethnicity										
Non-Hispanic	31,711	47.8	17,307	26.1	15,080	22.8	2,178	3.3	17,258	26.0
Hispanic	10,652	48.5	5,883	26.8	4,670	21.3	751	3.4	5,421	24.7
English learning (EL) participation										
Not EL eligible or monitored	39,519	48.0	21,513	26.1	18,509	22.5	2,790	3.4	21,299	25.9
EL eligible or monitored	2,844	48.2	1,677	28.4	1,241	21.0	139	2.4	1,380	23.4

Table 7.8

*Mathematics Performance Level Distributions by Demographic Subgroup (*N *= 88,052)*

	Emerging		Approaching		At Target		Advanced		At Target + Advanced	
Subgroup	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Gender										
Male	33,317	55.8	12,291	20.6	9,122	15.3	4,980	8.3	14,102	23.6
Female	16,801	59.4	6,164	21.8	3,797	13.4	1,503	5.3	5,300	18.8
Nonbinary/undesignated	56	72.7	12	15.6	8	10.4	1	1.3	9	11.7
Race										
White	27,964	56.4	10,484	21.2	7,442	15.0	3,649	7.4	11,091	22.4
African American	10,647	56.4	4,044	21.4	2,739	14.5	1,455	7.7	4,194	22.2
Two or more races	6,924	59.3	2,445	21.0	1,641	14.1	658	5.6	2,299	19.7
Asian	3,124	62.3	815	16.2	641	12.8	438	8.7	1,079	21.5
American Indian	1,134	48.4	561	23.9	396	16.9	253	10.8	649	27.7
Native Hawaiian or Pacific Islander	245	58.3	87	20.7	59	14.0	29	6.9	88	21.0
Alaska Native	136	76.4	31	17.4	9	5.1	2	1.1	11	6.2
Hispanic ethnicity										
Non-Hispanic	37,750	57.1	13,976	21.1	9,701	14.7	4,714	7.1	14,415	21.8
Hispanic	12,424	56.7	4,491	20.5	3,226	14.7	1,770	8.1	4,996	22.8
English learning (EL) participation										
Not EL eligible or monitored	46,958	57.2	17,224	21.0	12,035	14.6	5,945	7.2	17,980	21.9
EL eligible or monitored	3,216	54.6	1,243	21.1	892	15.1	539	9.2	1,431	24.3

7.3. Mastery Results

As described above, the student performance levels are determined by applying cut points to the total number of linkage levels mastered in each subject. In this section, we summarize student mastery of assessed EEs and linkage levels, including how students demonstrated mastery from among three scoring rules and the highest linkage level students tended to master.

7.3.1. Mastery Status Assignment

As described in Chapter 5 of this manual, student responses to assessment items are used to estimate the posterior probability that the student mastered each of the assessed linkage levels using diagnostic classification modeling. Students with a posterior probability of mastery greater than or equal to .80 are assigned a linkage level mastery status of 1, or mastered. Students with a posterior probability of mastery less than .80 are assigned a linkage level mastery status of 0, or not mastered. Maximum uncertainty in the mastery status occurs when the probability is .5 and maximum certainty when the probability approaches 0 or 1. After considering the risk of false positives and negatives and preliminary data analyses, and based on input from the DLM Technical Advisory Committee (TAC), the threshold used to determine mastery classifications was set at .80. In addition to the calculated probability of mastery, students could be assigned mastery of linkage levels within an EE in two other ways: correctly answering 80% of all items administered at the linkage level or through the *two-down* scoring rule. The two-down scoring rule was implemented to guard against students assessed at the highest linkage levels being overly penalized for incorrect responses. When a student did not demonstrate mastery of the assessed linkage level, mastery was assigned at two linkage levels below the level that was assessed.

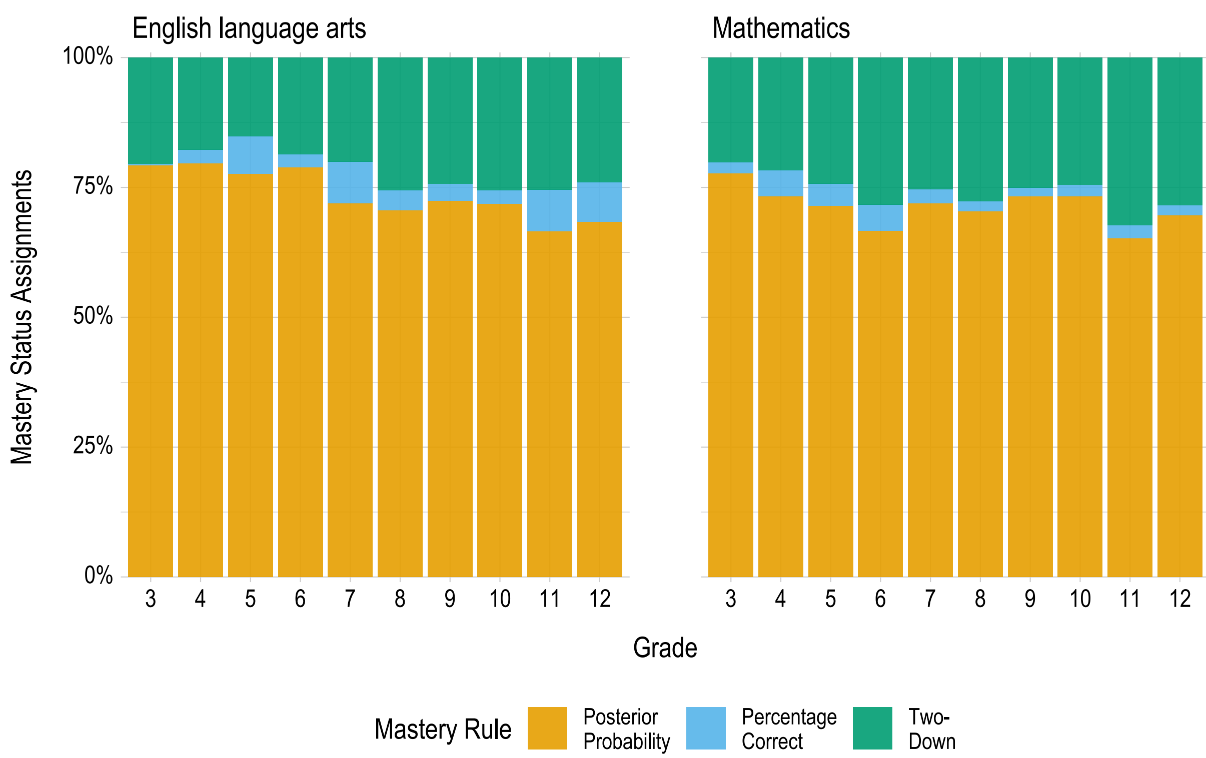
Take, for example, a student who tested only on the Target linkage level of an EE. If the student demonstrated mastery of the Target linkage level, as defined by the .80 posterior probability of mastery cutoff or the 80% correct rule, then all linkage levels below and including the Target level would be categorized as mastered. If the student did not demonstrate mastery on the tested Target linkage level, then mastery would be assigned at two linkage levels below the tested linkage level (i.e., the Distal Precursor), rather than showing no evidence of mastery at all. The two-down rule is based on linkage level ordering evidence and the underlying learning map structure (see Chapter 2 of this manual), and it is designed to not penalize students who test at higher linkage levels. That is, in the example above, the student is assigned mastery of the Distal Precursor, rather than being assigned no mastery for the EE.

To evaluate the degree to which each mastery assignment rule contributed to students' linkage level mastery status during the 2021–2022 administration of DLM assessments, the percentage of mastery statuses obtained by each scoring rule was calculated, as shown in Figure 7.1. Posterior probability was given first priority. That is, if multiple scoring rules agreed on the highest linkage level mastered within an EE (e.g., the posterior probability and 80% correct both indicate the Target linkage level as the highest mastered), the mastery status was counted as obtained via the posterior probability. If mastery was not demonstrated by meeting the posterior probability threshold, the 80% scoring rule was imposed, followed by the two-down rule. This means that EEs that were assessed by a student at the lowest two linkage levels (i.e., Initial Precursor and Distal Precursor) are never categorized as having mastery assigned by the two-down rule. This is because the student would either master the assessed linkage level and have the EE counted under the posterior probability or 80% correct scoring rule, or all three scoring rules would agree on the score (i.e., no evidence of mastery), in which case preference is given to the posterior probability. Across grades and subjects, approximately 65%–80% of mastered linkage levels were derived from the posterior probability obtained from the modeling procedure. Approximately <1%–8% of linkage levels were assigned mastery status by the percentage correct rule. The remaining approximately 15%–32% of mastered linkage levels were determined by the minimum mastery, or two-down rule.

Because correct responses to all items measuring the linkage level are often necessary to achieve a posterior probability above the .80 threshold, the percentage correct rule overlapped considerably (but was second in priority) with the posterior probabilities. The percentage correct rule did, however, provide mastery status in those instances where correctly responding to all or most items still resulted in a posterior probability below the mastery threshold. The agreement between these two methods was quantified by examining the rate of agreement between the highest linkage level mastered for each EE for each student. For the 2021–2022 operational year, the rate of agreement between the two methods was 87%. However, in instances in which the two methods disagreed, the posterior probability method indicated a higher level of mastery (and therefore was implemented for scoring) in 68% of cases. Thus, in some instances, the posterior probabilities allowed students to demonstrate mastery when the percentage correct was lower than 80% (e.g., a student completed a four-item testlet and answered three of four items correctly).

Figure 7.1

Linkage Level Mastery Assignment by Mastery Rule for Each Subject and Grade



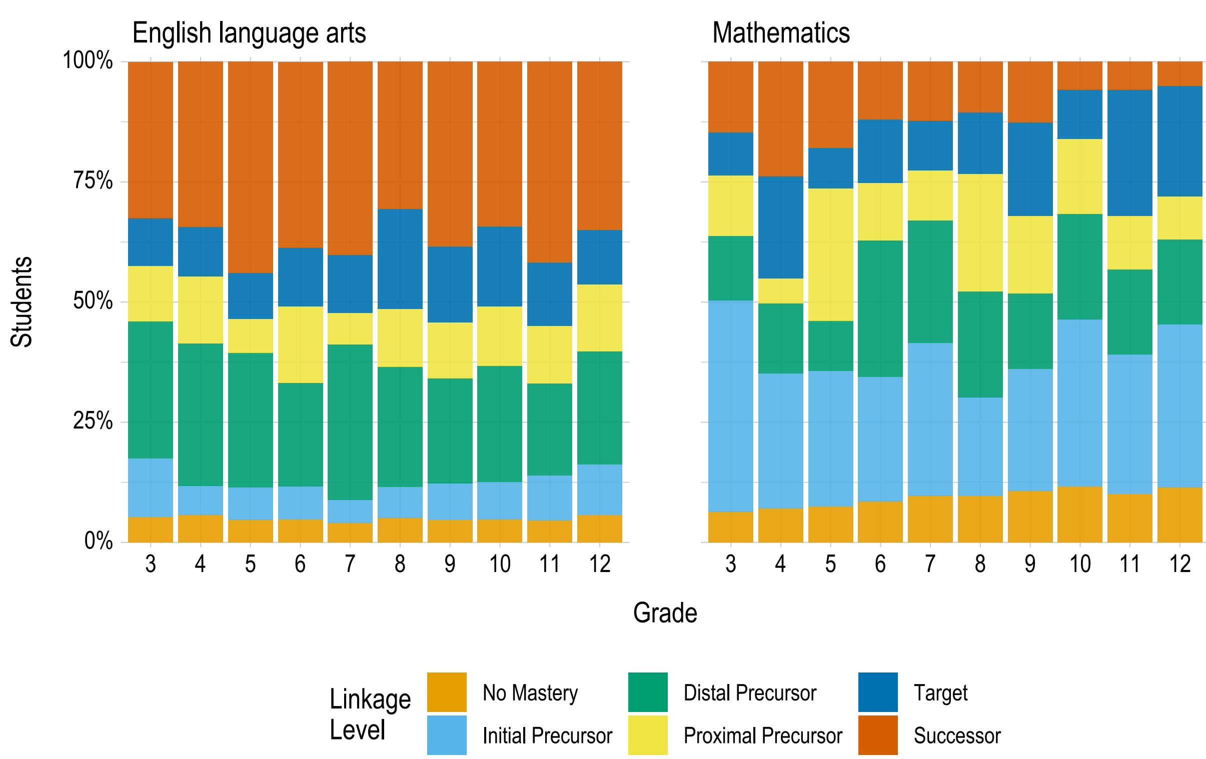
7.3.2. Linkage Level Mastery

Scoring for DLM assessments determines the highest linkage level mastered for each EE. The linkage levels are (in order): Initial Precursor, Distal Precursor, Proximal Precursor, Target, and Successor. A student can be a master of zero, one, two, three, four, or all five linkage levels, within the order constraints. For example, if a student masters the Proximal Precursor level, they also master all linkage levels lower in the order (i.e., Initial Precursor and Distal Precursor). This section summarizes the distribution of students by highest linkage level mastered across all EEs. For each student, the highest linkage level mastered across all tested EEs was calculated. Then, for each grade and subject, the number of students with each linkage level as their highest mastered linkage level across all EEs was summed and then divided by the total number of students who tested in the grade and subject. This resulted in the proportion of students for whom each linkage level was the highest level mastered.

Figure 7.2 displays the percentage of students who mastered each linkage level as the highest linkage level across all assessed EEs for ELA and mathematics. For example, across all grade 3 mathematics EEs, the Initial Precursor level was the highest level that students mastered 44% of the time. The percentage of students who mastered the Target or Successor linkage level ranged from approximately 42% to 55% in ELA and ranged from approximately 16% to 45% in mathematics.

Figure 7.2

Students' Highest Linkage Level Mastered Across English Language Arts and Mathematics Essential Elements by Grade



7.4. Additional Scoring Evidence

This section describes additional scoring evidence for DLM assessments, including preliminary evidence from a pilot survey collecting information on how educators define skill mastery, interrater reliability of test administrator scoring of student writing samples, and evidence that students who achieve At Target on DLM assessments are on track to pursue postsecondary opportunities.

7.4.1. Pilot Survey Defining Student Mastery

In 2019, DLM staff piloted a survey to obtain information about how educators view skill mastery for students with significant cognitive disabilities who take DLM assessments. The survey was completed by 95 educators across six states and included an open-ended item asking educators to describe how they define student mastery. The survey responses revealed that many educators defined mastery as 75%–80% successful trials on the skills, which is common in special education, but other educators shared ideas around consistent skill demonstration over a longer period of time, generalization and transfer, independence and speed of demonstration, and students' ability to explain the concept (Nehler & Clark, 2019). While these definitions differ from how mastery is determined on DLM assessments, they provide some insight as to how educators may interpret mastery on DLM assessments.

7.4.2. Writing Sample Scoring

Testlets measuring ELA writing EEs require the student to work outside of Student Portal and interact with the test administrator. The test administrator directs the student to engage in writing tasks and then evaluates the student writing tasks in the online system. Scores do not result from a high-inference process common in large-scale assessment, such as applying analytic or holistic rubrics. Instead, writing samples are evaluated for text features that are easily perceptible to a fluent reader and require little or no inference on the part of the rater (e.g., correct syntax, orthography). The test administrator is presented with an onscreen selected-response item and is instructed to choose the option(s) that best matches the student's writing sample. Only test administrators rate writing samples, and their item responses are used to determine students' mastery of linkage levels for writing and some language EEs on the ELA blueprint. Because scores for the writing samples are determined by the test administrator, we annually collect a subset of writing samples and evaluate agreement of writing sample scoring to provide evidence for the quality of writing EE mastery decisions. In the subsequent sections, we describe the process for collecting and scoring samples and report interrater agreement among test administrators. For a complete description of writing testlet design, including example items, see Chapter 3 of this manual.

7.4.2.1. Preparing the Data

Educators administer writing testlets at two levels: emergent and conventional. Emergent testlets measure nodes at the Initial Precursor and Distal Precursor levels, while conventional testlets measure nodes at the Proximal Precursor, Target, and Successor levels. The number of items that evaluate the writing sample per grade-level testlet is summarized in Table 7.9.

Testlets include one to six items evaluating the sample, administered as either multiple-choice or multiple-select multiple-choice items. Because each answer option could correspond to a unique linkage level and/or EE, writing items are dichotomously scored at the option level. Each item, which includes four to nine answer options, was scored as a separate writing item. For this reason, writing items are referred to as writing tasks in the following sections, and the options were scored as individual items. The dichotomous option responses (i.e., each scored as an item) were the basis for the evaluation of interrater agreement.

Table 7.9

Number of Items That Evaluate the Writing Product per Testlet by Grade

Grade	Emergent testlet	Conventional testlet
3	*	3
4	1	4
5	*	2
6	*	4
7	1	4
8	*	4
9	1	8
10	1	8
11	1	8

Note. Items varied slightly by blueprint model; the maximum number of items per testlet is reported here.
[*] The testlet at this grade included only items evaluating the writing process, with no evaluation of the sample.

The scoring process for DLM writing testlets is as follows. Data are extracted from the database that houses all DLM data. For writing items, the response-option identifiers are treated as item identifiers so that each response option can be scored as correct or incorrect for the EE and linkage level it measures. Also, response-option dependencies are built in, based on scoring direction provided by the ELA test development team, to score as correct response options that are subsumed under other correct response options. Once the data structure has been transformed and response-option dependencies are accounted for, the writing data are combined with all other data to be included in the scoring process. For more information on the scoring process for linkage level mastery, see Section 7.3.2.

7.4.2.2. Recruitment

Recruitment for the evaluation of interrater agreement of writing samples included the submission of student writing samples and direct recruitment of educators to serve as raters.

7.4.2.2.1. Samples

Educators are annually asked to submit student writing samples. Requested submissions included papers that students used during testlet administration, copies of student writing samples, or photographs of student writing samples. To allow the sample to be matched with test administrator response data, each sample was submitted in the online system to enable direct matching to the administered testlet, including original test administrator responses.

In 2022, DLM staff evaluated interrater reliability for writing samples collected during the 2019–2020 and 2020–2021 administrations. A total of 1,914 student writing samples were submitted from districts in 12 states across the spring assessment windows in 2020 and 2021. In several grades, the emergent writing testlet does not include any tasks that evaluate the writing sample (as shown in Table 7.9); therefore, emergent samples submitted for these grades were not included in the interrater reliability analysis (e.g., Grade 3 emergent writing samples). To lessen the burden on the raters, we sampled approximately 20 writing samples from each grade and writing level (excluding the emergent levels in certain grades described above). This resulted in the assignment of 309 writing samples to raters for evaluation of interrater agreement.

7.4.2.2.2. Raters

Recruited educators were required to have experience administering and rating DLM writing testlets to ensure they had already completed required training and were familiar with how to score the writing samples. Participants were screened and selected using the Accessible Teaching, Learning, and Assessment Systems (ATLAS) recruitment database platform in MemberClicks. Volunteers complete a demographic survey to be considered for ATLAS events, and this information is stored in the database for recruitment purposes. In total, 18 were selected to participate.

Raters had a range of teaching experience, as indicated in Table 7.10. Most had taught ELA and/or students with the most significant cognitive disabilities for at least 6 years. Furthermore, eight raters (44%) reported experience as DLM external reviewers (see Chapter 3 for a description of external review).

Table 7.10

*Raters' Teaching Experience (*N *= 18)*

	1–5 years		6–10 years		> 10 years	
Teaching experience	*n*	%	*n*	%	*n*	%
English language arts	2	11.1	2	11.1	14	77.8
Students with significant cognitive disabilities	1	5.6	4	22.2	13	72.2

Demographic information for participants in this study is summarized in Table 7.11. Participating raters were mostly female (78%), white (78%), and non-Hispanic/Latino (89%). Raters came from a variety of teaching settings.

Table 7.11

*Raters' Demographic Information (*N *= 18)*

Subgroup	*n*	%
Gender		
Female	14	77.8
Male	4	22.2
Race		
White	14	77.8
Black or African American	2	11.1
American Indian	1	5.6
Asian	1	5.6
Hispanic ethnicity		
Non-Hispanic/Latino	16	88.9
Hispanic/Latino	2	11.1
Teaching setting		
Rural	7	38.9
Urban	6	33.3
Suburban	5	27.8

7.4.2.3. Sample Ratings

All ratings occurred during a virtual event in July 2022. Raters were provided with deidentified PDF versions of student writing samples via an online Moodle course, which they could no longer access following the completion of ratings. They were also provided a link to a Qualtrics survey that included the writing tasks corresponding to the grade and level (i.e., emerging or conventional) of the assigned writing sample. Raters submitted all ratings online.

Writing samples were assigned to raters in batches of 13–24, using a partially crossed matrix design to assign each sample to a total of three raters. Thus, educators rated between 60 and 83 writing samples. Table 7.12 summarizes the number of samples that were rated at each grade and level.

Table 7.12

Student Writing Samples With Ratings by Grade (N = 309)

Grade	Number of writing samples: Emergent	Number of writing samples: Conventional	Total number of samples
3	*	28	28
4	31	33	64
5	*	19	19
6	*	24	24
7	24	22	46
8	*	25	25
9	14	13	27
10	13	15	28
11	23	25	48
Total	*105*	*204*	*309*

* The testlet at this grade included only items evaluating the writing process, with no evaluation of the sample.

Ratings submitted in Qualtrics were combined with the original spring 2020 and spring 2021 student data from when the writing sample was rated by the student's educator, resulting in four ratings for each of the 309 student writing samples.

7.4.2.4. Interrater Reliability

Because each writing sample was evaluated by multiple and different raters, interrater reliability was summarized by Fleiss's kappa and intraclass correlation (ICC) values. The purpose of Fleiss's kappa is to provide a measure of absolute agreement across two or more raters. Fleiss's kappa (Fleiss, 1981) is defined as

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{7.1}$$

where the denominator gives the degree of absolute agreement attainable above chance, and the numerator gives the degree of absolute agreement actually achieved above chance.

The purpose of the ICC is to provide a means for measuring rater agreement and consistency. For interrater reliability studies, rater agreement is of most interest. For this study, a one-way, random-effects model using the average kappa rating was selected because each writing sample was rated by a rater who was randomly selected from the pool of available raters. Using this model, only absolute agreement is measured by the ICC.

To summarize global agreement across all student writing samples, educators' original ratings (from spring

2022 operational administration) were compared against the additional three ratings obtained during the July 2022 event. Interrater agreement results are presented in Table 7.13. Results are also provided separately for emergent and conventional testlets.

Based on the guidelines specified by Cicchetti (1994), ICC agreement fell in the *excellent* range ($\geq$ .75), and Fleiss's kappa fell in the *good* (.60 – .74) to *excellent* range. Agreement was slightly higher for conventional testlets.

Table 7.13

*Interrater Agreement for Writing Samples (*N *= 309)*

Group	*n*	ICC	ICC lower bound	ICC upper bound	Fleiss's κ
Overall	309	.967	.966	.969	.881
EW	105	.901	.887	.914	.695
CW	204	.969	.967	.970	.886

Note. ICC = intraclass correlation; EW = emergent writing; CW = conventional writing.

Educator-administered testlets measuring reading and mathematics were not included in the study. Also, although student writing samples were evaluated, the student writing process was not. Additional data collection related to educator fidelity, including fidelity in educator-administered testlets in each subject, is provided in the Test Administration Observations section of Chapter 4 of this manual.

7.4.3. Alignment of At Target Achievement With Postsecondary Opportunities

DLM academic achievement standards are intended to reflect rigorous expectations for students with the most significant cognitive disabilities. During 2021–2022, evidence was collected to evaluate the extent to which the DLM alternate academic achievement standards are aligned to ensure that a student who meets these standards is on track to pursue postsecondary education or competitive integrated employment.

Chapter 2 of this manual provides evidence of vertical alignment for the alternate academic achievement standards to reflect how academic content builds across grades to reflect high expectations for the student population. Here we describe the relationship of DLM alternate academic achievement standards (specifically the At Target performance level) to the knowledge, skills, and understandings needed to pursue postsecondary opportunities.

We developed two hypotheses about the expected relationship between meeting DLM alternate academic achievement standards and being prepared for a variety of postsecondary opportunities.

1. Nearly all academic skills will be associated with At Target performance level descriptors at a variety of grades between Grade 3 and high school. Few, if any, academic skills will first occur before Grade 3 At Target or after high school At Target.
2. Because academic skills may be associated with multiple opportunities and with soft skills needed for employment and education, we expected Hypothesis 1 to hold for academic skills associated with employment opportunities, education opportunities, and soft skills.

Similar to academic education for all students, academic skills for students with significant cognitive disabilities develops across grades. Individuals use academic skills at varying levels of complexity, depending on specific employment or postsecondary education settings. Therefore, academic skills associated with achieving At Target in lower grades demonstrate where students are able to apply the least complex version of the skill. Given the vertical alignment of DLM content and achievement standards, students are expected to continue learning new skills in subsequent grades and be prepared for more complex applications of the academic skills by the time they transition into postsecondary education and employment.

A panel of experts on secondary transition and/or education of students with significant cognitive disabilities identified postsecondary competitive integrated employment and education opportunities. Their goal was to identify an extensive sampling of opportunities rather than an exhaustive list. Panelists also considered the types of educational and employment opportunities currently available to students with significant cognitive disabilities as well as opportunities that may be more aspirational (i.e., opportunities that may become available in the future). Panelists identified 57 employment opportunities and 7 postsecondary education opportunities. Employment opportunities spanned sectors including agriculture, arts, business, education, health sciences, hospitality, information technology, manufacturing, and transportation.

Panelists next identified the knowledge, skills, and understandings needed to fulfill the responsibilities for the employment opportunities as well as eight common responsibilities across all postsecondary education opportunities. Finally, the panel identified the knowledge, skills, and understandings within soft skills (e.g., social skills, self-advocacy) applicable across multiple postsecondary settings. Subject-matter experts in ELA and mathematics reviewed and refined the academic skill statements to provide clarity and consistency across skills. This resulted in 50 ELA academic skills and 41 mathematics academic skills to be used in the next phase of the study.

The second set of panels, one for each subject, examined the relationship between the academic skills and the types of academic knowledge, skills, and understandings typically associated with meeting the DLM alternate academic achievement standards (i.e., achieving At Target). By identifying the lowest grade where a student achieving At Target is likely to consistently demonstrate the academic skill, the second panel identified the first point where students would be ready to pursue postsecondary opportunities that required the least complex application of the skill.

Panels consisted of general educators and special educators who administered DLM assessments from across DLM states. Most panelists had expertise across multiple grade bands, and some had certification in both an academic subject and special education. Panels completed training and calibration activities prior to making independent ratings. Panels discussed ratings until consensus when there was not an initial majority agreement.

Panels identified the lowest grade in which students who achieve At Target on the DLM alternate assessment are at least 80% likely to be able to demonstrate each skill, showing the first point of readiness to pursue postsecondary opportunities that require the least complex application of academic skills. In ELA, students achieving At Target are expected to first demonstrate 66% of those skills by Grade 5 and 27% in middle grades (Grades 6–8). In mathematics, students meeting achievement standards are expected to first demonstrate 81% of the academic skills by Grade 5 and 19% of skills in middle grades.

Overall, findings from panels indicate that most academic skills needed to pursue postsecondary opportunities are first associated with meeting the DLM academic achievement expectations in earlier grades (i.e., 3–5). Given the vertical alignment of the DLM academic achievement standards, students who achieve At Target in early grades further develop these skills so that, by the time they leave high school, they are ready to pursue postsecondary opportunities that require more complex applications of the academic skills.

Panelists also participated in focus groups to share their perceptions of opportunities, skills, and expectations for students with significant cognitive disabilities. Panelists believed the academic skills were important to postsecondary opportunities for all students, not only those who take the DLM assessment. Panelists indicated that students who were At Target in high school on the DLM assessment were likely to possess the needed academic knowledge, skills, and understandings to pursue a range of postsecondary opportunities.

Evaluation of panelists' experiences from both panels and DLM TAC members' review of the processes and evaluation of results provide evidence that the methods and processes used achieved the goals of the study. See Karvonen et al. (2022) for the full version of the postsecondary opportunities technical report. Overall, this evidence shows that students who demonstrate proficiency on DLM assessments are on track to pursue postsecondary opportunities.

7.5. Data Files

DLM assessment results were made available to DLM state education agencies following the spring 2022 administration. The General Research File (GRF) contained student results, including each student's highest linkage level mastered for each EE and final performance level for the subject for all students who completed any testlets. In addition to the GRF, the states received several supplemental files. The special circumstances file provided information about which students and EEs were affected by extenuating circumstances (e.g., chronic absences), as defined by each state. State education agencies also received a supplemental file to identify exited students. The exited students file included all students who exited at any point during the academic year. In the event of observed incidents during assessment delivery, state education agencies are provided with an incident file describing students impacted; however, no incidents occurred during 2021–2022.

State education agencies were provided with a 2-week window following data file delivery to review the files and invalidate student records in the GRF. Decisions about whether to invalidate student records are informed by individual state policy. If changes were made to the GRF, state education agencies submitted final GRFs via Educator Portal. The final GRF was used to generate score reports.

7.6. Score Reports

Assessment results were provided to state education agencies to report to parents/guardians, educators, and local education agencies. Individual Student Score Reports summarized student performance on the assessment by subject. Several aggregated reports were provided to state and local education agencies, including reports for the classroom, school, district, and state.

7.6.1. *Individual Student Score Reports*

Individual Student Score Reports were developed through a series of focus groups, including a set conducted in partnership with The Arc, a community-based organization advocating for and serving people with intellectual and developmental disabilities and their families. First, several groups focused on parent/guardian perceptions of existing alternate assessment results and score reports (Nitsch, 2013). These findings informed the development of prototype DLM score reports. Prototypes were reviewed by the DLM Governance Board and revised based on multiple rounds of feedback. Refined prototypes were shared with parents/guardians, advocates, and educators through additional focus groups (Clark et al., 2015) before finalizing the 2015 reports.

Individual Student Score Reports are comprised of two parts: (1) the Performance Profile, which aggregates linkage level mastery information for reporting performance in the subject overall and on each conceptual area, and (2) the Learning Profile, which reports specific linkage levels mastered for each assessed EE. There is one Individual Student Score Report per student per subject. Figure 7.3 and Figure 7.4 display a sample page from the Performance Profile and Learning Profile for ELA, respectively.

Figure 7.3

Sample Page of the Performance Profile for ELA

REPORT DATE: 11-15-2022
SUBJECT: English language arts
GRADE: 7

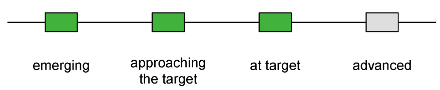
Individual Student End-of-Year Report
Performance Profile 2021-2022

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

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

Overall Results

Grade 7 English language arts allows students to show their achievement in 65 skills related to 13 Essential Elements. Student has mastered 37 of those 65 skills during Spring 2022. Overall, Student's mastery of English language arts fell into the third of four performance categories: **at target**. The specific skills Student has and has not mastered can be found in Student's Learning Profile.

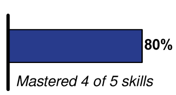


EMERGING:	The student demonstrates **emerging** understanding of and ability to apply content knowledge and skills represented by the Essential Elements.
APPROACHING THE TARGET:	The student's understanding of and ability to apply targeted content knowledge and skills represented by the Essential Elements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

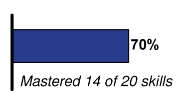
Area

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.

ELA.C1.1: Determine Critical Elements of Text



ELA.C1.2: Construct Understandings of Text



Page 1 of 5

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

© The University of Kansas. All rights reserved. For educational purposes only. May not be used for commercial or other purposes without permission. "Dynamic Learning Maps" is a trademark of The University of Kansas.

Figure 7.4

Sample Page of the Learning Profile for ELA

REPORT DATE: 11-15-2022
SUBJECT: English language arts
GRADE: 7

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

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

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

Student's performance in 7th grade English language arts Essential Elements is summarized below. This information is based on all of the DLM tests Student took during Spring 2022. Student was assessed on 13 out of 13 Essential Elements and 4 out of 4 Areas expected in 7th grade.

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.

Area	Essential Element	Estimated Mastery Level 1	2	3	4 (Target)	5
ELA.C1.1	ELA.EE.RI.7.5	Understand the functions of objects	Identify concrete details in an informational text	Recognize how titles reflect text structure and text purpose	Understand sequencing	Understand how parts of the text affect overall text structure
ELA.C1.2	ELA.EE.RL.7.1	Differentiate between text and pictures	Identify characters, setting, and major events	Identify words that answer explicit questions	Identify where explicit information is stated and where inferences can be drawn	Identify explicit and implicit information
ELA.C1.2	ELA.EE.RL.7.4	Understand words for absent objects and people	Identify definition of words explicitly defined in a sentence	Identify word meaning of multiple-meaning words using context clues	Determine the meaning of idioms and figures of speech	Determine the connotative meaning of words and phrases
ELA.C1.2	ELA.EE.RI.7.2	Match a picture representation with a real object	Identify concrete details in an informational text	Identify the implicit main idea in an informational text	Identify multiple main ideas in an informational text	Summarize a familiar informative text

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

This report is intended to serve as one source of evidence in an instructional planning process. Results are based only on item responses from the end of year spring assessment. 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.

© The University of Kansas. All rights reserved. For educational purposes only. May not be used for commercial or other purposes without permission. "Dynamic Learning Maps" is a trademark of The University of Kansas.

Page 3 of 5

Additional interviews and focus groups with DLM educators provided evidence of score report effectiveness and yielded additional information for refining score reports (Clark et al., 2018; Karvonen et al., 2017; Karvonen et al., 2016). These studies indicated educators were able to appropriately interpret score report contents, found them instructionally useful, and used them to reflect on their instructional decisions. Educators also suggested additional improvements, for instance, to improve clarity of language. Minor adjustments to score report structure is implemented on an annual basis based on feedback from the field and input from the DLM scoring and reporting ad hoc committee and the larger DLM Governance Board. Sample reports for 2021–2022 are displayed in the Scoring and Reporting Resources webpage (http://dynamiclearningmaps.org/srr/ye-2022). This page contains an overview of scoring, score report delivery, and data files. Because of continued impacts of the COVID-19 pandemic on instruction and assessment, during 2021–2022, state education agencies were given the option to add a cautionary statement to Individual Student Score Reports, which indicated that the results may reflect the continued effects of the COVID-19 pandemic on student performance. Four states opted to include the cautionary statement on their individual score reports.

7.6.2. Aggregate Reports

Student results are also aggregated into several other types of reports. At the classroom and school levels, roster reports list individual students with the number of EEs assessed, number of linkage levels mastered, and final performance level. District- and state-level reports provide frequency distributions, by grade level and overall, of students assessed and achieving at each performance level in ELA and mathematics. Sample aggregate reports are provided in the Scoring and Reporting Resources webpage (http://dynamiclearningmaps.org/srr/ye-2022).

7.6.3. Interpretation Resources

At the onset of the DLM assessment, the Theory of Action for the assessment set forth four tenets for score interpretation and use:

1. Scores represent what students know and can do.
2. Performance level descriptors provide useful information about student achievement.
3. Inferences regarding student achievement and progress can be drawn for individual conceptual areas.
4. Assessment scores provide information that can be used to guide instructional decisions.

To these ends, multiple supports are provided to aid score interpretation:

- The *Parent Interpretive Guide* is designed to provide definition and context to student score reports.
- Parent/guardian letter templates are designed to be used by educators and state superintendents to introduce the student reports to parents/guardians.
- The *Talking with Parents Guide* is designed to support educators' discussions and build understanding for parents/guardians and other stakeholders.
- The *Scoring and Reporting Guide for Administrators* targets building- and district-level administrators.
- A set of four short score-report interpretation videos, known as helplets, are available to help key stakeholders interpret assessment results as intended.
- All of the resources listed above and described in further detail below are compiled on the Scoring and Reporting Resources webpage (http://dynamiclearningmaps.org/srr/ye-2022). The score report videos are available on the "DLM Score Report Videos for Year-End States" webpage (https://dynamiclearningmaps.org/score-report-videos-ye).

7.6.3.1. Parent Interpretive Guide

The *Parent Interpretive Guide* uses a sample individual student report and text boxes to explain that the assessment measures student performance on alternate achievement standards for students with the most significant cognitive disabilities—the DLM EEs. The guide first provides a high-level explanation of the DLM alternate assessment before delving into specifics about interpreting the Performance Profile and the Learning Profile.

7.6.3.2. Parent Letters

Templates were developed for explanatory letters that educators and state superintendents could use to introduce parents/guardians to the student reports. These letters provided context for the reports, including a description of the DLM assessment, when the assessment was administered, and what results indicate about student performance.

The letter from the state superintendent emphasized that setting challenging and achievable academic goals for each student is the foundation for a successful and productive school year. The letter acknowledged that students have additional goals that parents/guardians and the students' Individualized Education Plan teams have established.

7.6.3.3. Talking with Parents Guide

An interpretive guide was provided for educators who would discuss results with parents/guardians or other stakeholders. The guide walked educators through directions for getting ready for a parent/guardian meeting, discussing the score report, and explaining other aspects of the DLM assessment.

7.6.3.4. Guide to Scoring and Reporting for Data Managers

The guide designed for local administrators, such as superintendents, district test coordinators, and other district staff, covered how assessment results were calculated and each type of report provided for DLM assessments (including class- and school-level results as well as district- and state-level results) and explained how reports would be distributed. The guide explained the contents of each report and provided hints about interpretation.

7.6.3.5. Score Report Videos

Short videos, called helplets, found on the DLM website[50] describe score report contents and ways educators can use results to inform instruction. The video content went through multiple rounds of internal review, followed by TAC review, and external review by educators from six states. Feedback from each group was incorporated into the final videos. The available helplets are:

- What do the DLM Alternate Assessments Measure?
- What is Skill Mastery?
- What Information is Contained in a Score Report?
- How can Score Reports be Used?

7.7. Quality-Control Procedures for Data Files and Score Reports

Quality-control procedures are implemented for all three data file types. To ensure that formatting and the order of columns are identical, column names in each file were compared with the data dictionary that is provided to states. Additional file-specific checks are conducted to ensure accuracy of all data files.

To allow quality-control checks to be performed more rapidly and efficiently, R programs were developed to perform quality-control procedures on the GRF and on Individual Student Score Reports.

The first program written to perform automated checks is designed to perform quality-control on the GRF. This program conducts a series of checks that can be organized into four main steps.

1. Check the data for reasonableness (checks described below).
2. Ensure the number of linkage levels mastered for each student is less than or equal to the maximum

[50] https://dynamiclearningmaps.org/score-report-videos-ye

possible value for that grade and subject.
3. Check all EE scores against the original scoring file.
4. Ensure all necessary fields are present and in the correct order. The automated program checks each row of data in the GRF and generates errors for review by the psychometrics team.

The reasonableness checks ensure that the GRF column names accurately match the data dictionary provided to states and check the columns to ensure that data match defined parameters. If invalid values are found, they are corrected as necessary by DLM staff and/or state education agencies during their 2-week review period.

An automated program is used to support manual review of all Individual Student Score Reports. The program was written to check key values used to generate Individual Student Score Reports. As the program generates reports, it creates a proofreader file containing the values that are used to create each score report. These values are then checked against the GRF to ensure that they are being accurately populated into score reports.

Demographic values including student name, school, district, grade level, state, and state student identifier are checked to ensure a precise match. Values of skills mastered, performance levels, conceptual areas tested and mastered, and EEs mastered and tested are also checked to ensure the correct values are populated. Values referring to the total number of skills, EEs, or conceptual areas available are checked to ensure they are the correct value for that grade and subject. In addition to the data files, individual student and aggregated score reports are generated within Kite® Educator Portal and checked for quality. Given the large number of score reports generated, a random sample of approximately 5%–10% of the score reports generated are checked manually.

For this sample, both the Performance Profile and the Learning Profile portions of the Individual Student Score Reports are checked for accuracy. Performance Profiles are checked to make sure the correct performance level displayed and matched with the value in the GRF. The percentage of skills mastered in the Performance Profile is compared against the GRF and the Learning Profile portion of the student score report to ensure that all three contain the same values. Additionally, the number of conceptual areas listed in the Performance Profile is compared with the blueprint. For each EE on the student's Learning Profile, the highest linkage level mastered is compared with the value for the EE in the GRF. For both the Performance Profile and Learning Profile, the number of EEs listed on the report is compared against the number listed in the blueprint for that subject and grade. Demographic information in the header of the Performance Profile and Learning Profile is checked to ensure that it matches values in the GRF. Formatting and text within each report is given an editorial review as well.

Aggregated reports undergo similar checks, including the comparison of header information to GRF data and verification that all students rostered to an educator or school (for class and school reports, respectively) are present and that no extraneous students are listed. Performance levels (for class and school reports) and the number of students with a given performance level (for district and state reports) are checked against the GRF.

Once all reports are checked, all files to be disseminated to states undergo a final set of checks to ensure that all files are present. This last set of checks involves higher-level assurances that the correct number of student files are present for each type of report according to the expected number calculated from the GRF, that file naming conventions are followed, and that all types of data files are present.

Any errors identified during quality-control checks are corrected prior to distribution of data files and score reports to states.

7.8. Conclusion

Results for DLM assessments include students' overall performance levels and linkage level mastery decisions for each assessed EE and linkage level. During spring 2022, assessments were administered to 88,404 students in 14 states. Between 11% and 37% of students achieved at the At Target or Advanced levels across all grades and subjects. Of the three scoring rules, linkage level mastery status was most frequently assigned by the diagnostic classification model, and students tend to demonstrate mastery of the Target or Successor level at higher rates in ELA than in mathematics. Additional scoring evidence indicates that writing tasks are reliably scored by test administrators, and students who achieve at the At Target performance level are on track to pursue postsecondary opportunities.

Following the spring 2022 administration, three data files were delivered to state education agencies: GRF, special circumstance code file, and exited students file. Score reports summarized assessment results at the individual student, class, school, district, and state levels. Supplementary materials to support score report interpretation and use are available on the DLM website.

8. Reliability

The Dynamic Learning Maps® (DLM®) Alternate Assessment System reports student results as discrete mastery classifications derived from diagnostic classification models (DCMs) and aggregations of the individual classifications. As such, methods for estimating the reliability must reflect the innovative assessment design and unique reporting structure of the DLM System. The approach draws on previous research from classification-based assessments to describe the reliability of student results. This chapter discusses the methods used to estimate reliability, the factors that are likely to affect variability in reliability results, and an overall summary of reliability evidence.

8.1. Background Information on Reliability Methods

Reliability estimates quantify the degree of precision in a test score. Expressed another way, a reliability index specifies how likely scores are to vary from one test administration to another due to chance. Historically, reliability has been quantified using indices such as the Guttman-Cronbach alpha (Cronbach, 1951; Guttman, 1945), which provides an index of the proportion of variance in a test score that is due to variance in the trait. Values closer to 1.0 indicate variation in test scores comes from individual differences in the trait, whereas values closer to 0.0 indicate variation in test scores comes from random error.

Many traditional measures of reliability exist; their differences are due to assumptions each measure makes about the nature of the data from a test. For instance, the Spearman-Brown reliability formula assumes items are parallel, contributing equal amounts of information about the trait and having equal variance. The Guttman-Cronbach alpha assumes tau-equivalent items (i.e., items with equal information about the trait but not necessarily equal variances). As such, the alpha statistic is said to subsume the Spearman-Brown statistic, meaning that if the data meet the stricter definition of Spearman-Brown, then alpha will be equal to Spearman-Brown. As a result, inherent in any discussion of reliability is the fact that the metric of reliability is accurate to the extent that the assumptions of the test are met.

The DLM assessments are scored using diagnostic classification models (DCMs), which assume that students' knowledge, skills, and understandings are represented by discrete mastery statuses, rather than a continuous latent trait that characterizes more traditional classical test theory and item response theory models. As such, reliability-estimation methods based on item response theory estimates of ability are not applicable for the DLM assessments. Therefore, the reliability evidence may appear different from that reported when test scores are produced using traditional psychometric techniques such as classical test theory or item response theory. However, interpretation of indices for DLM assessments is consistent with traditional approaches. When a test is perfectly reliable (i.e., it has an index value of 1), any variation in test scores comes from individual differences in the trait within the sample in which the test was administered. When a test has zero reliability, then any variation in test scores comes solely from random error.

DCMs are models that produce classifications based on probability estimates for students. For the DLM System, the classification estimates are based on the set of attributes across the alternate achievement standards on which each student was assessed. The alternate achievement standards are themselves organized into larger content strands. In DLM terms, each content strand is called a conceptual area, each of which is made up of standards called Essential Elements (EEs). Each EE is available at five linkage levels of complexity, which are the attributes in the DCM[51]: Initial Precursor, Distal Precursor, Proximal

[51] For more information on the specification of the DCMs, see Chapter 5 of this manual.

Precursor, Target, and Successor.

DLM testlets are written with items measuring the linkage level. Because of the DLM administration design, students do not take testlets outside of a single linkage level within an EE. Students take a single testlet measuring a given EE and linkage level; consequently, data obtained when students respond to testlets at adjacent linkage levels within an EE are sparse. Therefore, a linkage level DCM is used to score the assessment (i.e., estimate mastery proficiency; see Chapter 5 in this manual for more information).

The DCMs produce student-level posterior probabilities for each linkage level for which a student is assessed, with a threshold of 0.8 specified for demonstrating mastery. To guard against the model being overly influential, two additional scoring rules are applied. Students can also demonstrate mastery by providing correct responses to at least 80% of the items measuring the EE and linkage level.[52] Furthermore, because students are not assessed at more than one linkage level within an EE, students who do not meet mastery status for any assessed linkage level are assigned mastery status for the linkage level two levels below the lowest level on which they are assessed (unless the lowest level tested is either the Initial Precursor or Distal Precursor levels, in which case students are considered nonmasters of all linkage levels within the EE). See Chapter 7 of this manual for a complete description of scoring rules for the DLM assessments.

The DLM score reports display linkage level mastery for each EE. Linkage level results are also aggregated for EEs within each conceptual area in ELA and mathematics. Score reports also summarize overall performance in each subject with a performance level classification. The classification is determined by summing all linkage levels mastered in each subject and comparing the value with cut points determined during standard setting. For more information on cut points, see Chapter 6 of this manual. For more information on score reports, see Chapter 7 of this manual.

Consistent with the levels at which DLM results are reported, this chapter provides reliability evidence at six levels: (1) the classification accuracy of each linkage level within each EE (linkage level reliability), (2) the classification accuracy summarized for the five linkage levels (conditional evidence by linkage level), (3) the number of linkage levels mastered within each EE (EE reliability), (4) the number of linkage levels mastered within each conceptual area in ELA and mathematics (conceptual area reliability), (5) the total number of linkage levels mastered in each subject (subject reliability), and (6) the classification to overall performance level (performance level reliability). As described in the next section, reliability evidence comes from simulated retests in which assessment data are generated for students with the estimated model parameters and student mastery probabilities.

The reliability methods and evidence presented in this chapter adhere to guidance given in the *Standards for Educational and Psychological Testing* (AERA et al., 2014). Where relevant, evidence provided in accordance with specific standards is noted.

8.2. Methods of Obtaining Reliability Evidence

Because the DLM psychometric model produces complex mastery results summarized at multiple levels of aggregation (linkage level, EE, conceptual area, subject, and performance level), rather than a traditional raw or scale score value, methods for evaluating reliability were based on simulated retests. For a

[52] For a description of how often each scoring rule is utilized, see Chapter 7 of this manual.

simulation-based method of computing reliability, we generate simulated examinees with known characteristics, simulate assessment data using calibrated-model parameters, score the simulated retests using calibrated-model parameters, and compare estimated examinee characteristics with those characteristics known to be true in the simulation. For DLM assessments, the known characteristics of the simulated examinees are the set of linkage levels the examinee has mastered and not mastered.

Most methods for estimating the reliability of assessments scaled with DCMs are limited to attribute-level summaries of reliability (for a review see Sinharay & Johnson, 2019). Accordingly, these methods do not generalize to aggregated summaries of mastery classifications, such as those reported for DLM assessments. Simulated retests offer one method for estimating the reliability of DCM-based assessments at multiple levels of aggregation. At the attribute level (i.e., individual mastery classifications), simulated retests provide reliability estimates that are highly consistent with nonsimulation-based methods (W. J. Thompson, 2020). However, unlike the nonsimulation-based methods, simulated retests are able to support the evaluation of reliability for aggregations of individual mastery classifications (W. J. Thompson et al., 2019). In addition to supporting the evaluation of reliability evidence at multiple levels of reporting, simulated retests provide results consistent with classical reliability metrics in that perfect reliability is evidenced by consistency in classification, and zero reliability is evidenced by a lack of classification consistency.

The simulated retests used to estimate reliability for DLM versions of scores and classifications consider the unique design and administration of DLM assessments. Students typically take only 3–5 items per EE. Simulated retests are based on a replication of the administration process, including adaptive routing between testlets, and uses students' known mastery classifications from the operational assessment. Therefore, students may not receive the same testlets in the simulation as they did during their actual assessment (i.e., routing decisions may be different or different testlets may be assigned from the pool of available testlets). This means that the simulated retest offers a genuine approximation of actual retest assignment for any given student. Simulated retests replicate results of DLM assessments from actual examinees based on administration procedures specific to the DLM assessments. However, the use of simulation produces approximate estimates of reliability, which are contingent on the accuracy of the current scoring model. That is, reliability estimates are an upper bound on the true reliability. For the 2021–2022 administration, on advice of the DLM Technical Advisory Committee, the procedure for simulating the retest was updated to incorporate additional uncertainty into the simulation-assigned mastery status and testlets for each resampled student. The remaining sections of this chapter describe the current procedures and results with these updates, which provide a better estimate of the true reliability.

Simulated retests were conducted to assemble reliability evidence according to the *Standards'* assertion that "the general notion of reliability/precision is defined in terms of consistency over replications of the testing procedure" (AERA et al., 2014, p. 35). The reliability evidence reported here supports "interpretation for each intended score use," as Standard 2.0 recommends (AERA et al., 2014, p. 42). The "appropriate evidence of reliability/precision" (AERA et al., 2014, p. 42) was assembled using a methodology that aligns to the design of the assessment and interpretations of results. The procedures used to assemble reliability evidence align with all applicable standards.

8.2.1. Reliability Sampling Procedure

The simulation design that was used to obtain the reliability estimates uses a resampling design to mirror DLM assessment data. In accordance with Standard 2.1, the sampling design uses the entire set of operational assessment data to generate simulated examinees (AERA et al., 2014, p. 42). Using this process guarantees that the simulation takes on characteristics of the DLM operational assessment data that are likely to affect reliability results. For one simulated examinee, the process is as follows:

1. Draw with replacement the student record of one student from the operational assessment data (i.e., the spring assessment window). Use the student's originally scored linkage level mastery probabilities as the true values for the simulated student data. For linkage levels the drawn student was assessed on during the operational assessment, generate the mastery status from a Bernoulli distribution with a probability equal to the mastery probability. For linkage levels the student was not assessed on during the operational assessment, the probability is either fixed to 1 if the student has mastered a higher linkage level for the EE during the operational assessment, or is defined as the base rate of class membership for linkage levels higher than those assessed operationally (see Chapter 5 of this manual).
2. Simulate a new assessment based on administration rules. In practice, this means simulating one testlet at a time and applying routing rules (see Chapter 4 of this manual) to assign subsequent simulated testlets. Item responses are simulated for the assigned testlets from calibrated model parameters,[53] conditional on the linkage level mastery status determined in Step 1.
3. Score the simulated item responses using the operational DLM scoring procedure, estimating linkage level mastery or nonmastery for the simulated student. See Chapter 7 of this manual for more information.[54]
4. Calculate the aggregated summaries of linkage level mastery for the simulated retests (i.e., EE, conceptual area, subject, and performance level).
5. Compare the estimated linkage level mastery and aggregated summaries from the simulated retests to the values reported for the drawn student on the operational assessment.

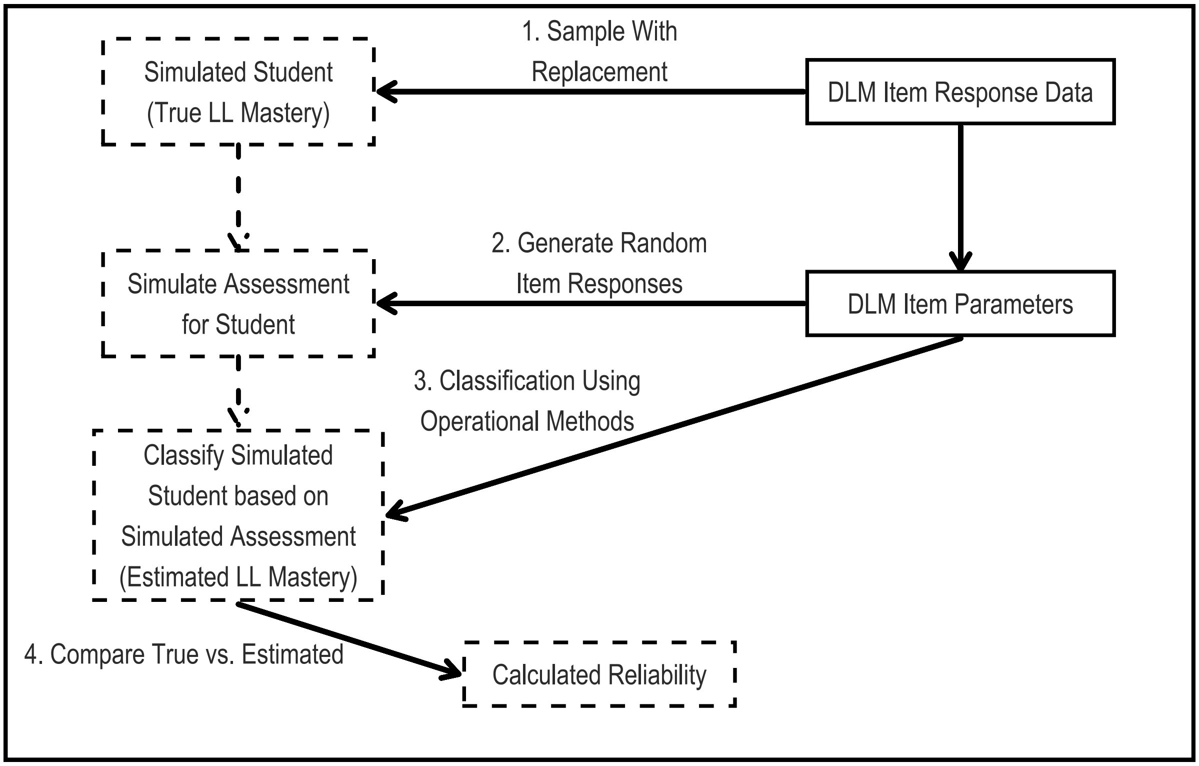
Steps 1 through 5 are then repeated 100,000 times for each grade and subject to create the full simulated retests data set. Figure 8.1 shows the steps of the simulation process as a flow chart.

[53] Calibrated parameters were treated as true and fixed values for the simulated retests.

[54] All three scoring rules were implemented when scoring the simulated retests to be consistent with the operational scoring procedure.

Figure 8.1

Simulation Process for Creating Reliability Evidence



Note. LL = linkage level.

8.3. Reliability Evidence

This chapter provides reliability evidence for six levels of data: (1) linkage level reliability, (2) conditional reliability by linkage level, (3) EE reliability, (4) conceptual area reliability, (5) subject reliability, and (6) performance level reliability, which ensures that the simulation and resulting reliability evidence are aligned with Standard 2.2 (AERA et al., 2014, p. 42). Additionally, providing reliability evidence for each of the six levels ensures that these reliability estimation procedures meet Standard 2.5 (AERA et al., 2014, p. 43). With 163 EEs, each comprising 5 linkage levels, the procedure includes 815 analyses to summarize reliability results. Because of the number of analyses, this chapter includes a summary of the reported evidence. The website version of this report[55] provides a full report of reliability evidence for all 815 linkage levels and 163 EEs. The full set of evidence is provided in accordance with Standard 2.12 (AERA et al., 2014, p. 45).

Reliability evidence at each level is reported using various correlation coefficients. Correlation estimates mirror estimates of reliability from contemporary measures such as the Guttman-Cronbach alpha. For linkage level and conditional evidence by linkage level reliability, the tetrachoric correlation estimates the relationship between true and estimated linkage level mastery statuses. The tetrachoric correlation is a

[55] https://2022-ye-techmanual.dynamiclearningmaps.org/8-reliability

special case of the polychoric in which the variables are discrete. Both the polychoric and tetrachoric correlations provide more useful estimates of relationships between ordinal and discrete variables that would otherwise be attenuated using the standard correlation (i.e., the Pearson coefficient). For EE and performance level reliability, the polychoric correlation estimates the relationship between two ordinal variables: the true performance level or true number of linkage levels mastered and the corresponding estimated value. Finally, for subject and conceptual area reliability, the Pearson correlation estimates the relationship between the true and estimated numbers of linkage levels mastered.

Reliability evidence at each level is also reported using correct classification rates (raw and chance corrected), indicating the proportion of estimated classifications that match true classifications. The chance-corrected classification rate, kappa, represents the proportion of error reduced above chance. Kappa values above .6 indicate substantial-to-perfect agreement between estimated and true values (Landis & Koch, 1977). However, Cohen's kappa may be limited in this context. Numerous studies have shown that the kappa statistic tends to be too conservative when there are unbalanced categories (Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990; O'Leary et al., 2014; Pontius & Millones, 2011). This is often the case for DLM linkage levels, where the base rate of class membership can be less than .3 or greater than .7.[56] Thus, it is important to interpret the kappa statistic with caution and within the context of the other reporting metrics.

8.3.1. Linkage Level Reliability Evidence

Evidence at the linkage level comes from comparing the true and estimated mastery status for each of the 815 linkage levels in the operational DLM assessment. This level of reliability reporting is the finest grain of reporting, and while it does not have a directly comparable classical test theory or item response theory analogue, its inclusion is important because it is the level at which mastery classifications are made for DLM assessments. All reported summary statistics of linkage level reliability are based on the resulting contingency tables: the comparison of true (operational assessment) and estimated (simulated retest) mastery statuses across all simulated examinees.

In addition to summary statistics from the simulated retests, we also calculated the classification consistency metric, $\hat{\mathrm{P}}_{\mathrm{C}}$, described by Johnson and Sinharay (2018). As the name implies, the classification consistency index is a measure of how consistent the student-level classifications are for each linkage level, and it is calculated from the estimated DCM parameters (see Chapter 5 of this manual for a description of the model parameters). The classification consistency metric is based on the estimated model parameters, and thus is only applicable to the linkage level, which is the unit of model estimation.[57] This metric is not based on simulated retests, and thus provides a measure of reliability independent from the simulation.

For each statistic, figures are given comparing the results of all 815 linkage levels. We report linkage level reliability evidence based on three summary statistics from the simulated retests and the nonsimulation-based classification consistency:

1. the tetrachoric correlation between estimated and true mastery status,
2. the classification agreement for the mastery status of each linkage level,

[56] See Chapter 5 of this manual for a summary of base rates of class membership.
[57] See Chapter 5 of this manual for a complete description of the model specification.

3. the classification consistency (Johnson & Sinharay, 2018), and
4. the classification agreement Cohen's kappa for the mastery status of each linkage level.

As there are 815 total linkage levels across all 163 EEs, the summaries reported herein are based on the proportion and number of linkage levels that fall within a given range of an index value (results for individual linkage levels can be found in the website version of this report[58]). Results are given in both tabular and graphical forms. Table 8.1 and Figure 8.2 provide proportions and the number of linkage levels, respectively, that fall within prespecified ranges of values for the four linkage level reliability summary statistics (i.e., tetrachoric correlation, classification agreement rate, classification consistency, and Cohen's kappa).

The correlations and classification agreement rates show reliability evidence for the classification of mastery at the linkage level. Across all linkage levels, 42 (5%) had a tetrachoric correlation below .6, 0 (0%) had a percent classification agreement below .6, 6 (1%) had a classification consistency below .6, and 444 (54%) had a Cohen's kappa below .6. As previously described, Cohen's kappa may be limited in this context due to unbalanced class categories. Thus, the other three metrics provide a more useful evaluation of linkage level reliability.

Notably, Johnson and Sinharay (2018) recommend a cutoff of .7 for fair classification consistency. Overall, 731 (90%) linkage levels meet this cutoff, indicating that the linkage level classifications show a high degree of reliability.[59]

Table 8.1

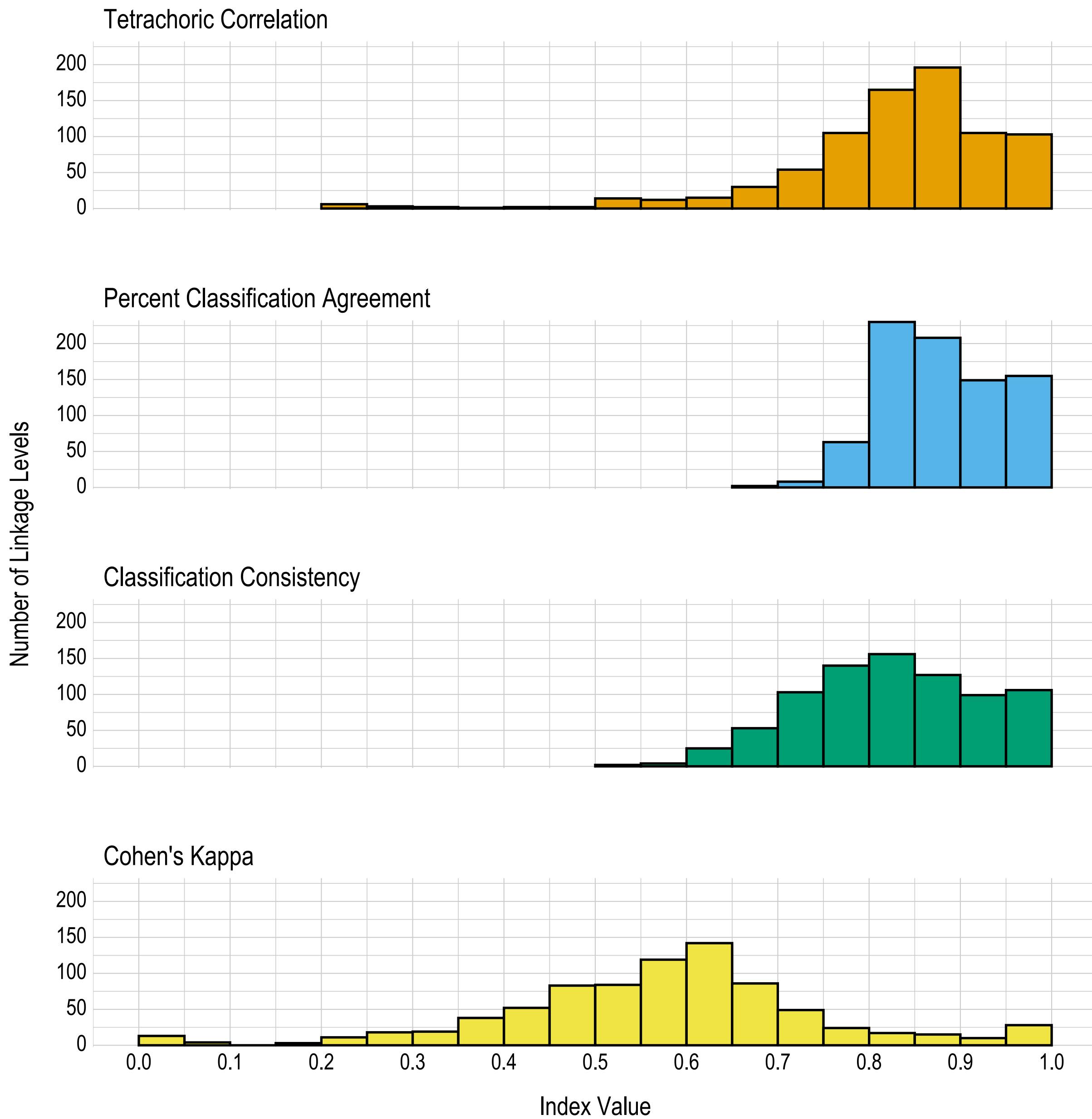
Reliability Summaries Across All Linkage Levels: Proportion of Linkage Levels Falling Within a Specified Index Range

Reliability index	Index range								
	0.00–0.59	0.60–0.64	0.65–0.69	0.70–0.74	0.75–0.79	0.80–0.84	0.85–0.89	0.90–0.94	0.95–1.00
Tetrachoric correlation	.052	.018	.037	.066	.129	.202	.240	.129	.126
Percent classification agreement	.000	.000	.002	.010	.077	.282	.255	.183	.190
Classification consistency	.007	.031	.065	.126	.172	.191	.156	.121	.130
Cohen's kappa	.545	.174	.106	.060	.029	.021	.018	.012	.034

58 https://2022-ye-techmanual.dynamiclearningmaps.org/8-reliability
59 For a summary of the closely related classification accuracy, see Chapter 5 of this manual.

Figure 8.2

Summaries of Linkage Level Reliability



8.3.2. Conditional Reliability Evidence by Linkage Level

Traditional assessment programs often report conditional standard errors of measurement to indicate how the precision of measurement differs along the score continuum. The DLM assessment system does not report total or scale score values. Therefore, traditional measures of conditional reliability are not applicable. In particular, standard errors of measurement (inversely related to reliability) that are conditional on a continuous trait are based on the calculation of Fisher's information, which involves taking the second derivative-model likelihood function with respect to the latent trait. When classifications are the latent traits, however, the likelihood is not a smooth function regarding levels of the trait and therefore cannot be differentiated (Henson & Douglas, 2005; Templin & Bradshaw, 2013). In other words, because diagnostic classification modeling does not produce a total score or scale score, traditional methods of calculating conditional standard errors of measurement are not appropriate. However, because DLM assessments were designed to span the continuum of students' varying knowledge, skills, and understandings as defined by the five linkage levels, evidence of reliability can be summarized for each linkage level to approximate conditional evidence over all EEs, similar to a conditional standard error of measurement for a total score.

Conditional reliability evidence by linkage level is based on the true and estimated mastery statuses for each linkage level, as summarized by each of the five levels. Results are reported using the same four statistics used to summarize the overall linkage level reliability evidence (i.e., tetrachoric correlation, classification agreement rate, classification consistency, and Cohen's kappa).

Table 8.2 and Figure 8.3 provide the proportions and the number of linkage levels, respectively, that fall within prespecified ranges of values for each linkage level for the four reliability summary statistics (i.e., tetrachoric correlation, classification agreement rate, classification consistency, and Cohen's kappa). The correlations and classification agreement rates generally indicate that all five linkage levels provide reliable classifications of student mastery; results are fairly consistent across all linkage levels for each of the four statistics reported.

Table 8.2

Reliability Summaries Across All Linkage Levels: Proportion of Linkage Levels Falling Within a Specified Index Range

Reliability index	Index range								
	0.00–0.59	0.60–0.64	0.65–0.69	0.70–0.74	0.75–0.79	0.80–0.84	0.85–0.89	0.90–0.94	0.95–1.00
Initial Precursor									
Tetrachoric correlation	.000	.000	.012	.055	.129	.258	.239	.141	.166
Percent classification agreement	.000	.000	.000	.000	.147	.350	.282	.147	.074
Classification consistency	.000	.000	.000	.043	.110	.307	.288	.178	.074
Cohen's kappa	.319	.258	.153	.092	.031	.043	.043	.000	.061
Distal Precursor									
Tetrachoric correlation	.043	.012	.043	.049	.098	.245	.325	.117	.067
Percent classification agreement	.000	.000	.012	.049	.135	.515	.227	.031	.031
Classification consistency	.006	.037	.098	.209	.184	.160	.061	.104	.141
Cohen's kappa	.368	.258	.209	.086	.025	.025	.000	.012	.018
Proximal Precursor									
Tetrachoric correlation	.031	.018	.006	.074	.104	.239	.252	.147	.129
Percent classification agreement	.000	.000	.000	.000	.074	.331	.350	.166	.080
Classification consistency	.000	.037	.098	.202	.196	.172	.135	.086	.074
Cohen's kappa	.485	.178	.110	.061	.061	.018	.025	.018	.043
Target									
Tetrachoric correlation	.092	.037	.049	.067	.184	.141	.196	.110	.123
Percent classification agreement	.000	.000	.000	.000	.031	.172	.245	.294	.258
Classification consistency	.012	.043	.061	.110	.196	.172	.166	.092	.147
Cohen's kappa	.687	.098	.043	.043	.031	.012	.025	.025	.037
Successor									
Tetrachoric correlation	.092	.025	.074	.086	.129	.129	.190	.129	.147
Percent classification agreement	.000	.000	.000	.000	.000	.043	.172	.276	.509

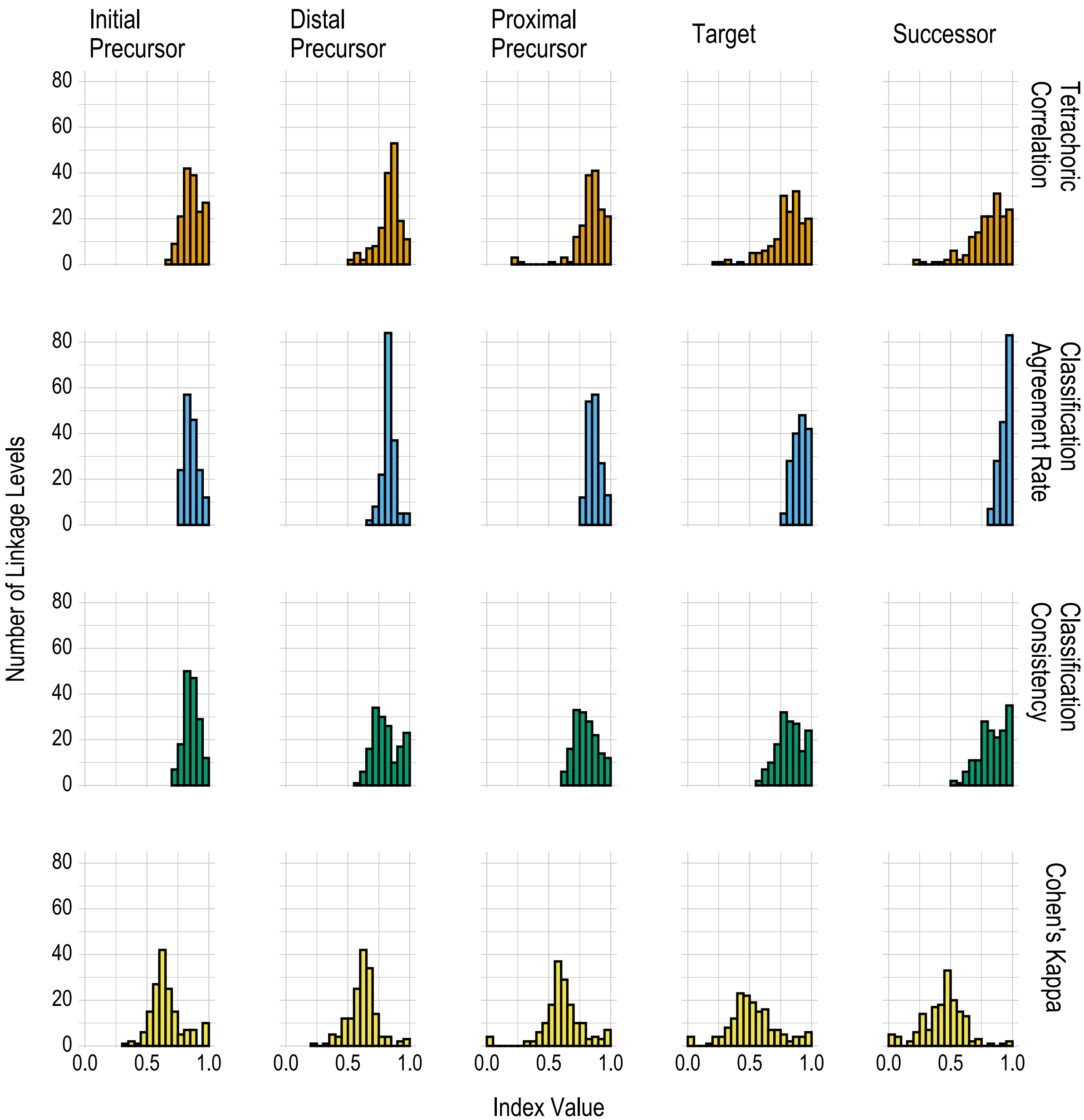
Table 8.2

Reliability Summaries Across All Linkage Levels: Proportion of Linkage Levels Falling Within a Specified Index Range (continued)

Reliability index	0.00–0.59	0.60–0.64	0.65–0.69	0.70–0.74	0.75–0.79	0.80–0.84	0.85–0.89	0.90–0.94	0.95–1.00
Classification consistency	.018	.037	.067	.067	.172	.147	.129	.147	.215
Cohen's kappa	.865	.080	.012	.018	.000	.006	.000	.006	.012

Figure 8.3

Conditional Reliability Evidence Summarized by Linkage Level



8.3.3. *Essential Element Reliability Evidence*

The first level of linkage level aggregation is the EE. EE-level results are reported as the highest linkage level mastered for each EE. Because EE-level results are an aggregation of the individual linkage level classifications, more traditional measures of the reliability (e.g., the classification consistency used for linkage levels) are not possible. Therefore, reliability results are only reported based on the simulated retests, which do offer a method for evaluating the reliability of aggregated classifications (W. J. Thompson et al., 2019).

Three statistics are used to summarize reliability evidence for EEs:

1. the polychoric correlation between true and estimated numbers of linkage levels mastered within an EE,
2. the classification agreement rate for the number of linkage levels mastered within an EE, and
3. the classification agreement Cohen's kappa for the number of linkage levels mastered within an EE.

Because there are 163 EEs, the summaries are reported herein according to the number and proportion of EEs that fall within a given range of an index value (results for individual EEs can be found in the website version of this report[60]). Results are given in both tabular and graphical forms. Table 8.3 and Figure 8.4 provide the proportions and the number of EEs, respectively, falling within prespecified ranges of values for the three reliability summary statistics (i.e., classification agreement rate, kappa, correlation). Across all EEs, 2 (1%) had a polychoric correlation below .6, 90 (55%) had a percent classification agreement below .6, and 10 (6%) had a Cohen's kappa below .6. It should be noted the classification agreement rate is measuring the exact agreement between the reported results and simulated retests. Due to the adaptive routing, students are often assigned different linkage levels in the simulated retests than were administered operationally. The highest linkage level mastered for each EE is therefore often different in the simulated retests. Thus, this measure of reliability is likely less useful than the other two metrics for evaluating EE reliability. In general, the reliability summaries show strong evidence for reliability for the number of linkage levels mastered within EEs.

Table 8.3

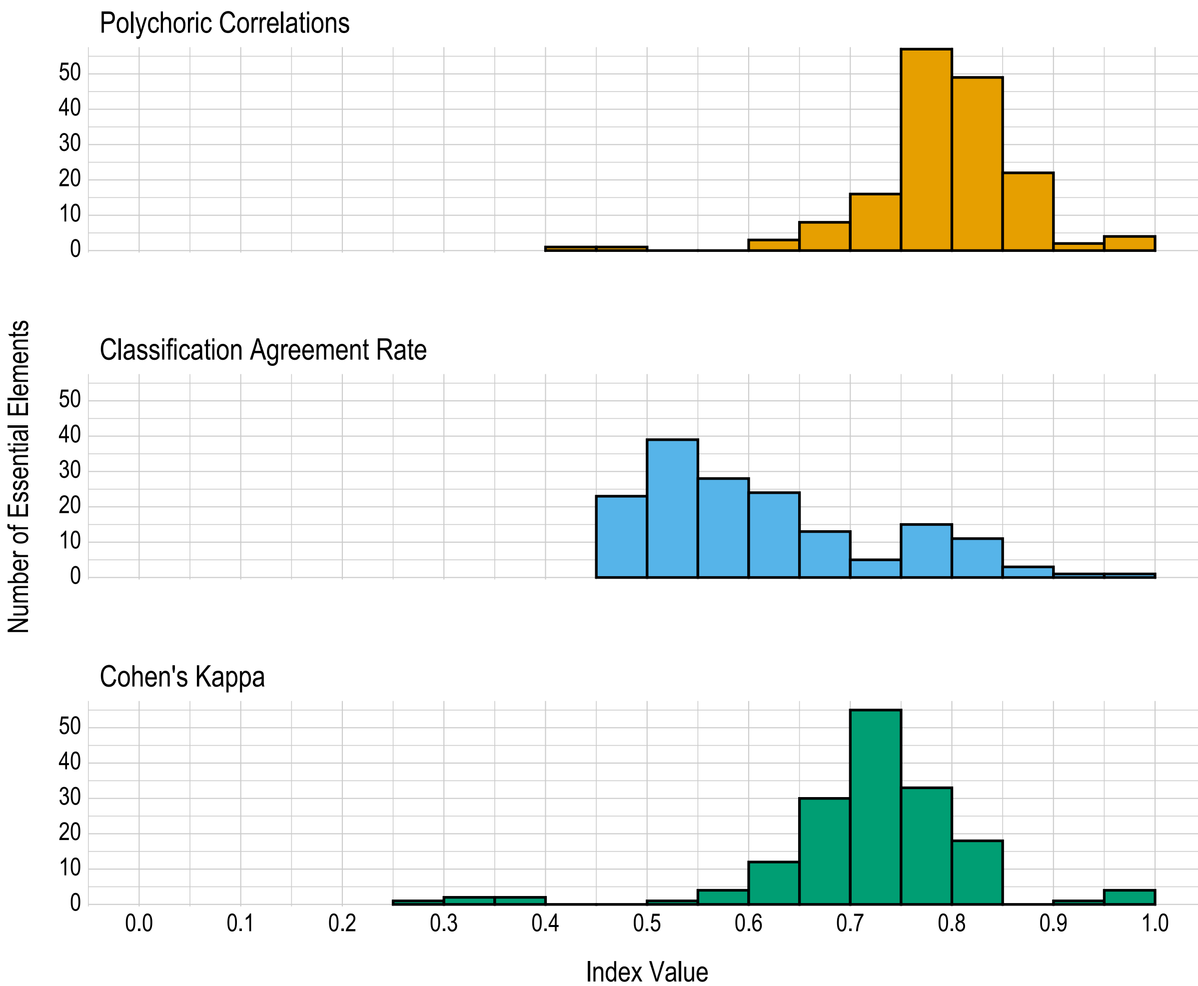
Reliability Summaries Across All Essential Elements: Proportion of Essential Elements Falling Within a Specified Index Range

	Index range								
Reliability index	0.00–0.59	0.60–0.64	0.65–0.69	0.70–0.74	0.75–0.79	0.80–0.84	0.85–0.89	0.90–0.94	0.95–1.00
Polychoric correlation	.012	.018	.049	.098	.350	.301	.135	.012	.025
Percent classification agreement	.552	.147	.080	.031	.092	.067	.018	.006	.006
Cohen's kappa	.061	.074	.184	.337	.202	.110	.000	.006	.025

60 https://2022-ye-techmanual.dynamiclearningmaps.org/8-reliability

Figure 8.4

Number of Linkage Levels Mastered Within Essential Element Reliability Summaries



8.3.4. Conceptual Area Reliability Evidence

ELA and mathematics EEs are organized into conceptual areas, which are akin to content strands for other assessments. These collections of related EEs describe the overarching sections of the learning map that are the foundation of the development of DLM assessments.[61] Because Individual Student Score Reports summarize the number and percentage of linkage levels students mastered in each conceptual area,[62] we also provide reliability evidence for each conceptual area, in accordance with Standard 2.2, which indicates that reliability evidence should be provided consistent with the level(s) of scoring (AERA et al., 2014, p. 42).

Conceptual area reliability provides consistency evidence for the number of linkage levels mastered across all EEs in each conceptual area for each grade and subject. Conceptual area reliability evidence compares

[61] For more information, see Chapter 2 of this manual.
[62] See Chapter 7 of this manual for more information.

the true and estimated number of linkage levels mastered across all tested levels for each conceptual area. Reliability is reported with three summary numbers:

1. the Pearson correlation between the true and estimated number of linkage levels mastered within a conceptual area,
2. the classification agreement rate for which linkage levels were mastered as averaged across all simulated students for each conceptual area, and
3. the classification agreement Cohen's kappa for which linkage levels were mastered, as averaged across all simulated students for each conceptual area.

Table 8.4 shows the three summary values for each conceptual area. Classification rate information is provided in accordance with Standard 2.16 (AERA et al., 2014, p. 46). The correlation and Cohen's kappa summary statistics included in Table 8.4 also align with Standard 2.19 (AERA et al., 2014, p. 47). The correlations range from .356 to .923 in ELA and from .346 to .849 in mathematics, the average classification agreement rates range from .889 to .997 in ELA and from .948 to .998 in mathematics, and the average Cohen's kappa values range from .675 to .996 in ELA and from .868 to .998 in mathematics. This indicates that, overall, the conceptual area results provided on score reports are reliable.

Table 8.4

Summary of Conceptual Area Reliability Evidence

Grade	Conceptual area	Linkage levels mastered correlation	Average student classification agreement	Average student Cohen's kappa
English language arts				
3	ELA.C1.1	.796	.960	.912
3	ELA.C1.2	.804	.950	.881
3	ELA.C1.3	.547	.983	.974
3	ELA.C2.1	.781	.984	.973
4	ELA.C1.1	.800	.935	.841
4	ELA.C1.2	.816	.924	.801
4	ELA.C1.3	.518	.986	.978
4	ELA.C2.1	.923	.991	.984
5	ELA.C1.1	.577	.997	.996
5	ELA.C1.2	.843	.905	.712
5	ELA.C1.3	.619	.942	.874
5	ELA.C2.1	.893	.995	.992
6	ELA.C1.1	.356	.968	.948
6	ELA.C1.2	.851	.908	.745
6	ELA.C1.3	.623	.987	.981
6	ELA.C2.1	.863	.992	.987
7	ELA.C1.1	.427	.980	.968
7	ELA.C1.2	.820	.931	.824

Table 8.4

Summary of Conceptual Area Reliability Evidence (continued)

Grade	Conceptual area	Linkage levels mastered correlation	Average student classification agreement	Average student Cohen's kappa
7	ELA.C1.3	.687	.928	.818
7	ELA.C2.1	.884	.974	.946
8	ELA.C1.2	.820	.889	.675
8	ELA.C1.3	.684	.966	.937
8	ELA.C2.1	.832	.968	.934
9	ELA.C1.2	.843	.913	.771
9	ELA.C1.3	.755	.909	.768
9	ELA.C2.1	.743	.968	.938
9	ELA.C2.2	.847	.994	.992
10	ELA.C1.2	.841	.913	.771
10	ELA.C1.3	.757	.910	.770
10	ELA.C2.1	.725	.967	.935
10	ELA.C2.2	.840	.994	.992
11	ELA.C1.2	.805	.896	.722
11	ELA.C1.3	.748	.927	.826
11	ELA.C2.1	.833	.958	.908
11	ELA.C2.2	.688	.989	.983
Mathematics				
3	M.C1.1	.705	.979	.958
3	M.C1.3	.551	.993	.991
3	M.C2.2	.521	.978	.964
3	M.C3.1	.497	.993	.990
3	M.C3.2	.628	.993	.991
3	M.C4.1	.547	.997	.996
3	M.C4.2	.609	.992	.990
4	M.C1.1	.596	.988	.983
4	M.C1.3	.788	.998	.998
4	M.C2.1	.577	.991	.987
4	M.C2.2	.524	.986	.979
4	M.C3.1	.849	.948	.868
4	M.C4.2	.543	.987	.981
5	M.C1.1	.485	.988	.984
5	M.C1.2	.736	.973	.946
5	M.C1.3	.633	.996	.996
5	M.C2.1	.668	.960	.921
5	M.C2.2	.574	.989	.984

Table 8.4

Summary of Conceptual Area Reliability Evidence (continued)

Grade	Conceptual area	Linkage levels mastered correlation	Average student classification agreement	Average student Cohen's kappa
5	M.C3.2	.547	.989	.985
6	M.C1.2	.604	.959	.915
6	M.C2.2	.671	.960	.913
6	M.C3.2	.520	.989	.985
6	M.C4.1	.573	.974	.952
7	M.C1.1	.503	.984	.976
7	M.C1.3	.847	.982	.961
7	M.C2.1	.650	.983	.973
7	M.C2.2	.486	.992	.989
7	M.C4.1	.538	.991	.987
8	M.C1.1	.555	.991	.988
8	M.C1.3	.537	.984	.976
8	M.C2.1	.677	.991	.987
8	M.C2.2	.536	.987	.981
8	M.C3.2	.562	.986	.980
8	M.C4.1	.451	.994	.993
8	M.C4.2	.602	.975	.947
9	M.C1.3	.829	.979	.958
9	M.C2.1	.664	.961	.923
9	M.C4.1	.734	.967	.936
10	M.C1.3	.491	.991	.987
10	M.C2.1	.516	.993	.991
10	M.C3.1	.470	.995	.994
10	M.C3.2	.699	.988	.980
10	M.C4.1	.503	.994	.993
10	M.C4.2	.610	.971	.940
11	M.C1.3	.642	.972	.948
11	M.C2.1	.641	.988	.983
11	M.C3.2	.346	.983	.977
11	M.C4.2	.638	.967	.935

8.3.5. Subject Reliability Evidence

The next level of aggregation of linkage level mastery is for the subject overall. Subject reliability provides consistency evidence for the number of linkage levels mastered across all EEs for a given grade and subject. Because students are assessed on multiple linkage levels across the assessed EEs in each subject, subject reliability evidence is similar to reliability evidence for testing programs that use summative assessments to describe overall performance in a subject. That is, the number of linkage levels mastered within a subject is analogous to the number of items answered correctly (i.e., total score) in a different type of testing program.

Subject reliability evidence compares the true and estimated number of linkage levels mastered across all tested levels for a given subject. Because subject-level reporting summarizes the total number of linkage levels a student mastered, the statistics reported for subject reliability are the same as those reported for conceptual area reliability. Reliability is reported with three summary values:

1. the Pearson correlation between the true and estimated number of linkage levels mastered within a subject,
2. the classification agreement rate for which linkage levels were mastered, as averaged across all simulated students, and
3. the classification agreement Cohen's kappa for which linkage levels were mastered, as averaged across all simulated students.

Table 8.5 shows the three summary values for each grade and subject. The correlation between true and estimated number of linkage levels mastered ranges from .758 to .914. Students' average classification agreement rates range from .896 to .946 and average Cohen's kappa values range from .691 to .825. These values indicate that the total linkage levels mastered in a subject are reliably determined.

Table 8.5

Summary of Subject Reliability Evidence

Grade	Linkage levels mastered correlation	Average student classification agreement	Average student Cohen's kappa
English language arts			
3	.901	.915	.739
4	.902	.909	.731
5	.914	.907	.711
6	.901	.913	.753
7	.892	.901	.716
8	.893	.899	.707
9	.912	.903	.726
10	.912	.904	.726
11	.907	.896	.691
Mathematics			
3	.835	.941	.780
4	.901	.925	.739
5	.865	.923	.745
6	.803	.920	.739
7	.860	.939	.781
8	.855	.927	.752
9	.873	.930	.779
10	.792	.946	.813
11	.758	.938	.825

8.3.6. Performance Level Reliability Evidence

The final level of linkage level mastery aggregation is at the overall performance level. Results for DLM assessments are reported using four performance levels. The scoring procedure sums the linkage levels mastered across all EEs in each subject, and cut points are applied to distinguish between the four performance categories.[63]

Performance level reliability provides evidence for how reliably students are classified to the four performance levels for each subject and grade level. Because the performance level is determined by the total number of linkage levels mastered, large fluctuations in the number of linkage levels mastered, or fluctuation around the cut points, could affect how reliably students are assigned into performance categories. The performance level reliability evidence is based on the observed and estimated performance levels from the simulated retests (i.e., based on the estimated total number of linkage levels

[63] See Chapter 6 of this manual for details on the standard setting procedure to determine the cut points.

mastered and predetermined cut points). Three statistics are included to provide a comprehensive summary of results:

1. the polychoric correlation between the true and estimated performance levels within a grade and subject,
2. the classification agreement rate between the true and estimated performance levels within a grade and subject, and
3. the classification agreement Cohen's kappa between the true and estimated performance levels within a grade and subject.

Table 8.6 presents this information across all grades and subjects. Polychoric correlations between true and estimated performance level range from .789 to .928. Classification agreement rates range from .640 to .793, and Cohen's kappa values are between .665 and .848. These results indicate that the DLM scoring procedure of reporting performance levels based on total linkage levels mastered results in reliable classification of students to performance level categories.

Table 8.6

Summary of Performance Level Reliability Evidence

Grade	Polychoric correlation	Classification agreement rate	Cohen's kappa
English language arts			
3	.920	.772	.794
4	.910	.771	.792
5	.921	.745	.831
6	.893	.705	.771
7	.898	.713	.803
8	.904	.754	.803
9	.914	.756	.813
10	.928	.793	.817
11	.916	.745	.811
Mathematics			
3	.887	.716	.735
4	.925	.737	.848
5	.889	.651	.774
6	.829	.657	.710
7	.897	.750	.776
8	.856	.706	.694
9	.880	.700	.770
10	.833	.739	.703
11	.789	.640	.665

8.4. Conclusion

In summary, reliability measures for the DLM assessment system address the standards set forth by AERA et al. (2014). The methods are consistent with assumptions of diagnostic classification modeling and yield evidence to support the argument for internal consistency of the program for each level of reporting. The results indicate high levels of reliability for the individual linkage level mastery classifications, as well as for all levels of aggregation for which results are reported (i.e., EE, conceptual area, subject, and overall performance level). Because the reliability results depend on the model used to calibrate and score the assessment, any changes to the model or evidence obtained when evaluating model fit also affect reliability results. As with any selected methodology for evaluating reliability, the current results assume that the model and model parameters used to score DLM assessments are correct. However, unlike other traditional measures of reliability that often require unattainable assumptions about equivalent test forms, the simulation method described in this chapter provides a replication of the same test administration process used in the operational assessment, which provides a rigorous evaluation of the variation in student results across simulated repeated assessment administrations.

9. Training and Professional Development

The Dynamic Learning Maps® (DLM®) Alternate Assessment System provides comprehensive support and training to state education agency staff and local educators. The type of support provided for local educators is twofold: DLM Required Test Administrator Training, and optional professional development for instruction. DLM Required Test Administrator Training provides test administrators with both the context and practical knowledge of the assessment system design, administration, and security practices to administer the assessment with fidelity. All required training is therefore aligned with the *Test Administration Manual* (DLM Consortium, 2021a).[64] The purpose of the professional development component is to provide professional learning opportunities to support instructional practices for test administrators and educators of the target population of students who participate in DLM assessments.

This chapter describes the training that was offered in 2021–2022 for state and local education agency staff as well as the DLM Required Test Administrator Training, which was originally developed for the first operational administration of the DLM assessments in 2014–2015. Since the initial release of training materials, numerous updates have been made, which are reflected in the description of the training that was available in 2021–2022. Additionally, this chapter describes the participation rates and evaluation results for the optional instructional professional development modules.

9.1. Training for State Education Agency Staff

State education agency staff are integral to the implementation of the DLM assessment system. While there is no formal comprehensive training program for this audience, the staff have opportunities to participate in training designed for local education agency staff and test administrators. Throughout the year, state education agency staff also receive instruction during regularly scheduled meetings and through written documentation from DLM staff on state-level support topics, such as monitoring test administrators' completion of required training, viewing and editing data in Educator Portal, and using data extracts from Educator Portal to monitor assessment administration.

9.2. Training for Local Education Staff

In 2021–2022, as in prior years, three local education staff roles supported implementation of the DLM System. These roles were typically held by one or more district-level staff members, but in some cases were fulfilled at the building level.

- The assessment coordinator oversees the assessment process, including managing staff roles and responsibilities, developing and implementing a comprehensive training plan, developing a schedule for test implementation, monitoring and supporting test preparations and administration, and developing a plan to facilitate communication with parents/guardians and staff.
- The data manager oversees educator, student, and roster data.
- Technology personnel verify that network and testing devices are prepared for test administration.

In August 2021, a prerecorded training video was posted to each state's page on the DLM website. This training applied to assessment coordinators, data managers, and technology personnel. It discussed the checklist tasks from each role's manual and explained how the roles work together to successfully

[64] See Chapter 4 of this manual for a complete description of assessment administration practices.

implement the DLM assessment in their district. A training flyer was provided to state education agencies to distribute and promote this training opportunity. The flyer included information regarding two follow-up questions and answer opportunities, whereby DLM staff answered questions via live chat.

In January 2022, another prerecorded training video was posted to each state's page on the DLM website. This video was provided for district superintendents and building principals. It provided an overview of the DLM assessment and explained the activities that should be ongoing in their buildings and districts to prepare for and monitor the spring assessment. Follow-up question and answer opportunities regarding this training were also offered via live chat.

9.3. Required Training for Test Administrators

Training is required annually for all educators who will administer the DLM alternate assessment to students. Test administrators are not allowed access to their student's login credentials for the assessment until all required training components are completed.

The DLM Required Test Administrator Training is conducted in Moodle. First-time test administrators are enrolled in a training course for new test administrators that takes approximately three hours to complete. Test administrators who completed the training for the previous school year are enrolled in a course for returning test administrators that takes approximately an hour to complete. Each trainee has a secure login to the Moodle training site. Directions for accessing and completing the training are provided on each state's page on the DLM website.

The training for new test administrators consists of four modules, each with a post-test. Trainees have to pass each post-test with a score of at least 80% to receive the certificate of completion and gain access to students' login credentials for the assessment. The training for new test administrators is offered in two formats: self-directed, whereby the trainee navigates through the training course independently and completes the post-tests, or facilitated, whereby a trained facilitator presents or guides trainees through the training and then directs them to complete the post-tests on their own in Moodle.

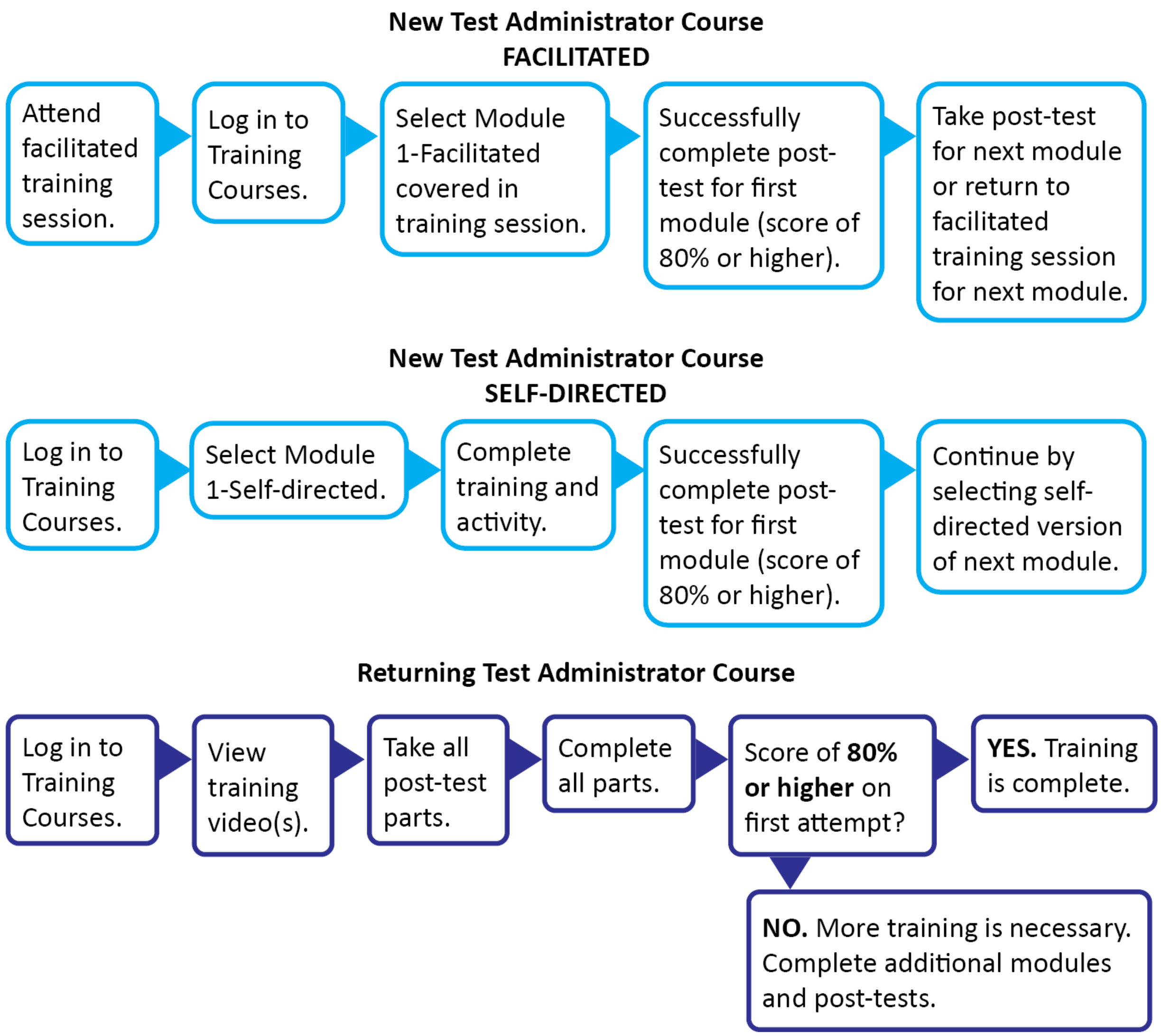
For returning test administrators, the training consists of a single module that reviews the most salient aspects of the DLM assessment and test administration procedures. After the review, a four-part post-test is required. Each part corresponds to one of the four modules for new test administrators. Any returning test administrator who does not pass a part of their post-test with a score of at least 80% on the first attempt is redirected to the training for new test administrators for remediation before continuing the returning test administrator course. The training for returning test administrators is available in the self-directed format only.

Some states include one or more DLM professional development modules in their required test administrator training courses. When included, these additional modules are accessed in Moodle and are required to be completed to generate the certificate of completion, unless the state asked that the additional modules be included as optional modules. For a summary of participation in the professional development modules as part of the required training, see section 9.4.1.1 of this chapter. Some states also include one or two optional videos in their required test administrator training courses. These videos are posted on the DLM website but were accessed in Moodle if included in the required training. They do not have a post-test. Completing them is not required to generate the certificate of completion.

Figure 9.1 illustrates the differences between the two training formats for new test administrators.

Figure 9.1

Required Training Processes Flows for Facilitated and Self-Directed Training



9.3.1. Facilitated Training

The facilitated modules for new test administrators are intended to use with groups. This version of the modules is designed to meet the need for face-to-face training without requiring a train-the-trainers approach or requiring the facilitator to have deep expertise in the subject matter. Each state determines its own policy guidance regarding who serves as facilitators. Examples of individuals who serve as facilitators include district- and building-level test coordinators, district special education coordinators, instructional coaches, lead educators, state education agency staff, and trainers from regional education agencies that are responsible for professional development.

Facilitators are provided an agenda, a detailed guide, handouts, videos, and instructions to facilitate a meaningful face-to-face training. Facilitators play the DLM-produced videos and implement learning activities as described in their facilitator directions. Facilitators who wish to add to the training contents or deliver the content themselves rather than via video also have access to the PowerPoint slides and transcripts.

Facilitators for these sessions prepare for the training by reviewing all videos and all sections of the *Test Administration Manual* (DLM Consortium, 2021a) addressed in the training. States also recommend (and some require) facilitators to complete the training requirements themselves; facilitators who are also test administrators are required to pass the post-tests. Facilitators are asked to ensure that participants have Educator Portal and Moodle accounts and access to them prior to the facilitated training session. Their responsibilities include setting up the training area with equipment, delivering the facilitated training modules, and directing users to return all equipment. Finally, facilitators direct test administrators to take each module post-test in Moodle with support from the *Guide to DLM Required Test Administrator Training* (DLM Consortium, 2021c) for details and access procedures. Facilitated training is flexibly structured so post-tests could be taken on-site during training sessions (e.g., in a computer lab) or independently after the training session was complete.

9.3.2. Self-Directed Training

The self-directed modules are designed to meet the needs of educators who are unable to attend facilitated sessions and those who otherwise need access to on-demand training. Some states and districts do not provide a facilitated training, so test administrators in those areas have to take the self-directed course. Self-directed modules combine videos, text, and online learning activities to engage educators with a range of content, strategies, and supports, as well as the opportunity to reflect upon and apply what they learned. The videos are identical to those used in facilitated training. The post-test for each module is the same as those for trainees who attend a facilitated training.

The self-directed training is completed entirely within Moodle with support from the *Guide to DLM Required Test Administrator Training* (DLM Consortium, 2021c) for detail and access procedures, including directions for completing all post-tests.

9.3.3. Training Content

Several pertinent resources provided on the DLM website are referenced throughout the training modules, including the *Test Administration Manual*, *Accessibility Manual*, *Educator Portal User Guide*, and the *Guide to Practice Activities and Released Testlets*. The individual modules that made up the training are described in more detail in the following sections.

9.3.3.1. Module 1: Overview of the Dynamic Learning Maps Alternate Assessment System

Module 1 of the test administrator training provides an overview of the DLM system components. The focus is on familiarizing trainees with DLM vocabulary, such as Essential Elements, linkage levels, assessment blueprints, claims and conceptual areas, and testlets. The security standards are also addressed.

9.3.3.2. Module 2: Understanding and Delivering Testlets

In Module 2, participants are shown examples of DLM testlet items. Testlet types (computer-delivered and educator-administered) are also explained. Accessibility supports and practices allowed and not allowed are described.

9.3.3.3. Module 3: Test Administration and Scoring

Module 3 focuses on the tasks test administrators complete in Educator Portal, such as verifying student data and completing students' First Contact surveys and Personal Needs and Preferences Profiles. Then, the test administration process is explained, including how to access Testlet Information Pages and track testing completion. Finally, Module 3 provides an overview of the DLM scoring model.

9.3.3.4. Module 4: Preparing to Administer the Assessment

Module 4 concludes the training by explaining test administrator responsibilities. Considerations are discussed for technology preparations, utilizing the practice activities and released testlets, scheduling assessment sessions, and devising back-up plans to ensure all students complete the assessment before the close of the assessment window. Module 4 concludes with information about the DLM professional development modules offered at https://dlmpd.com.

9.3.4. Completion of All Modules

Each of the four required training modules for new test administrators includes a post-test. Participants are required to complete each post-test with at least 80% accuracy, and they have to score at least an 80% to be able to move on to the next module. Participants are allowed to retake the post-test as many times as necessary to achieve a passing score. Feedback is provided in Moodle for any incorrect responses. Trainees always have access to the module's contents, and the recommendation is that they review the module's contents before attempting the post-test again. Whether participants complete the self-directed or facilitated version of the required training, post-tests are completed independently in Moodle. Details of participants' progress are made available to state educational agencies through a data extract in Educator Portal.

9.4. Instructional Professional Development

The DLM professional development system is built to support educators in their efforts to teach English language arts and mathematics to their students who participate in the DLM System. In addition to instructional support, the instructional professional development modules also teach educators about the DLM System. While the modules were originally intended for educators who administer DLM assessments, demographic information suggests that preservice educators, related service providers, parents, and others also accessed and completed the modules.

The professional development system is built in WordPress, an open-source website content management system. The professional development modules and instructional support materials are available for anyone's use at https://dlmpd.com or through a direct link from the DLM website. These DLM professional development modules address instruction in English language arts and mathematics. The modules also address processes educators can apply to create Individual Education Programs that are aligned with the DLM Essential Elements and supports they can provide to address the communication needs of the students they teach. Finally, the modules help them understand the components of the DLM assessment system more completely.

To support state and local education agencies in providing continuing education credits to educators who complete the modules, each module also includes a time-ordered agenda, learning objectives, and biographical information regarding the faculty who developed the training modules. There are a total of 51 modules, which are described in section 9.4.1.2 of this chapter.

The 51 modules are available in both self-directed and facilitated formats. The self-directed modules are available online, on-demand. The interactive modules include a combination of video-based content, embedded activities, and, for participants who would like to receive a certificate documenting their successful completion of the module, a five-item pre- and post-test. These certificates are sent directly to each participant's email when they score 80% or higher on the post-test.

Modules in the facilitated format were created for groups and by individuals who prefer to learn by reading the contents rather than interacting with video and other package contents. The modules are organized on the DLM professional development website[65] by claim and conceptual area and alphabetically.[66] The modules have also been packaged into groups[67] that are comprised of 5- and 10-hour focused packages for schools and school systems to use across the school year. These packages include the following:

- **Foundations of Instruction in DLM.** This package provides educators with five hours of foundational content regarding the DLM System and the students who participate in DLM alternate assessments. Divided across five modules, the content can be split across five sessions or combined to meet local needs. The five modules are (1) Who are Students with Significant Cognitive Disabilities? (2) College and Career Readiness Standards: An Overview, (3) DLM Essential Elements Overview, (4) DLM Claims and Conceptual Areas, and (5) Individual Education Programs Linked to the DLM Essential Elements.
- **Getting Started with Students Working at an Initial Precursor Linkage Level.** This package provides educators with 10 hours of important content and pedagogical knowledge pertaining to students with the most complex disabilities who are working at an Initial Precursor linkage level in the DLM alternate assessments in English language arts and mathematics. Divided across 10 modules, the content can be split across 10 sessions or combined to meet local needs. The 10 modules are (1) Beginning Communicators, (2) Symbols, (3) DLM Core Vocabulary and Communication, (4) Speaking and Listening, (5) Shared Reading, (6) Writing with Alternate Pencils, (7) Emergent Writing, (8) Predictable Chart Writing, (9) Unitizing, and (10) Forms of Number.
- **From Shared Reading to Reading Comprehension.** This package provides educators with five hours of professional development focused on reading instruction, beginning with supporting engagement and interaction during shared reading and moving to reading and listening with comprehension. Divided across five modules, the content can be split across five sessions or combined to meet local needs. The five modules are (1) Shared Reading, (2) Teaching Text Comprehension: Anchor-Read-Apply, (3) Generating Purposes for Reading, (4) DR-TA and Other Text Comprehension Approaches, and (5) Supporting Participation in Discussion.

[65] https://dlmpd.com
[66] For more information on DLM claims and conceptual areas, see Chapter 2 of this manual.
[67] https://www.dlmpd.com/professional-development-packages/

- **From Emergent Writing to Written Composition.** This package provides educators with 10 hours of professional development focused on writing instruction, beginning with supporting access to writing tools and early engagement in writing, and moving through writing for various purposes and audiences. Divided across 10 modules, the content can be split across five sessions or combined to meet local needs. The 10 modules are (1) Universal Design for Learning, (2) Writing with Alternate Pencils, (3) Emergent Writing, (4) Writing: Text Types and Purposes, (5) Writing Information and Explanation Texts, (6) Writing: Production and Distribution, (7) Writing: Getting Started with Narrative Writing, (8) Writing: Getting Started in Writing Arguments, (9) Writing: Research and Range of Writing, and (10) Strategies and Formats for Presenting Ideas.
- **Moving Students from Initial to Distal Precursor Linkage Levels in Mathematics.** This package provides educators with 10 hours of professional development focused on emergent mathematics understandings for students with complex disabilities participating in the DLM alternate assessment in mathematics at the Initial Precursor or Distal Precursor levels. Divided across 10 modules, the content can be split across five sessions or combined to meet local needs. The 10 modules are (1) Unitizing, (2) Forms of Number, (3) Composing and Decomposing Shapes and Area, (4) Composing, Decomposing, and Comparing Numbers, (5) Counting and Cardinality, (6) The Power of Ten-Frames, (7) Basic Geometric Shapes and Their Attributes, (8) Measuring and Comparing Lengths, (9) Time and Money, and (10) Patterns and Sequence.

In addition to the 51 modules, the instructional professional development site provides instructional resources for educators. These include DLM Essential Element unpacking documents; extended descriptions of the Initial and Distal Precursor linkage levels and how they relate to grade-level Essential Elements; links to dozens of texts that are at an appropriate level of complexity for students who take DLM assessments and are linked to the texts that are listed in Appendix B of the Common Core State Standards[68]; vignettes that illustrate shared reading with students with the most complex needs across the grade levels; supports for augmentative and alternative communication for students who do not have a comprehensive, symbolic communication system; alternate pencils for educators to download and use with students who cannot use a standard pen, pencil, or computer keyboard; and links to Pinterest boards and other online supports.

9.4.1. Professional Development Participation and Evaluation

There are two ways in which test administrators and educators may complete professional development modules: required test administrator training or optional professional development. As described in section 9.3, some states require professional development modules to be completed as part of the required test administrator training. States can require certain modules be completed by new test administrators, returning test administrators, or both. Test administrators completing professional development modules as part of the required test administrator training access the modules through the Moodle training site where the rest of the required test administration training is located. The second way in which professional development modules are completed is through the DLM professional development website[69]. The modules on the professional development website may be accessed by anyone and can be completed at any time. Additionally, participants completing modules on the professional development website are

[68] http://www.corestandards.org/assets/Appendix_B.pdf
[69] https://dlmpd.com

administered a short evaluation survey following the module. With the exception of the evaluation survey, the content of the modules is identical between the required test administrator training and the professional development website.

9.4.1.1. Required Professional Development Participation

A total of nine states required at least one professional development module as part of their required test administrator training. Table 9.1 shows the number of modules required, by state, for new and returning test administrators, as well as the total number of modules completed. In total, 21,888 professional development modules were completed by 4,431 new and 3,033 returning test administrators as part of the required training.

Table 9.1

Number of Professional Development Modules Completed as Part of the Required Test Administrator Training

State	New test administrators		Returning test administrators		Total modules completed
	Required modules	Test Administrators	Required modules	Test Administrators	
Alaska	1	84	1	130	214
Delaware	2	359	—	—	718
Maryland	4	1,466	—	—	5,864
New Hampshire	1	152	—	—	152
New Mexico	3	824	—	—	2,472
Oklahoma	5	649	5	1,305	9,770
Rhode Island	1	110	1	190	300
West Virginia	2	203	1	364	770
Wisconsin	1	584	1	1,044	1,628

Table 9.2 shows which modules were required for new and returning test administrators across all states choosing to include professional development modules in the required training. For example, the *DLM Essential Elements* module was required for new test administrators in four states and was required for returning test administrators in two states.

Table 9.2

Professional Development Modules Selected for Inclusion in Required Test Administrator Training

Module	States requiring for new test administrators	States requiring for returning test administrators	Total modules completed
DLM Claims and Conceptual Areas	1	—	1,466
DLM Essential Elements	4	2	3,993
Effective Instruction in Mathematics	3	2	3,987
Individual Education Programs Linked to the DLM Essential Elements	4	2	3,437
Principles of Instruction in English Language Arts	3	2	5,048
Universal Design for Learning	1	—	824
Who are Students With Significant Cognitive Disabilities?	4	1	3,133

9.4.1.2. Optional Professional Development Participation

Table 9.3 shows the number of individuals who completed optional professional development modules through the DLM professional development website[70], as well as the total number of test administrators from each state who had a student assigned to them for the DLM assessment. In total, 5,987 modules were completed in the self-directed format from August 1, 2021, to July 31, 2022. Since the first module was launched in the fall of 2012, a total of 77,061 modules have been completed on the professional development website.

[70] https://dlmpd.com

Table 9.3

*Number of Self-Directed Modules Completed in 2021–2022 by Educators in DLM States and Other Localities (*N *= 5,987)*

State	Participants	DLM test administrators	Total modules completed
Alaska	20	180	42
Arkansas	24	785	62
Colorado	250	1,161	882
Delaware	83	259	201
Illinois	89	3,254	208
Iowa	78	939	358
Kansas	138	895	619
Maryland	125	1,405	448
Missouri	38	1,564	381
New Hampshire	9	328	28
New Jersey	79	2,928	456
New Mexico	15	643	100
New York	84	5,550	271
North Dakota	12	265	40
Oklahoma	16	1,484	121
Pennsylvania	53	4,380	525
Rhode Island	40	248	78
Utah	81	782	287
West Virginia	10	489	26
Wisconsin	178	1,543	846
Non-DLM state and other locations	4	—	8

Note. Participant counts may include individuals who are not educators or test administrators (e.g., pre-service educators).

To evaluate educator perceptions of the utility and applicability of the modules, DLM staff asked educators to respond to a series of evaluation questions upon completion of each self-directed module. Four questions asked about importance of content, whether new concepts were presented, the utility of the module, and whether educators planned to use what they learned. Educators responded using a four-point scale ranging from *strongly disagree* to *strongly agree*. During the 2021–2022 year, educators completed the evaluation questions 86% of the time. The responses were consistently positive, as illustrated in Table 9.4. Across all modules, 82% of respondents either agreed or strongly agreed with each statement.

To evaluate the consistency in the ratings for each module, we calculated Cronbach's (1951) alpha from the four items for each module using all ratings from fall 2012 through the 2021–2022 year. Across all modules, alpha ranged from .85 to .96 with an average value of .92, suggesting high internal consistency in responses.

Table 9.4

Response Rates and Rate of Agree *or* Strongly Agree *on 2021–2022 Self-Directed Module Evaluation Questions*

Module	Total modules completed (*n*)	Response rate	The module addressed content that is important for professionals working with SWSCDs. (%)	The module presented me with new ideas to improve my work with SWSCDs. (%)	Completing this module was worth my time and effort. (%)	I intend to apply what I learned in the module to my professional practice. (%)
Algebraic Thinking	62	95.2	90.3	87.1	88.7	88.7
Basic Geometric Shapes and Their Attributes	71	93.0	83.1	80.3	74.6	81.7
Beginning Communicators	251	77.3	74.9	73.3	73.7	74.9
Calculating Accurately With Addition	82	90.2	89.0	86.6	86.6	86.6
Calculating Accurately With Division	41	95.1	92.7	90.2	92.7	92.7
Calculating Accurately With Multiplication	44	88.6	84.1	86.4	86.4	79.5
Calculating Accurately With Subtraction	48	91.7	91.7	91.7	89.6	89.6
College and Career Readiness Standards	177	87.0	81.9	79.1	79.1	81.9
Composing and Decomposing Shapes and Area	46	91.3	89.1	84.8	87.0	84.8

Table 9.4

Response Rates and Rate of Agree *or* Strongly Agree *on 2021–2022 Self-Directed Module Evaluation Questions (continued)*

Module	Total modules completed (*n*)	Response rate	The module addressed content that is important for professionals working with SWSCDs. (%)	The module presented me with new ideas to improve my work with SWSCDs. (%)	Completing this module was worth my time and effort. (%)	I intend to apply what I learned in the module to my professional practice. (%)
Composing, Decomposing, and Comparing Numbers	108	93.5	90.7	92.6	88.9	90.7
Counting and Cardinality	115	93.0	89.6	89.6	87.0	91.3
DLM Claims and Conceptual Areas	191	90.1	87.4	85.3	84.8	86.4
DLM Core Vocabulary and Communication	236	85.6	83.9	83.5	83.9	84.3
DLM Essential Elements	400	86.5	82.5	81.8	81.2	82.8
DR-TA and Other Text Comprehension Approaches	153	80.4	77.1	76.5	76.5	77.8
Effective Instruction in Mathematics	148	85.1	83.8	83.8	81.8	83.8
Emergent Writing	301	83.7	81.1	81.4	81.4	82.1
Exponents and Probability	24	91.7	91.7	91.7	91.7	87.5
Forms of Number	90	82.2	81.1	77.8	76.7	78.9

Table 9.4

Response Rates and Rate of Agree *or* Strongly Agree *on 2021–2022 Self-Directed Module Evaluation Questions (continued)*

Module	Total modules completed (*n*)	Response rate	The module addressed content that is important for professionals working with SWSCDs. (%)	The module presented me with new ideas to improve my work with SWSCDs. (%)	Completing this module was worth my time and effort. (%)	I intend to apply what I learned in the module to my professional practice. (%)
Fraction Concepts and Models Part I	25	88.0	84.0	80.0	84.0	84.0
Fraction Concepts and Models Part II	27	92.6	92.6	85.2	88.9	88.9
Functions and Rates	19	94.7	94.7	94.7	94.7	94.7
Generating Purposes for Reading	149	88.6	85.9	85.9	84.6	84.6
Individual Education Programs Linked to the DLM Essential Elements	198	80.3	75.3	73.7	69.7	75.3
Measuring and Comparing Lengths	28	85.7	85.7	82.1	78.6	82.1
Organizing and Using Data to Answer Questions	24	87.5	83.3	87.5	87.5	87.5
Patterns and Sequence	29	96.6	96.6	96.6	93.1	89.7
Perimeter, Volume, and Mass	26	92.3	88.5	92.3	84.6	88.5
Place Value	75	74.7	72.0	70.7	66.7	72.0

Table 9.4

Response Rates and Rate of Agree *or* Strongly Agree *on 2021–2022 Self-Directed Module Evaluation Questions (continued)*

Module	Total modules completed (*n*)	Response rate	The module addressed content that is important for professionals working with SWSCDs. (%)	The module presented me with new ideas to improve my work with SWSCDs. (%)	Completing this module was worth my time and effort. (%)	I intend to apply what I learned in the module to my professional practice. (%)
Predictable Chart Writing	67	85.1	83.6	83.6	83.6	83.6
Principles of Instruction in English Language Arts	140	82.1	80.0	79.3	78.6	80.7
Properties of Lines and Angles	18	88.9	88.9	88.9	88.9	83.3
Shared Reading	423	83.9	80.9	79.7	78.5	81.1
Speaking and Listening	75	77.3	77.3	76.0	77.3	77.3
Strategies and Formats for Presenting Ideas	75	85.3	82.7	82.7	82.7	82.7
Supporting Participation in Discussions	121	86.0	75.2	74.4	71.9	71.1
Symbols	82	78.0	65.9	64.6	64.6	65.9
Teaching Text Comprehension: Anchor-Read-Apply	191	83.2	81.7	81.2	80.6	81.2
The Power of Ten-Frames	65	92.3	89.2	87.7	89.2	87.7
Time and Money	74	87.8	85.1	83.8	85.1	85.1

Table 9.4

Response Rates and Rate of Agree *or* Strongly Agree *on 2021–2022 Self-Directed Module Evaluation Questions (continued)*

Module	Total modules completed (*n*)	Response rate	The module addressed content that is important for professionals working with SWSCDs. (%)	The module presented me with new ideas to improve my work with SWSCDs. (%)	Completing this module was worth my time and effort. (%)	I intend to apply what I learned in the module to my professional practice. (%)
Unitizing	76	89.5	89.5	86.8	85.5	89.5
Units and Operations	27	92.6	92.6	88.9	92.6	88.9
Universal Design for Learning	221	87.8	85.1	84.6	81.4	85.5
Who are Students With Significant Cognitive Disabilities?	647	91.8	89.6	82.7	86.1	87.3
Writing Information and Explanation Texts	58	77.6	77.6	77.6	77.6	77.6
Writing With Alternate Pencils	163	85.9	84.0	84.7	84.0	84.7
Writing: Getting Started With Narrative Writing	51	84.3	84.3	84.3	84.3	84.3
Writing: Getting Started Writing Arguments	47	78.7	78.7	78.7	76.6	78.7
Writing: Production and Distribution	47	85.1	85.1	85.1	85.1	85.1
Writing: Research and Range of Writing	57	89.5	84.2	84.2	82.5	84.2

Table 9.4

Response Rates and Rate of Agree *or* Strongly Agree *on 2021–2022 Self-Directed Module Evaluation Questions (continued)*

Module	Total modules completed (*n*)	Response rate	The module addressed content that is important for professionals working with SWSCDs. (%)	The module presented me with new ideas to improve my work with SWSCDs. (%)	Completing this module was worth my time and effort. (%)	I intend to apply what I learned in the module to my professional practice. (%)
Writing: Text Types and Purposes	74	81.1	81.1	81.1	81.1	81.1
Total	*5,987*	*86.2*	*83.3*	*81.8*	*81.4*	*82.8*

Note. SWSCDs = students with significant cognitive disabilities.

9.5. Conclusion

The DLM System makes training and instructional professional development modules available to test administrators and educators as well as state and local education agency staff. The Required Test Administrator Training is designed to ensure that test administrators understand the DLM System in order to administer the assessment with fidelity. Complementary to the required training are the professional development modules, which provide participants learning opportunities to support instructional practices aligned with content measured by DLM assessments. In 2021–2022, participants provided consistently positive feedback regarding the importance and relevance of the professional development modules.

10. Validity Argument

The Dynamic Learning Maps® (DLM®) Alternate Assessment System is based on the core belief that all students should have access to challenging, grade-level academic content. Therefore, DLM assessments provide students with the most significant cognitive disabilities the opportunity to demonstrate what they know and can do. This technical manual provides evidence evaluating the propositions and assumptions that undergird the assessment system as described in the DLM Theory of Action (Figure 10.1; for more information, see description in Chapter 1).

This chapter synthesizes the evidence provided in this technical manual and places it within a validity framework to assess the program's overall success at producing results that reflect their intended meaning and can be used for their intended purposes. In addition, we discuss future research studies as part of ongoing and iterative processes of continuous improvement.

10.1. Validity Framework

Validation is the process of evaluating the evidence and theory presented in the overall validity argument. All aspects of the validity argument must be carefully evaluated (Lissitz, 2009; Sireci, 2009). The purpose of the assessment and support for intended use of results is critical to the overall validity argument because it is indicative of the model from which the assessment was originally designed (Mislevy, 2009). It follows, therefore, that the evidence collected throughout the entire development, administration, and reporting processes should point to a clear and persuasive link between the assessment purpose and the intended uses and interpretations of the results and overall project outcomes. Cohesion between what can be observed (e.g., student responses to assessment tasks) and what must be inferred (e.g., student achievement in the subject) must inform the validity and interpretative arguments (Kane, 2006). It is equally important that the dimensions and organization of the overall validity argument include not only the content sampled and procedural bases of the assessment, but also evidence for the underlying construct being assessed, any assessment elements that are irrelevant to the construct, and the relative importance of the consequences of the resulting scores (Linn, 2009; Messick, 1989).

The DLM program adopted an argument-based approach to assessment validation. The validity argument process began by specifying, with the DLM Governance Board, the intended, supported, and unsupported uses of DLM assessment results, consistent with expectations described in the *Standards for Educational and Psychological Testing* (*Standards* hereafter), which are the professional standards broadly used to inform development and evaluation of educational assessments (AERA et al., 2014). We followed this with a three-tiered approach to validation, which included specification of 1) a Theory of Action with defining statements[71] in the validity argument that must be in place to achieve the goals of the system and the chain of reasoning between those statements, 2) an interpretive argument defining propositions that must be examined to evaluate each component of the Theory of Action, and 3) validity studies to evaluate each proposition in the interpretive argument.

We also drew on the organization of evidence according to the *Standards* (AERA et al., 2014). The *Standards* define five types of evidence for validity arguments, including evidence based on content, response process, internal consistency, relation to other variables, and consequences. Evidence collected

[71] The term "statement" is used here to mean a claim within the overall validity argument. The term "claim" is reserved in this technical manual for use specific to content claims (see Chapter 3 of this manual).

for propositions in the DLM validity argument is classified according to type.

10.2. Intended Uses

In 2013, the DLM Governance Board determined uses of DLM assessment results. Table 10.1 shows intended, supported, and unsupported uses of DLM assessment results.

Table 10.1

Uses of Dynamic Learning Maps Assessment Results

Type of scoring	Intended uses	Supported uses (optional—states' decision to use and responsibility to provide evidence)	Uses that are NOT supported or intended
Mastery determinations obtained throughout the year from optional instructionally embedded assessments	1. Instructional planning, monitoring, and adjustment	1. One source of information for evaluations of educator and principal effectiveness	1. Determining disability eligibility 2. Placement 3. Retention 4. Inclusion in state accountability model to evaluate school and district performance (unless supported by state-conducted research) 5. Direct comparisons with results on general education assessments
Summative reporting	1. Reporting achievement within the instructed content aligned to grade-level content standards, to a variety of audiences, including educators and parents 2. Inclusion in state accountability model to evaluate school and district performance 3. Planning instructional priorities and program improvement for following school year	1. One source of information for evaluations of educator and principal effectiveness 2. Graduation (in states that use AA-AAS as exit exam)	1. Determining disability eligibility 2. Placement 3. Retention 4. Graduation (sole source of evidence) 5. Direct comparisons with results on general education assessments

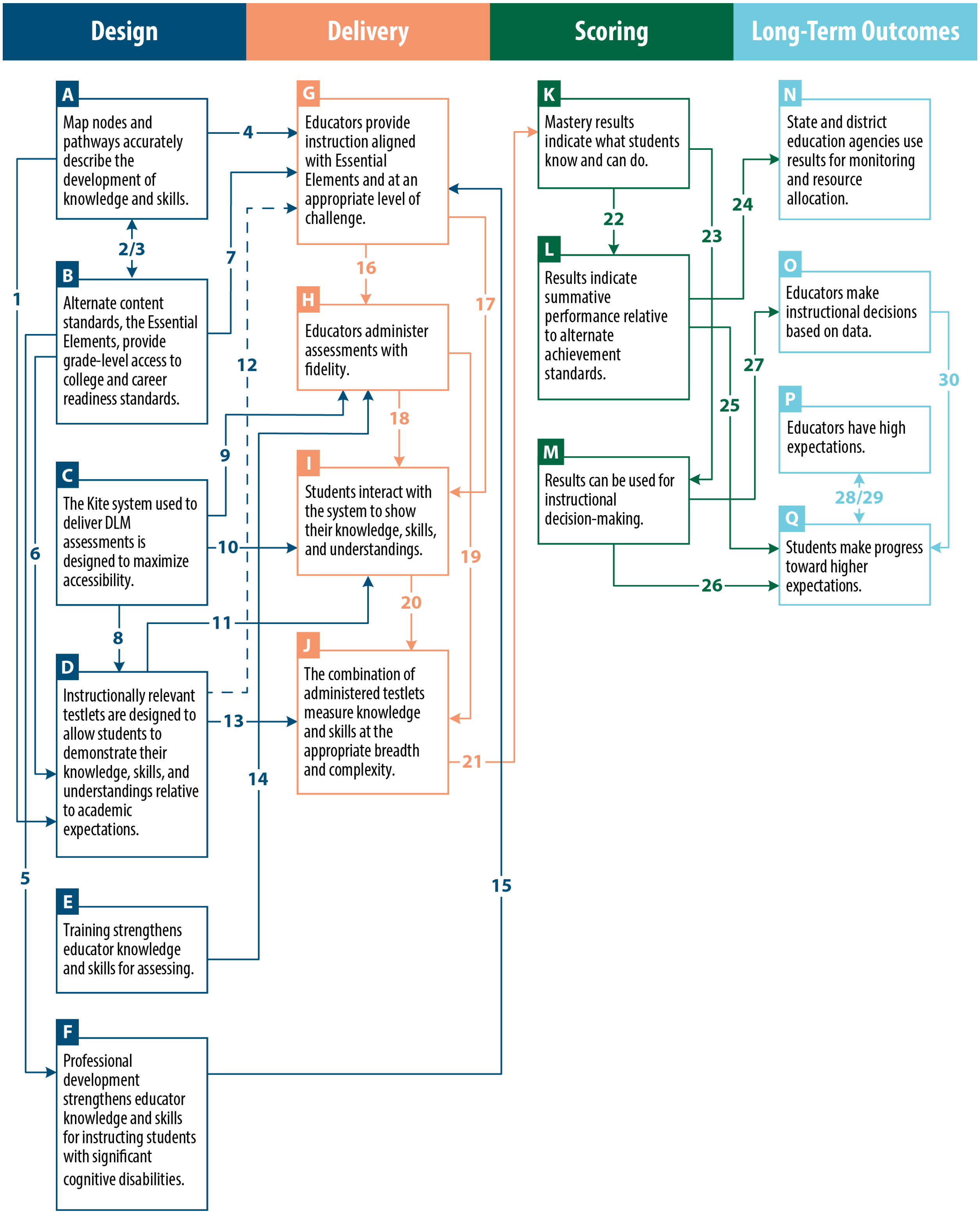
Note. AA-AAS = Alternate Assessment based on Alternate Achievement Standards.

10.3. Theory of Action

The DLM Theory of Action includes statements about the assessment program spanning from its design, delivery, and scoring, to long-term intended outcomes. While validity arguments typically center on validation of scores for intended uses, the DLM Theory of Action also specifies intended long-term outcomes associated with use of the system (consistent with the definition given by the National Council on Measurement in Education [NCME], 2018). The Theory of Action is depicted in Figure 10.1. Letters are assigned to each statement. Connections between statements, referred to as inputs in this chapter, are designated by numbers. Dashed lines represent connections that are present when the optional instructionally embedded assessments are utilized. For more information on the DLM Theory of Action, see Chapter 1 of this manual.

Figure 10.1

Dynamic Learning Maps Theory of Action Diagram



10.4. Propositions and Validity Evidence

The three-tiered approach to validation relies on propositions underlying the statements in the Theory of Action and evidence evaluating those statements. For each statement in the Theory of Action, we summarize the underlying propositions and associated procedural and empirical evidence informing evaluation of the proposition, indicate the type of evidence (e.g., content, response process, consequences), and identify the chapters in this manual containing a full description of the evidence. We also describe how connections in the Theory of Action affect the extent to which statements are supported and defensible. We retain the organizational structure of the Theory of Action when describing the statements, propositions, and evidence, organizing them according to elements of design, delivery, scoring, and long-term outcomes.

10.4.1. Design

The design components of the DLM assessment system include the learning maps, Essential Elements (EEs), Kite® Suite, assessments, training, and professional development. Specifically, we evaluate the following six statements related to the design of DLM assessments.

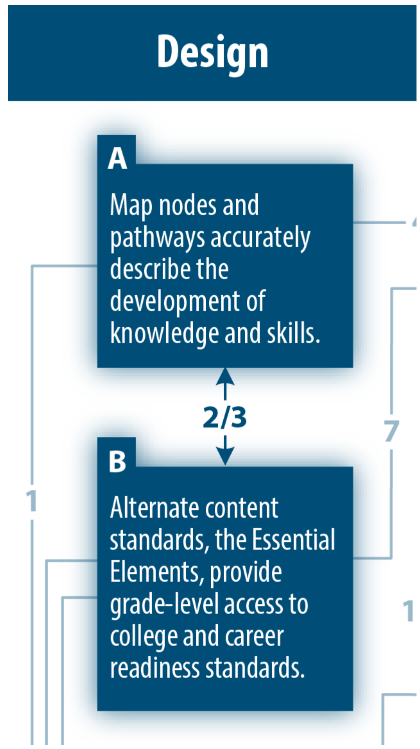
A. Map nodes and pathways accurately describe the development of knowledge and skills.
B. Alternate content standards, the Essential Elements, provide grade-level access to college and career readiness standards.
C. The Kite system used to deliver DLM assessments is designed to maximize accessibility.
D. Instructionally relevant testlets are designed to allow students to demonstrate their knowledge, skills, and understandings relative to academic expectations.
E. Training strengthens educator knowledge and skills for assessing.
F. Professional development strengthens educator knowledge and skills for instructing and assessing students with significant cognitive disabilities.

10.4.1.1. A: Map Nodes and Pathways Accurately Describe the Development of Knowledge and Skills

Learning maps are one of the critical content structures in the DLM assessment system. Together with EEs, they define the knowledge, skills, and understandings measured by the assessments, as shown in Figure 10.2. Map nodes span early foundational skills to college and career readiness skills, with connections between nodes defining how skills develop over time. The nodes and connections provide all students with an access point for learning rigorous grade-level academic content.

Figure 10.2

Theory of Action Inputs for Learning Maps and Essential Elements



Three propositions describe the accuracy of map nodes and pathways in the DLM assessment system. Table 10.2 summarizes the propositions and evidence evaluating the accuracy of nodes and pathways in the learning maps. Because EEs serve as an input into map design (as shown in Figure 10.2), some of the evidence in the table relies on evidence collected for EE development. Rather than restate the evidence, the overlap is noted with a dagger (†) where applicable. We applied this principle throughout the chapter when describing evidence derived from inputs in the Theory of Action.

Table 10.2

Propositions and Evidence for Learning Maps

Proposition	Procedural evidence	Empirical evidence	Type	Chapter(s)
Nodes are specified at the appropriate level of granularity and are distinct from other nodes	Description of map development process, mini-maps, Essential Element Concept Maps (EECMs)†	External review ratings	Content	2
Nodes for linkage levels are correctly prioritized and are adequately spaced within breadth of map	Description of map development process, mini-maps, EECMs†		Content	2, 3
Nodes are sequenced according to acquisition order, given population variability	Description of map development process, mini-maps, EECMs†	External review ratings, modeling analyses, alignment study (linkage level vertical articulation)	Content	2, 3

† Relies on evidence from Theory of Action input statements, as shown in Figure 10.2.

For the map to accurately describe how students develop knowledge, skills, and understandings, nodes must be specified at the right grain size (i.e., neither too large nor too small) and be distinct from other nodes. Documentation describes how nodes were developed, including extensive literature syntheses and rounds of internal review to evaluate node size and distinction from other nodes. Essential Element Concept Maps (EECMs) and mini-maps available in Educator Portal depict the final selection of nodes assessed at each linkage level. Content and special education experts provided ratings on whether nodes were of correct size and distinct from other nodes. In the limited instances in which nodes were rated as being at the incorrect size or indistinct from other nodes, map developers reviewed the feedback and adjusted the map structure. Together, these sources of evidence support the proposition that nodes are of the appropriate grain size and are distinct from one another.

To demonstrate that the map is accurately specified, the nodes measured by each linkage level should be correctly prioritized from among the full breadth of nodes in the map and be adequately spaced throughout the map structure (i.e., neither bunched together too closely nor spread too far apart). Documentation describes the process that map developers used to select linkage level nodes, which included rounds of internal review examining node selection and spacing within the map. The EECMs and mini-maps in Educator Portal depict the final selection of nodes assessed at each linkage level. Because of the size and complexity of the full map, external review did not evaluate the selection of nodes prioritized for linkage levels. While additional evidence could be collected for this proposition, the current evidence provides moderate support for nodes being correctly prioritized and spaced through the map.

Finally, for the map to be accurately specified, nodes must be correctly sequenced. Documentation describes how map connections were developed, including extensive literature syntheses and rounds of internal review evaluating the correct ordering of nodes within the map structure. The EECMs used for test development and the mini-maps in Educator Portal depict the finalized node connections. Content and special education experts rated the ordering of node connections. In the limited instances of identified node connection reversals, the map developers adjusted the learning map connections. Additional indirect evidence of correct map sequencing comes from analyses evaluating linkage level ordering for each EE (linkage levels measure one or more nodes in the map). Modeling analyses indicated more than 80% of EEs showed evidence of correct hierarchical ordering (W. J. Thompson & Nash, 2019). The Target and Successor levels were most flagged for potential incorrect sequence. These levels also had the smallest available sample sizes. External alignment study ratings of the ordered linkage levels demonstrated correct ordering for most EEs. In ELA, panelists rated 76 of 95 EEs (80%) as showing a clear progression from Precursor to Successor levels. In mathematics, panelists rated 64 of 66 EEs (97%) as showing a clear progression. Collectively, the evidence supports the proposition that nodes are correctly sequenced.

10.4.1.2. B: Alternate Content Standards, the Essential Elements, Provide Grade-Level Access to College and Career Readiness Standards

The EEs are rigorous academic content standards measured by DLM assessments. Together with the learning map, the EEs make up the content structures for the DLM assessment system, as shown in Figure 10.2. Because students with significant cognitive disabilities have historically had limited opportunities to learn academic content (e.g., Karvonen, Flowers, & Wakeman, 2013), the DLM program prioritized adoption of rigorous academic content standards that provide adequate breadth and complexity of content.

Four propositions describe the EEs measured by DLM assessments. Table 10.3 summarizes the propositions and evidence for EEs providing grade-level access to rigorous academic expectations.

Table 10.3

Propositions and Evidence for Essential Elements

Proposition	Procedural evidence	Empirical evidence	Type	Chapter(s)
The grain size and description of Essential Elements (EEs) are sufficiently well-defined to communicate a clear understanding of the targeted knowledge, skills, and understandings	Description of EE development, including state and content expert review		Content	2
EEs capture what students should know and be able to do at each grade level to be prepared for postsecondary opportunities, including college, career, and citizenship	Description of EE development and review, vertical articulation evidence	Alignment study (EE to Common Core State Standards), postsecondary opportunities ratings	Content	2, 7
The collection of EEs in each grade sufficiently sample the domain	Description of development and review of EEs and blueprint		Content	2
EEs have appropriately specified linkage levels measuring map nodes	Helplet videos, Essential Element Concept Maps, description of development and review of EEs and map†	External review ratings, alignment study (EE to target node)	Content	2

† Relies on evidence from Theory of Action input statements, as shown in Figure 10.2.

For EEs to provide students with significant cognitive disabilities access to grade-level college and career readiness standards, the EEs must be well defined and must clearly articulate the targeted knowledge, skills, and understandings. Documentation describes the EE development process, which included instructions for test developers and educators to follow when specifying the targeted knowledge, skills, and understandings of each EE. The DLM Governance Board and other content experts reviewed the draft EEs and provided feedback on their clarity and grain size before adoption. While this proposition was not evaluated with empirical evidence such as rating data, the rigorous development process and rounds of internal and external review provide support for this proposition.

The EEs should also build across grades to prepare students for postsecondary opportunities. Documentation describes the development process, including procedures for establishing and reviewing the vertical articulation of EE content across grades and alignment with college and career expectations. The DLM Governance Board and other content experts reviewed the EEs for sequencing of content across grades. Empirical evidence includes external ratings on the alignment of EEs to the Common Core State Standards in each grade. Additional indirect evidence comes from a postsecondary opportunities study that evaluated the alignment of skills needed for postsecondary opportunities with the *At Target* performance level descriptors (PLDs) (Karvonen et al., 2022). Although panelists rated skills against grade- and content-specific PLDs, the PLDs are based on the types of skills students tend to master across EEs. Collectively, the evidence indicates that EEs capture what students should know and be able to do at each grade to be prepared for postsecondary opportunities.

For EEs to provide students with grade-level access to rigorous standards, the EEs must sufficiently sample the domain. Evidence includes a description of the process for developing EEs to sample the breadth of Common Core State Standards. The DLM Governance Board and other content experts reviewed the EEs for adequate sampling breadth. Documentation also describes the original blueprint development process in which blueprints were designed to sample content across the full range of claims and conceptual areas. Additional documentation describes the blueprint revision, which preserved the breadth of coverage across claims and conceptual areas. The DLM Governance Board and Technical Advisory Committee (TAC) reviewed both the EEs and the assessment blueprints for breadth of coverage before adoption. While this proposition was not evaluated with empirical evidence, the procedural evidence demonstrates that EEs sufficiently sample the domain.

Finally, to provide students with access to grade-level standards, EEs must have appropriately specified linkage levels, measuring nodes in the learning map. Documentation describes the process for simultaneously developing and revising the EEs and the learning maps to reflect each other and how linkage levels were developed. Short helplet videos describe how testlets measure map nodes. The EECMs demonstrate the map nodes measured by the five linkage levels of each EE. Project staff conducted rounds of review to consider alignment between map nodes and the EE linkage levels. During an alignment study, all EEs were rated as aligned to target-level nodes, with most EEs rated as having a “near” link to the target-level node. The combination of evidence across EE and map development and review provides strong support for the alignment of EEs to linkage levels measuring map nodes.

10.4.1.3. C: The Kite System Used to Deliver DLM Assessments is Designed to Maximize Accessibility

For the DLM program to accurately measure student knowledge, skills, and understandings, the Kite Suite must be accessible to DLM students and their educators. There are no direct inputs into the Kite Suite design statement in the Theory of Action (see Figure 10.1).

Four propositions describe the design and accessibility of the Kite Suite used to deliver DLM assessments. Table 10.4 summarizes the propositions and evidence evaluating the accessible design of the Kite Suite.

Table 10.4

Propositions and Evidence for the Kite Suite

Proposition	Procedural evidence	Empirical evidence	Type	Chapter(s)
System design is consistent with accessibility guidelines (e.g., accessible portable item protocol standards and web content accessibility guidelines) and current best practices	System documentation		Content	3, 4
Supports needed by students are available within and outside of the assessment system	*Accessibility Manual*, First Contact and Personal Needs and Preferences helplet video	Personal Needs and Preferences Profile data, test administrator surveys, focus groups	Content	4
Item types support the range of students in presentation and response	Description of item types	Cognitive labs, test administrator survey	Response Process, Content	3, 4
Kite Suite is accessible and usable by educators	*Test Administration Manual*, Educator Portal helplet video	Test administrator survey, focus groups		4

For the Kite Suite to maximize accessibility, system design should be consistent with accessibility guidelines, accessible portable item protocol (APIP) standards, web content accessibility guidelines (WCAG), and current best practices. System documentation provides evidence that the Kite Suite meets the highest standards for system design.

For the Kite Suite to be fully accessible to all students, all supports needed by students within and outside of the assessment system should be available. Documentation describes the types and range of accessibility supports available. Test administrators record the accessibility supports used by the student in the Personal Needs and Preferences Profile (PNP). In 2021–2022, all available PNP supports were used, ranging from less than 1% of students (n = 109) using uncontracted braille to 79% (n = 75,846) using human read aloud. Test administrator surveys provide additional evidence that accessibility supports needed by the student are available within and outside the assessment system. Test administrator survey responses from 2017 to 2022 indicate that 92%–95% of students had access to all supports necessary to participate in the assessment. In spring 2020, DLM staff conducted focus groups with some of the educators who disagreed that students had all necessary supports. The findings revealed that most of the accessibility and system challenges that educators reported stemmed, in part, from uncertainty about allowable practices for assessment administration, rather than gaps in system functionality. This uncertainty included practices used during instruction that are not allowed during assessment (e.g., hand-over-hand support). Together, the evidence demonstrates that the supports needed by students are available within and outside the system.

System accessibility also means that item types support the full range of DLM students in presentation and response. Multiple item types, as well as both computer-delivered and educator-administered items, allow students to demonstrate their knowledge, skills, and understandings. Cognitive labs showed that students can respond to a range of item types. The 2022 test administrator survey findings show that educators agreed or strongly agreed that 89% of students responded to items to the best of their knowledge, skills, and understandings and that 86% of students were able to respond regardless of disability, behavior, or health concerns. These percentages are consistent with responses from the 2018 and 2019 test administrator surveys. Overall, the evidence suggests that the different item types support the full range of students in presentation and response.

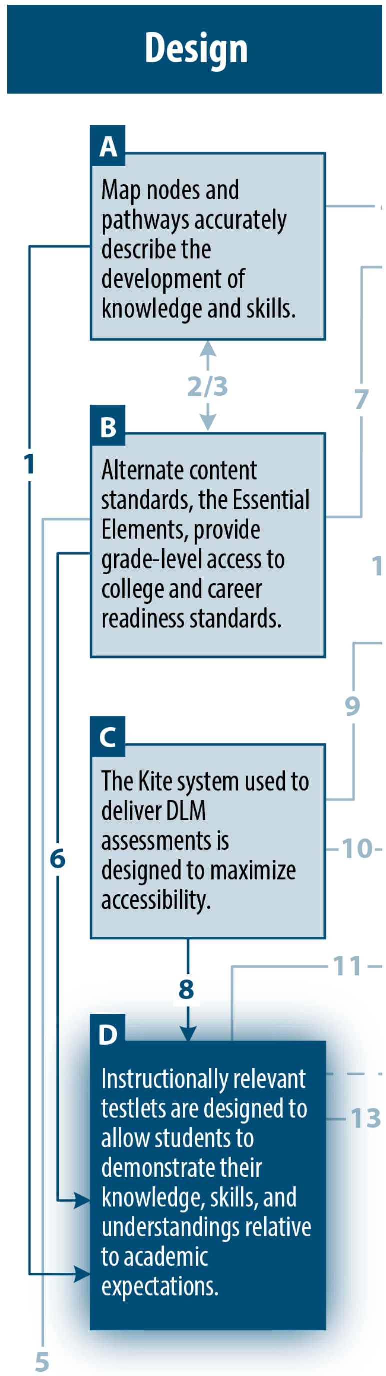
Finally, the Kite Suite must also be accessible to educators. Documentation describes how educators use Educator Portal, including the Instruction and Assessment Planner, to assign optional instructionally embedded testlets, manage student data, and access Testlet Information Pages. The documentation also describes how educators use Student Portal to administer the testlet to students. Short helplet videos describe how to use Educator Portal and Student Portal. Spring 2019 test administrator survey data indicate that most respondents found both Educator Portal and Student Portal easy to use. Educators rated their overall experience with Educator Portal as good or excellent in 85% of cases, and their overall experience with Kite Student Portal as good or excellent in 91% of cases. In addition, when asked about their experiences with Student Portal, educators found it to be either somewhat easy or very easy to enter the site in 95% of cases, somewhat or very easy to navigate within a testlet in 96% of cases, somewhat or very easy to record a response in 97% of cases, somewhat or very easy to submit a completed testlet in 97% of cases, and somewhat or very easy to administer testlets on various devices in 93% of cases. Focus group feedback also indicated that educators found the Kite Suite easy to use. Overall, there is strong evidence that the Kite Suite is accessible and usable for educators.

10.4.1.4. D: Instructionally Relevant Testlets are Designed to Allow Students to Demonstrate Their Knowledge, Skills, and Understandings Relative to Academic Expectations

To meet the needs of students with significant cognitive disabilities, DLM assessments are comprised of short, instructionally relevant testlets. Testlet development relies on inputs from the learning maps (Statement A), EEs (Statement B), and Kite Suite (Statement C), shown as arrows in the excerpt of the Theory of Action in Figure 10.3.

Figure 10.3

Theory of Action Inputs for Test Design



Seven propositions describe the design and development of DLM testlets and items, which are summarized in Table 10.5.

Table 10.5

Propositions and Evidence for Test Design

Proposition	Procedural evidence	Empirical evidence	Type	Chapter(s)
Items within testlets are aligned to linkage level	Item writing handbook, fungible linkage level parameters, ELA and mathematics testlet helplet videos, Essential Element Concept Maps (EECMs)†	Alignment data, item analyses	Content, Internal Structure	3, 4
Testlets are designed to be accessible to students	Item writer qualifications, item writing handbook, *Accessibility Manual*, EECMs†	Internal and external review, test administrator survey, focus groups	Content	3, 4
Testlets are designed to be engaging and instructionally relevant	Item writer qualifications, item writing handbook, *Accessibility Manual*	Internal and external review, test administrator survey, focus groups	Content	3, 4
Testlets are written at appropriate cognitive complexity for the linkage level	Item writing handbook, EECMs†	Cognitive process dimension ratings, text complexity, external review of passages, internal and external review	Content	3
Items are free of extraneous content	Item writing handbook, EECMs†	Internal and external review	Content	3
Items do not contain content that is biased against or insensitive to subgroups of the population	Item writing handbook	Internal and external review, differential item functioning analyses	Content, Internal Structure	3, 4
Items are designed to elicit consistent response patterns across different administration formats	Item writing handbook, *Accessibility Manual*, *Test Administration Manual*, EECMs†	Internal and external review, item statistics	Content, Internal Structure	3

† Relies on evidence from Theory of Action input statements, as shown in Figure 10.3.

For testlet design to allow students to demonstrate their knowledge, skills, and understandings, items must appropriately measure the linkage level. The EECMs for each EE support item writing by specifying the linkage levels and nodes assessed. The item writing handbook provides guidance to item writers for how to write items measuring the linkage level. Items are written to be fungible, meaning that all items written to an EE and linkage level should measure the linkage level equally well. In modeling analyses, all items measuring the linkage levels have equal parameters. Helplet videos describing ELA and mathematics testlets illustrate these principles. External alignment ratings show that 96% of ELA items and 100% of mathematics items were rated as meeting content centrality criteria.[72] Additionally, item analyses show that the *p*-value for 98% of ELA items and 99% of mathematics items fall within two standard deviations of the mean for the linkage level. This demonstrates that items consistently measure the linkage level, supporting the intended fungibility of items. Together, the evidence shows that items are aligned to the linkage level.

Testlets must also be accessible to students. Items are written by educators familiar with students with significant cognitive disabilities. The item writing handbook specifies how to write accessible testlet content, and the *Accessibility Manual* (DLM Consortium, 2021b) describes how item writing guidelines are based on the Universal Design for Learning. The EECMs specify accessibility considerations for the content being measured. Items are internally and externally reviewed for accessibility. Test administrator survey responses from 2022 indicate that 86% of students were able to respond regardless of disability, behavior, or health concerns, which is consistent with findings from 2018 and 2019. As previously described for Statement C, educator focus groups did not reveal gaps in system functionality that prevent accessibility. Together, the evidence suggests testlet content is accessible to students.

Testlets should be designed to be engaging and instructionally relevant. Items are written by educators familiar with instruction for students with significant cognitive disabilities. The item writing handbook describes how to write engaging and instructionally relevant content, and the *Accessibility Manual* (DLM Consortium, 2021b) describes the engagement activities that are designed to motivate students, provide a context for the items, and activate prior knowledge. Internal and external review evaluates testlet content. On the 2022 test administrator survey, educators indicated that testlets aligned with instruction for approximately 71% of students in ELA and approximately 62% of students in mathematics. Focus group feedback reveals some variability in educator perception that content is engaging and instructionally relevant. Together, there is moderate evidence that testlets are engaging and instructionally relevant, with opportunity for additional data collection.

Testlet content must also be written at the appropriate cognitive complexity for the linkage level. The EECMs specify the cognitive complexity of the content, while the item writing handbook provides guidance for writing items that differentiate the linkage levels. All items are rated for cognitive process dimension and evaluated for text complexity. Rounds of internal and external review by content experts provide additional evidence that items and testlets are written at the appropriate level. Together, the evidence indicates that testlet content is generally written at an appropriate cognitive complexity for the linkage level.

Items should also be free of extraneous content. The EECMs define the node(s) and linkage levels being measured to focus item writing on necessary content. The item writing handbook provides further guidance for focusing item writing to reduce cognitive load for students with significant cognitive disabilities (e.g.,

[72] Content centrality is the degree of fidelity between the content of the learning map nodes and the assessment items. See Chapter 3 of this manual for details.

using concise language at an appropriate reading level). Rounds of internal review by ATLAS staff and external review by content panels further evaluate item content. Together, evidence supports the proposition that items are free of extraneous content.

Items should similarly be free of biased or sensitive content. The item writing handbook provides guidance for writing testlets, including consideration for diverse subgroups of the population and topics that may be sensitive or inaccessible for students with significant cognitive disabilities (e.g., references to mobility, tasks requiring sight). Internal and external review includes bias and sensitivity review, incorporating the evaluation of items, testlets, and ELA texts. Differential item functioning analyses conducted across years demonstrate that more than 99% of items have no or negligible evidence of differential functioning across student subgroups. Together, evidence shows items to be free of sensitive content and bias.

Finally, items should elicit consistent responses regardless of administration format. The EECMs document the content measured by each linkage level to promote consistency within and across testlets. The item writing handbook further describes procedures to promote consistency so that item writers develop items and testlets according to fungibility principles (i.e., all items measuring an EE and linkage level have equal parameters). Documentation describes administration procedures for promoting consistency regardless of form type (e.g., braille) or supports used (e.g., switch use). Internal review and external review evaluate item content. Item statistics demonstrate that items consistently measure the linkage level. While there is opportunity for additional data collection, the collected evidence indicates that items elicit consistent responses regardless of administration format.

10.4.1.5. E: Training Strengthens Educator Knowledge and Skills for Assessing

Required training prepares educators to administer DLM assessments with fidelity. Because the format of DLM assessments differs from other assessments (e.g., short testlets, computer-based rather than portfolio), required training covers critical concepts educators need to know to administer DLM assessments as intended. There are no direct inputs for the design of required training in the Theory of Action.

Three propositions pertain to the required test administrator training for DLM assessments. Table 10.6 summarizes the propositions and evidence evaluating required training for educators.

Table 10.6

Propositions and Evidence for Required Training

Proposition	Procedural evidence	Empirical evidence	Type	Chapter(s)
Required training is designed to strengthen educator knowledge and skills for assessing	Documentation of scope of training and process of developing, posttest-passing requirements		Content	9
Training prepares educators to administer DLM assessments	Documentation of scope of training, *Test Administration Manual* and *Accessibility Manual* as supplemental materials	Test administrator survey	Consequences	4, 9
Required training is completed by all test administrators	System requirement, description of self-directed and facilitated training options, state-provided guidance on training, Kite training status extracts, state and local monitoring	Data extracts	Consequences	9

Required training should strengthen educator knowledge and skills for assessing students. Documentation describes the scope and development process for identifying the set of critical knowledge and skills that test administrators need to administer DLM assessments. To demonstrate mastery of critical knowledge and skills, all test administrators must pass the module posttests with at least 80% accuracy. Since 2015, project staff have updated required training modules to reflect the most critical knowledge needed for administration while maintaining a short time commitment needed to complete. Together, the evidence supports the statement that required training is designed to strengthen educator knowledge for assessing students.

Training should prepare educators to administer assessments. Annual training covers topics educators need to know to administer DLM assessments. Test administrators also have access to the *Test Administration Manual* (DLM Consortium, 2021a) and the *Accessibility Manual* (DLM Consortium, 2021b) to prepare for test administration. In 2022, 89% of survey test administrators agreed or strongly agreed that test administrator training prepared them for test administrator responsibilities. These results are consistent with prior findings from 2017 to 2021, which ranged from 87% to 91%. These sources of evidence show that training prepares educators to administer assessments, with some opportunity for continuous improvement.

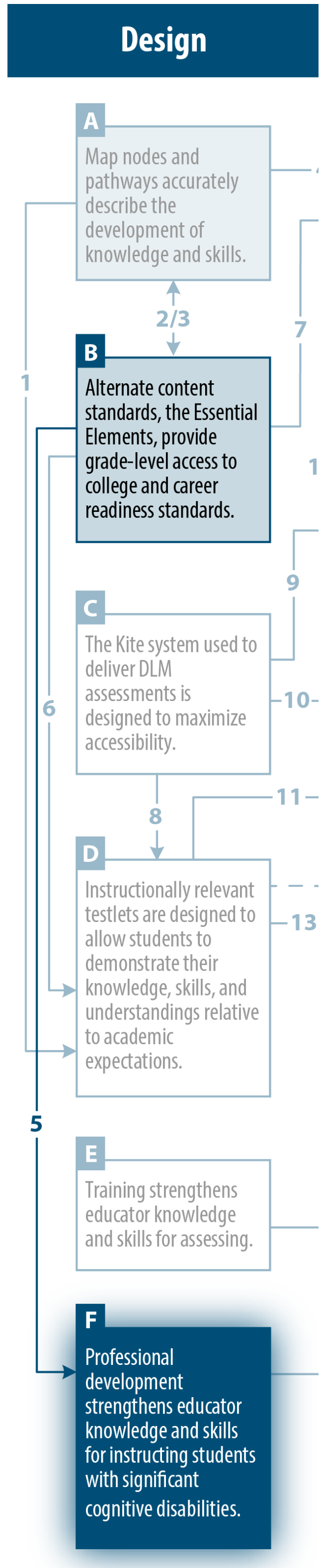
Test administrators must complete required training. The Kite Suite prevents test administration until required training is completed. Self-directed and facilitated training formats are available to meet diverse needs. States and districts determine the format of training and provide test administrators with additional guidance and resources as needed. State, district, and building staff can monitor training completion using the Training Status Extract generated by the Kite Suite. Data extracts show that all test administrators completed required training.

10.4.1.6. F: Professional Development Strengthens Educator Knowledge and Skills for Instructing Students With Significant Cognitive Disabilities

Professional development modules support educators in providing high-quality, rigorous academic instruction. Because students with significant cognitive disabilities have historically had limited opportunities to learn academic content (e.g., Karvonen, Flowers, & Wakeman, 2013), the DLM assessment program prioritized development of professional development resources and their inclusion in the Theory of Action. Professional development modules are optionally available in all states, with some states including a subset of modules in their required educator training. Professional development relies on one input (EEs, Statement B), as shown in Figure 10.4.

Figure 10.4

Theory of Action Inputs for Professional Development



Three propositions are related to the design and structure of professional development for educators who instruct students who take DLM assessments. Table 10.7 summarizes the evidence evaluating professional development.

Table 10.7

Propositions and Evidence for Professional Development

Proposition	Procedural evidence	Empirical evidence	Type	Chapter(s)
Professional development (PD) modules cover topics relevant to instruction	Description of approach to PD, list of PD modules, Essential Elements†	PD rating data	Consequences	9
Educators access PD modules	Description of process for accessing PD modules	PD prevalence data	Consequences	9
Educators implement the practices on which they have been trained	Description of PD content and its application to practice	PD rating data	Consequences	9

† Relies on evidence from Theory of Action input statements, as shown in Figure 10.4.

For professional development to strengthen educator knowledge and skills for instructing students with significant cognitive disabilities, it must cover topics relevant to instruction. The DLM professional development system includes 51 modules that address instruction in ELA and mathematics and support educators in creating Individual Education Programs (IEPs) that are aligned with the DLM EEs. Empirical evidence of the instructional relevance of the professional development modules comes from educator evaluations collected after completing the modules. During the 2021–2022 year, approximately 83% of educators completing the surveys either agreed or strongly agreed that the modules addressed content that is important for professionals working with students with significant cognitive disabilities, and 82% agreed that the modules presented new ideas to improve their work. This combined evidence shows that professional development covers topics relevant to instruction.

For professional development to strengthen knowledge, educators should access the modules. To facilitate access, the modules are available on the DLM professional development website in two formats, self-directed and facilitated. There are also bundles of related content. A subset of states requires that specific professional development modules be completed within the required training interface. In 2021–2022, nine states required at least one professional development module as part of their required test administrator training. Across all states administering DLM assessments in 2022, 21,888 modules were completed by 4,431 new test administrators and 3,033 returning test administrators as part of required training. Additionally, a total of 5,987 modules were completed by 1,426 individuals through the DLM professional development website. Since the first professional development module was launched in the fall of 2012, a total of 77,061 modules have been completed via the DLM professional development website. Overall, these findings show somewhat limited adoption of the professional development modules.

For professional development to strengthen educator knowledge, educators should implement the practices covered in the modules. Documentation (e.g., facilitator guides) describes application of professional development content to instructional practice. In 2021–2022, 83% of educators indicated that they intended to apply what they learned in the professional development modules to their professional practice. There is opportunity for additional evidence to be collected for this proposition.

10.4.2. Delivery

According to the chain of reasoning in the Theory of Action, the collective set of design statements informs delivery of DLM assessments. Theory of Action statements regarding delivery of DLM assessments pertain to administration and implementation of the DLM System. This portion of the Theory of Action encompasses statements about educators' instruction, assessment administration, student interactions with the system, and the combined set of administered testlets. The specific statements related to the delivery of DLM assessments are:

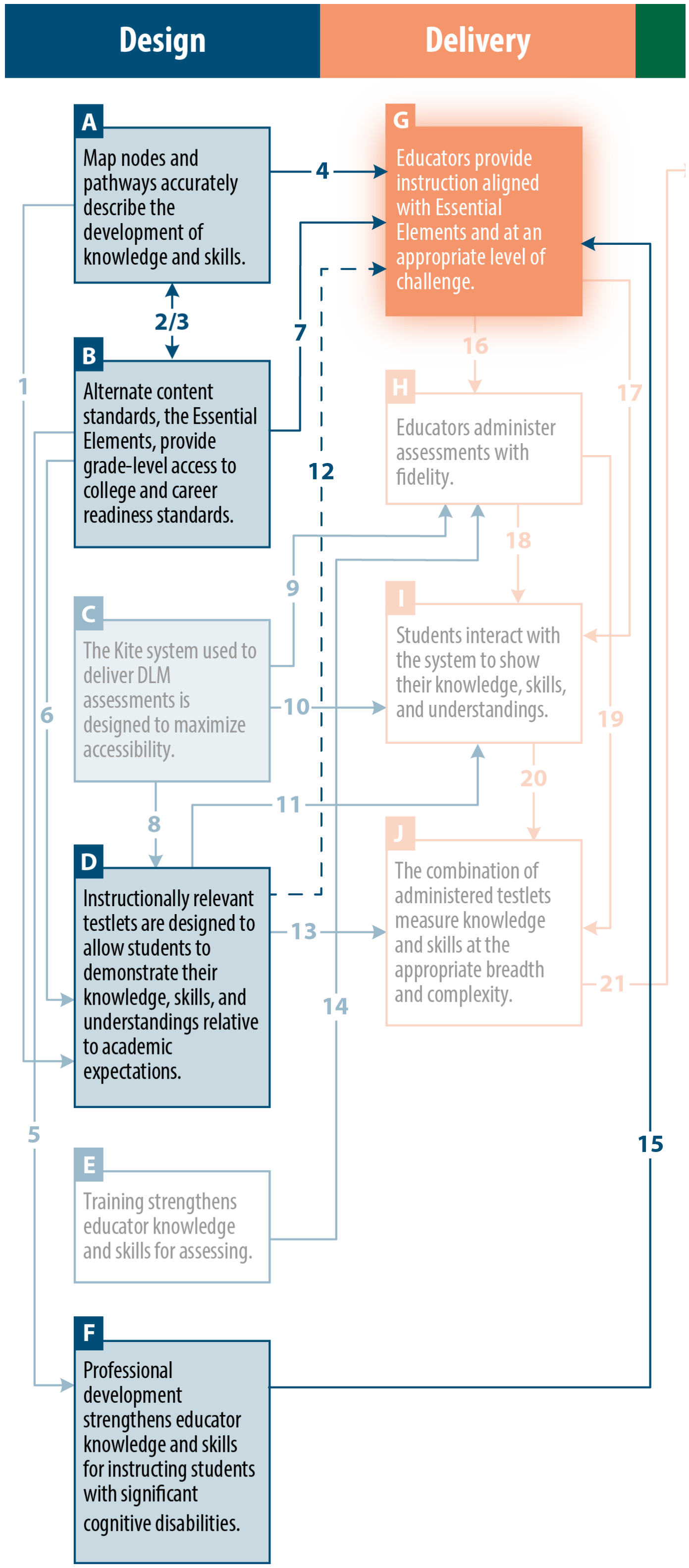
G. Educators provide instruction aligned with content standards and at an appropriate level of challenge.
H. Educators administer assessments with fidelity.
I. Students interact with the system to show their knowledge, skills, and understandings.
J. The combination of administered assessments measure knowledge and skills at the appropriate breadth and complexity.

10.4.2.1. G: Educators Provide Instruction Aligned With Content Standards and at an Appropriate Level of Challenge

Students with significant cognitive disabilities have historically had limited opportunity to learn the full breadth of grade-level academic content (e.g., Karvonen, Wakeman, et al., 2013; Karvonen et al., 2009, 2011). To address this, the DLM program prioritizes educators providing instruction aligned with content standards at an appropriate level of challenge. Several inputs in the Theory of Action inform instructional practice, including the availability of rigorous grade-level expectations as reflected in the map and EEs (Statements A and B), the content of instructionally relevant testlets (Statement D), and professional development that strengthens educator knowledge and skills for instructing the DLM population of students (Statement F), as shown in Figure 10.5.

Figure 10.5

Theory of Action Inputs for Aligned Instruction



Two propositions are related to the alignment of instruction with DLM content standards, as shown in Table 10.8.

Table 10.8

Propositions and Evidence for Aligned Instruction

Proposition	Procedural evidence	Empirical evidence	Type	Chapter(s)
Educators provide students with the opportunity to learn content aligned with the assessed grade-level Essential Elements	State and local guidance, blueprints†	Test administrator survey, First Contact survey	Content	4
Educators provide instruction at an appropriate level of challenge, based on knowledge of the student	Instructional resources, required training, mini-maps†	Focus groups, test administrator survey	Content	4

† Relies on evidence from Theory of Action input statements, as shown in Figure 10.5.

Educators should provide students with the opportunity to learn content aligned with the EEs. States and districts provide additional guidance to support educators in providing high-quality instruction aligned with grade-level EEs. Blueprints specify the breadth of EE coverage across claims and conceptual areas. The test administrator survey collects information on the approximate number of hours during the school year that educators spent on instruction for each of the DLM conceptual areas. For both ELA and mathematics, 51% of the test administrators provided 11 or more hours of instruction per conceptual area to their students in ELA, and 42% did so in mathematics. The results are consistent with prior survey data, showing variability in student opportunity to learn the full breadth of academic content. One factor that affects opportunity to learn is student engagement during instruction. Based on First Contact survey responses, 62% of the students who take DLM assessments demonstrate fleeting attention to educator-directed instruction, and 56% demonstrate fleeting attention to computer-directed instruction. A small percentage of students demonstrate little or no attention to educator-directed (14%) or computer-directed (13%) instruction. These results collectively indicate some variability in the amount of instruction and in the level of student engagement in instruction for the full breadth of academic content, with opportunity for continuous improvement.

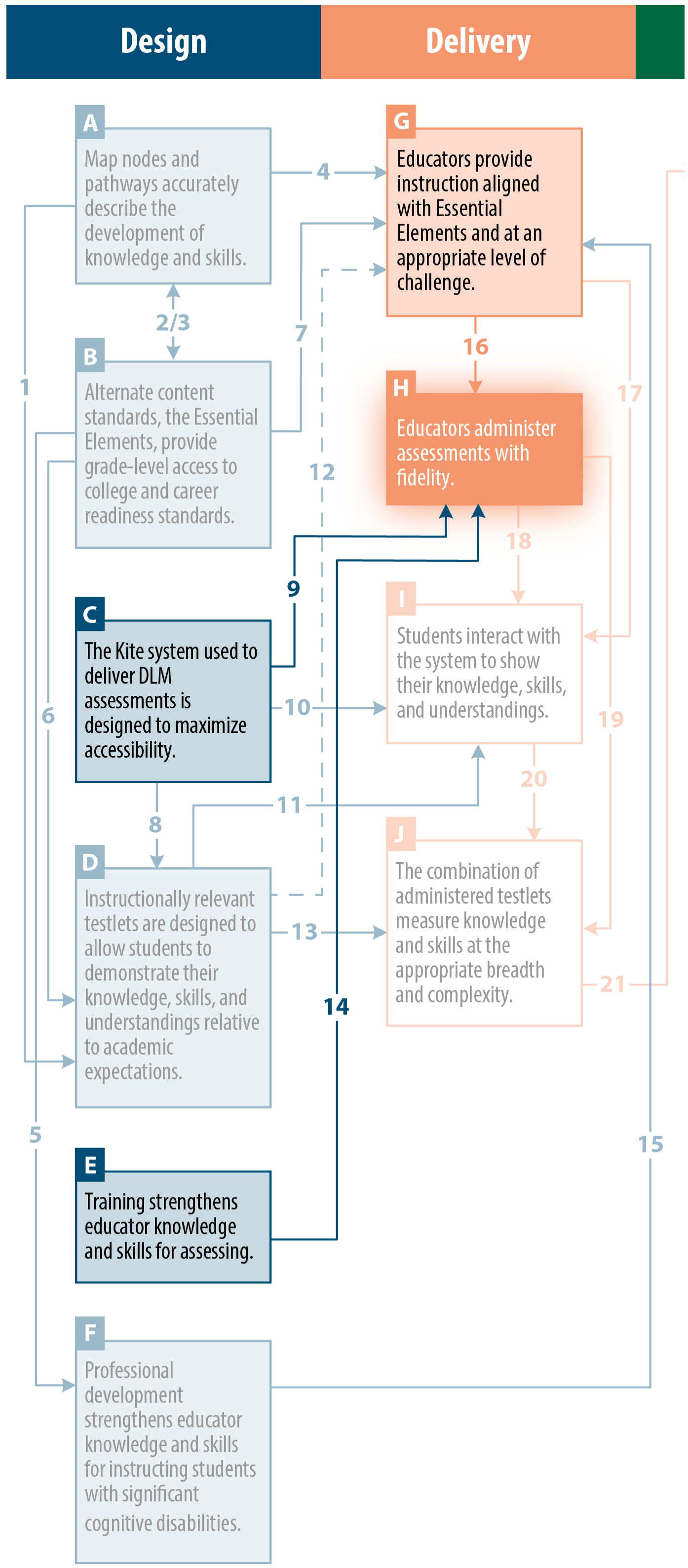
Educators should also provide instruction at an appropriate level of challenge based on knowledge of the student. Instructional resources on the DLM website are intended to be used to inform instruction prior to administration of DLM assessments. This includes directions for accessing mini-maps in Educator Portal to identify the map nodes measured at each linkage level to inform instruction. Spring 2022 test administrator survey responses indicated that most or all ELA testlets matched instruction for approximately 71% of students, and most or all mathematics testlets matched instruction for approximately 62% of students. These percentages have annually increased since 2017. While there is opportunity for additional data collection and continuous improvement in instructional practice, current data provides some evidence supporting the proposition that educators provide instruction at an appropriate level of challenge based on knowledge of the student.

10.4.2.2. H: Educators Administer Assessments With Fidelity

Educators should administer DLM assessments as intended. Administering assessments with fidelity relies on two inputs in the DLM Theory of Action, the Kite Suite design (Statement C) and required training content (Statement E), and aligned instruction, as shown in Figure 10.6.

Figure 10.6

Theory of Action Inputs for Administration Fidelity



Four propositions are related to fidelity of assessment administration, as shown in Table 10.9.

Table 10.9

Propositions and Evidence for Administration Fidelity

Proposition	Procedural evidence	Empirical evidence	Type	Chapter(s)
Educators are trained to administer testlets with fidelity (i.e., as intended)	*Test Administration Manual* (TAM), *Accessibility Manual*, Testlet Information Pages, required training†	Test administrator survey, test administration observations (TAOs)	Response Process, Content	4, 9
Educators enter accurate information about administration supports	TAM, Personal Needs and Preferences Profile (PNP) helplet video, required training,† PNP system functionality†	PNP selections	Response Process, Content	4, 9
Educators allow students to engage with the system as independently as they are able	TAM	PNP, TAOs, First Contact responses	Response Process, Content	4, 9
Educators enter student responses with fidelity (where applicable)	TAM	PNP, TAOs, writing interrater agreement	Response Process, Content	4, 9

† Relies on evidence from Theory of Action input statements, as shown in Figure 10.6.

For educators to administer assessments with fidelity, they must be appropriately trained, which relies in part on evidence from the required training input (Statement E) in the Theory of Action (Figure 10.6). The *Test Administration Manual* (DLM Consortium, 2021a), *Accessibility Manual* (DLM Consortium, 2021b), and Testlet Information Pages provide test administrators with information needed to administer assessments as intended. On the spring 2022 test administrator survey, educators agreed or strongly agreed that the required test administrator training prepared them for their responsibilities as test administrators in 89% of cases and agreed or strongly agreed that they were confident administering DLM testlets in 96% of cases. These results are consistent with responses received since 2017. From 2019 to 2022, educators agreed or strongly agreed that they used manuals and/or the DLM Educator Resource Page[73] materials in 89%–93% of cases. Indirect evidence comes from test administration observation data; observations indicated educators administered assessments as intended, suggesting that training

[73] https://dynamiclearningmaps.org/instructional-resources-ye

effectively prepared them to administer assessments. Together, the evidence shows that educators are trained to administer assessments with fidelity.

Fidelity of assessment administration also includes educators entering accurate information about administration supports used by the student. In addition to being covered in required training, documentation and a helplet video describe expectations for educators completing the student's PNP to indicate supports used during assessment administration. During the 2021–2022 academic year, 88% of students had at least one support indicated on their PNP. The most selected supports in 2021–2022 were human read aloud, test administrator enters responses for student, and spoken audio. However, data are not available to evaluate the accuracy of educator PNP selections (i.e., whether supports were actually used on the assessment) or their consistency with supports used during instruction. While there is opportunity for additional data collection, the evidence provides general support for educators entering accurate information about supports used by the student.

Administering DLM assessments with fidelity requires that educators allow students to engage with the system as independently as they are able. The *Test Administration Manual* (DLM Consortium, 2021a) provides guidance for administering computer-delivered assessments with flexibility to meet individual student needs. In 2022, 54% of student PNP records indicated the test administrator entered responses on the student's behalf. Data are not available to evaluate whether students in fact used or needed this support during administration. On the First Contact survey, educators indicated that 40% of students could independently use a computer with or without assistive technology. Test administration observations collect additional data on test administrator behaviors. In 2022, 60% of observed behaviors were classified as supporting (e.g., clarifying directions), 38% neutral (e.g., asking the student to clarify their response), and 2% nonsupporting actions (e.g., physically directing the student to a response). While there is opportunity for additional data collection, evidence generally supports the proposition that educators allow students to engage with the system as independently as they are able.

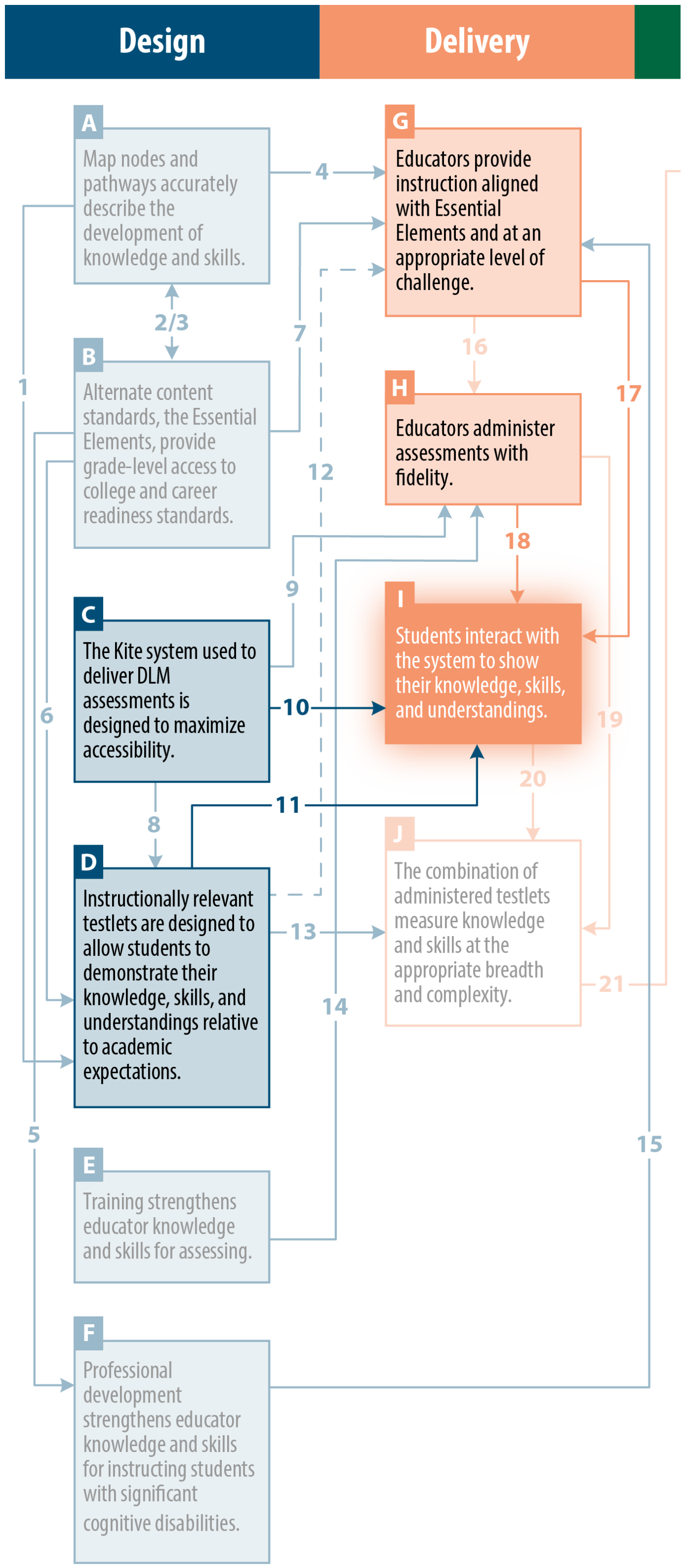
Finally, educators should enter student responses into the system with fidelity. While all testlets that are educator-administered require test administrators to enter students' responses, test administrators may also enter responses for students during computer-delivered testlets (i.e., if PNP support indicates the student needs the test administrator to enter responses). Resources provide educators with guidance for entering student assessment responses, including allowed and not-allowed practices. Test administration observations collected in 2021–2022 showed that test administrators entered responses for the student in approximately 18% of test observations (21 of 115), which is lower than the percentage of students whose PNP indicated the test administrator enters responses (54%). In approximately 86% of those cases (18 of 21 observations), observers indicated that the entered response matched the student's response. The remaining three observers either responded that they could not tell if the entered response matched the student's response, or the observer left the item blank. Additional fidelity evidence comes from evaluation of the consistency of scoring student writing samples, which are educator-administered and require the test administrator to enter responses in the system. Across years, interrater reliability of student writing samples scoring has demonstrated high levels of agreement, indicating educators enter student responses with fidelity. While evidence generally shows educators enter responses with fidelity, there is opportunity for additional data collection to further evaluate this proposition.

10.4.2.3. I: Students Interact With the System to Show Their Knowledge, Skills, and Understandings

Students should be able to interact with the system to show their knowledge, skills, and understandings. As shown in Figure 10.7, there are several relevant inputs in the Theory of Action that affect students' ability to show their knowledge and skills. These include system design (Statement C), testlet design (Statement D), instruction (Statement G), and administration fidelity (Statement H), as well as indirect inputs such as EEs (Statement B) and required training (Statement E).

Figure 10.7

Theory of Action Inputs for Student Interaction With the System



Three propositions correspond to students interacting with the system to show their knowledge, skills, and understandings on DLM assessments. Table 10.10 summarizes the propositions and evidence evaluating student interaction with the system.

Table 10.10

Propositions and Evidence for Student Interaction With the System

Proposition	Procedural evidence	Empirical evidence	Type	Chapter(s)
Students are able to respond to tasks, regardless of sensory, mobility, health, communication, or behavioral constraints	*Test Administration Manual* (TAM), *Accessibility Manual*, administration fidelity,† system and testlet design†	Test administrator surveys, Personal Needs and Preferences Profile/alternate form completion rates, test administration observations (TAOs), focus groups	Response Process, Content	3, 4
Student responses to items reflect their knowledge, skills, and understandings	System and testlet design,† administration fidelity,† received aligned instruction†	Cognitive labs, test administrator surveys	Response Process, Content	3, 4
Students are able to interact with the system as intended	TAM, *Accessibility Manual*, system design†	Cognitive labs, TAOs, test administrator surveys	Response Process, Content	4

† Relies on evidence from Theory of Action input statements, as shown in Figure 10.7.

To show their knowledge, skills, and understandings in the DLM System, all students should be able to respond to tasks regardless of sensory, mobility, health, communication, or behavioral constraints. Documentation describes procedures for selecting accessibility supports and administering the assessment in ways that align with individual student needs while also maintaining fidelity to intended practice. System and testlet accessibility design support students being able to show their knowledge, skills, and understandings. On the 2022 DLM test administrator survey, educators agreed or strongly agreed that 86% of students were able to respond regardless of disability, behavior, or health concerns. This percentage is consistent with responses since 2017. In 2022, educators indicated that 92% of students had access to all supports necessary to participate. This percentage is consistent with data since 2017. Test administration observation data showed that in 2021–2022, in 41 of 42 cases, students were able to respond without encountering difficulty using accessibility supports. The PNP selection data provide further evidence that educators select supports to meet individual student sensory, mobility, and communication needs, including enabling various display settings, switch use, or administration of alternate forms (e.g., braille). In 2021–2022, alternate forms for students with blindness or low vision were selected for 2,068 students (2%), and uncontracted braille was selected for 109 students (0.1%). As described for Statement C, accessibility focus groups revealed that the challenges educators reported stemmed in part from uncertainty about allowable practices for assessment administration rather than gaps in system functionality. This evidence shows that students, in general, can respond to tasks regardless of sensory, mobility, health, communication, or behavioral constraints.

Student responses to assessment items should reflect their knowledge, skills, and understandings rather than construct-irrelevant factors. Construct-irrelevant variance is minimized through system and testlet design and through administration fidelity (demonstrated by connection with Statements C, D, and H in Figure 10.7). Responses should similarly reflect students' knowledge, skills, and understandings after receiving rigorous aligned instruction (demonstrated by connection with Statement G). Cognitive labs provide evidence that items do not present barriers to the intended response process due to construct-irrelevant testlet features or item response demands. In addition, on the 2022 test administrator survey, 89% agreed that students responded to items to the best of their knowledge, skills, and understandings. This percentage is consistent with agreement rates in 2018 and 2019. Evidence generally indicates that student responses reflect their knowledge, skills, and understandings.

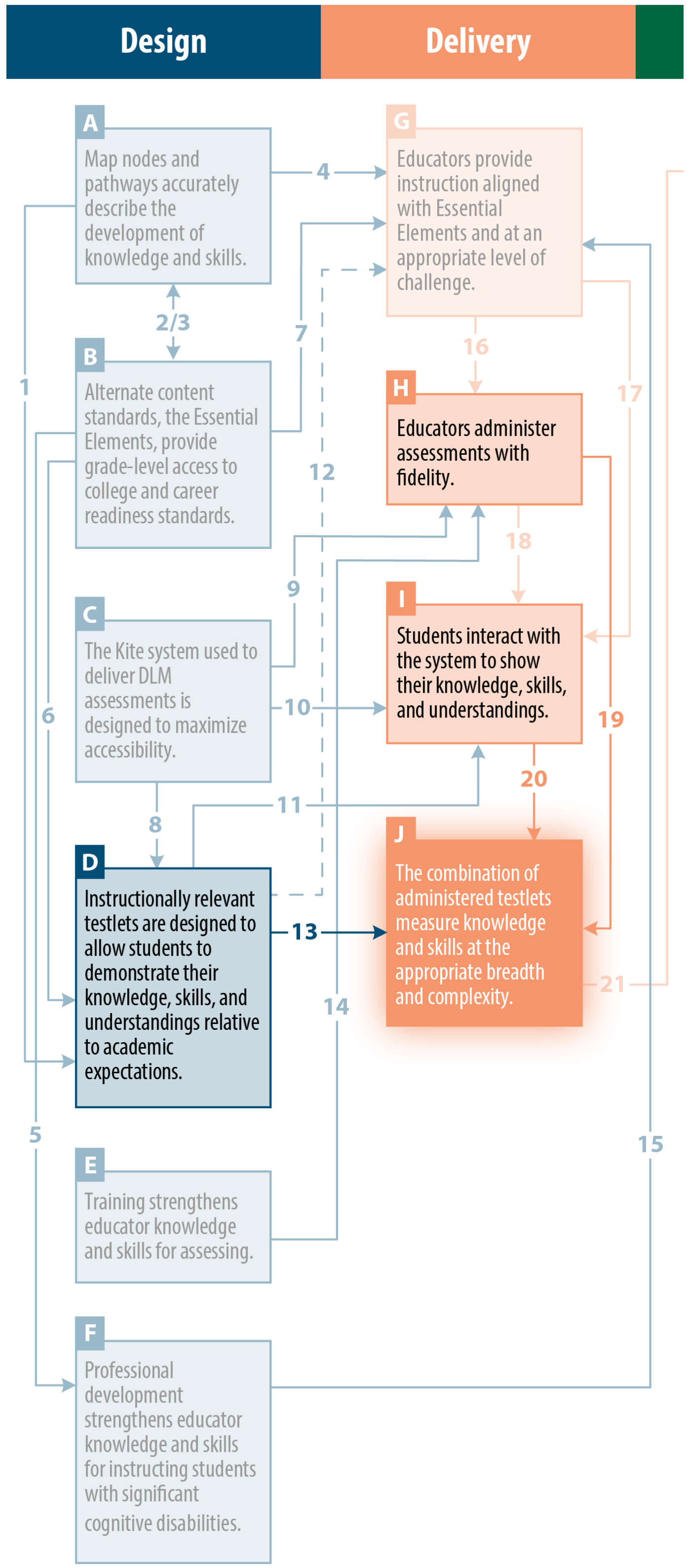
Finally, students should be able to interact with the system as intended. Documentation describes intended student interactions, including allowed and not-allowed practices. Confidence that students interact with the system as intended relies on system and testlet design and administration fidelity (demonstrated by connections with Statements C, D, and H in Figure 10.7). Cognitive labs, test administration observations, and test administrator survey data on Student Portal functionality provide evidence that students interact with assessments as intended. Furthermore, in 2019, more than 90% of educators indicated that Student Portal made it easy to navigate testlets, record responses, submit testlets, and administer testlets on a variety of devices (computer, iPad, etc.). Combined, this evidence supports the proposition that students are able to interact with the system as intended.

10.4.2.4. J: The Combination of Administered Assessments Measure Knowledge and Skills at the Appropriate Breadth and Complexity

The Theory of Action includes a statement that the combination of administered assessments measure knowledge and skills at the appropriate breadth and complexity. Administering testlets at the appropriate breadth and complexity relies on several direct inputs in the DLM Theory of Action, including testlet content being available for all EEs and linkage levels (Statement D), educators administering testlets with fidelity, and students interacting with the Kite Suite (Figure 10.8). Additionally, this statement includes indirect inputs like the map structure and EEs.

Figure 10.8

Theory of Action Inputs for Appropriate Combination of Testlets



Three propositions are related to the breadth and complexity of administered assessments, as summarized in Table 10.11.

Table 10.11

Propositions and Evidence for the Appropriate Combination of Testlets

Proposition	Procedural evidence	Empirical evidence	Type	Chapter(s)
First Contact survey correctly assigns students to appropriate complexity band	Description of First Contact survey design and algorithm development, First Contact helplet video	Pilot analyses, educator adjustment patterns	Content, Response Process	4
Administered testlets are at the appropriate linkage level	Administration fidelity,† mini-maps†	Adaptive routing patterns, linkage level parameters and item statistics, educator focus groups	Content	3, 4
Administered testlets cover the full blueprint	*Test Administration Manual*, monitoring extracts, blueprints,† administration fidelity†	Blueprint coverage extracts and analyses, Special Circumstance codes	Content	4

† Relies on evidence from Theory of Action input statements, as shown in Figure 10.8.

For the combination of administered assessments to measure knowledge and skills at the appropriate level of complexity, the First Contact survey should correctly assign students to a complexity band. Documentation and a helplet video describe expectations for educators accurately completing the First Contact survey. Documentation also details the algorithms that use a subset of First Contact items to assign students to complexity bands, which are used to assign students their first spring testlet in each subject. Evidence from the fixed-form pilot administration of DLM assessments demonstrated that complexity bands appropriately assigned students to linkage levels (Clark et al., 2014). Data from the 2022 spring assessment window indicate that 33% of ELA students and 41% of mathematics students did not adapt after their first testlet. In instances in which adaptations did occur, students adapted up and down at similar rates. The combination of evidence demonstrates some support that First Contact survey responses correctly assign students to complexity bands, with opportunity for additional research.

All administered testlets should be at the appropriate linkage level to allow students to demonstrate their knowledge, skills, and understandings at the appropriate level of complexity. This proposition assumes that testlets correctly measure linkage levels (Statement D in Figure 10.8). Documentation describes linkage level initialization and adaptive routing, whereby students are assigned their first testlet based on their complexity band, and subsequent testlets are assigned via adaptive routing based on performance on prior testlets. Routing data show some evidence of adaptation between testlets, which supports the proposition that assessments are at the appropriate linkage level. Linkage level modeling parameters (i.e., probability of masters and nonmasters providing correct responses, base rate of probability) and item statistics (i.e., *p*-values and standardized difference values) show that students perform as expected on assessed linkage levels, providing some evidence that administered testlets are at the appropriate level. However, educator feedback from focus groups indicates some variability in the perception that testlets are of appropriate difficulty. While evidence generally supports the proposition, there is opportunity for additional data collection.

For the combination of testlets to measure skills of appropriate breadth of content (i.e., adequate construct representation), administered assessments should cover the full blueprint. Documentation describes assignment procedures, including how frequently the system assigns subsequent testlets. Educators can monitor whether coverage requirements are met using Blueprint Coverage and Test Administration Monitoring extracts, which are available to educators and to building, district, and state users. Across all grades, 96% of students in ELA and 97% of students in mathematics were assessed on all of the EEs and met blueprint requirements. Special Circumstance codes entered in the system explain why some students do not meet coverage requirements (e.g., chronic absence). These results provide evidence that nearly all students are administered testlets to cover the full blueprint.

10.4.3. Scoring

According to the chain of reasoning demonstrated in the Theory of Action, the collective set of design and delivery statements inform scoring of DLM assessments. Scoring statements in the Theory of Action encompass students' mastery determinations, overall achievement, and usability of results. The specific statements are:

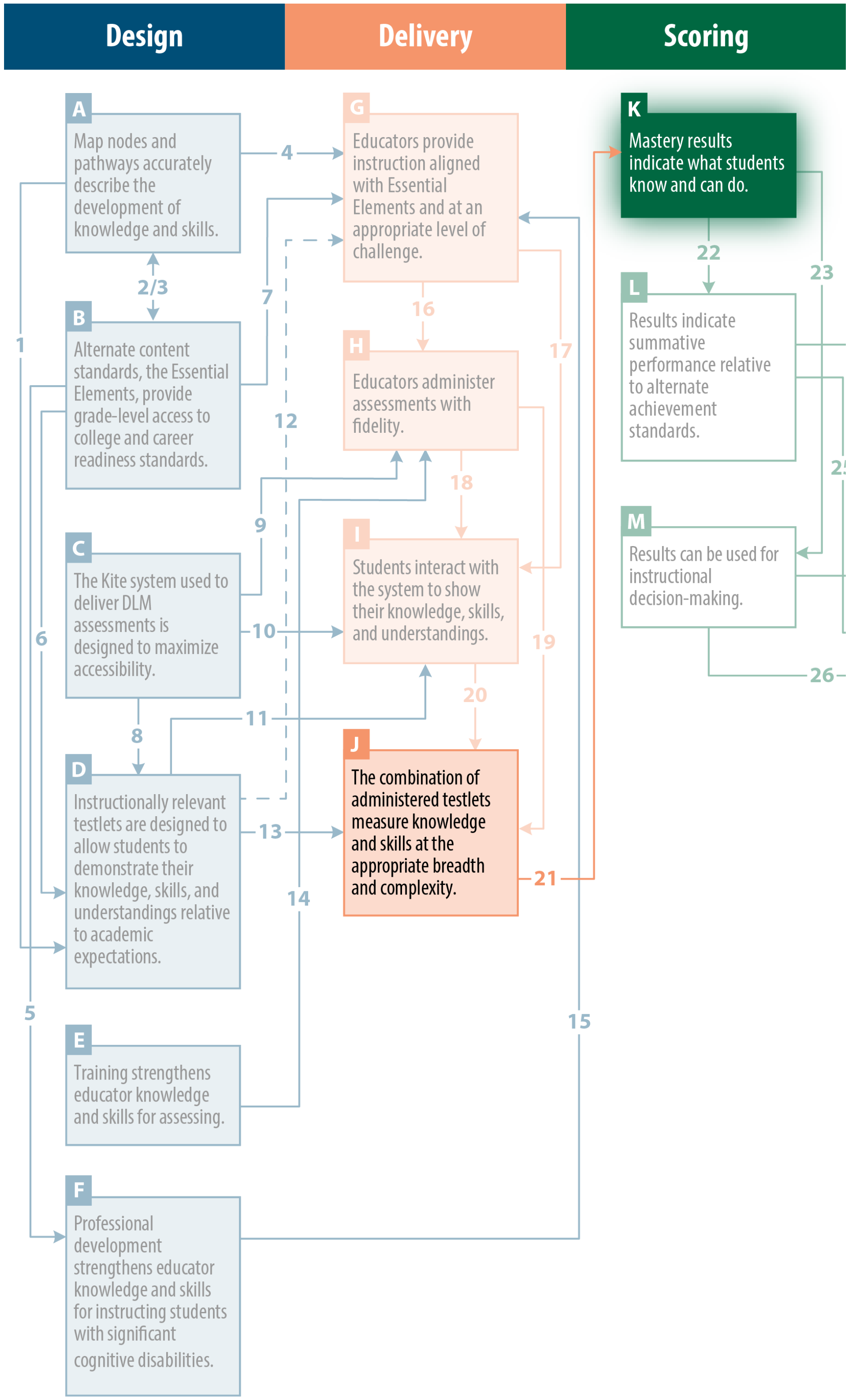
K. Mastery results indicate what students know and can do.
L. Results indicate summative performance relative to alternate achievement standards.
M. Results can be used for instructional decision-making.

10.4.3.1. K: Mastery Results Indicate What Students Know and Can Do

Because DLM assessments report results as the set of mastered skills across all EEs, mastery results should be accurate indications of students' knowledge, skills, and understandings. According to the Theory of Action, the only direct input informing mastery results the combination of administered testlets being at the appropriate breadth and complexity (Statement J). Indirect connections in the Theory of Action also influence mastery reporting (e.g., map structure, EEs, aligned instruction, administration fidelity, and students' interaction with the Kite Suite), as shown in Figure 10.9.

Figure 10.9

Theory of Action Inputs for Mastery Results



There are three propositions corresponding to mastery results accurately indicating what students know and can do. Table 10.12 summarizes the propositions and evidence evaluating mastery results.

Table 10.12

Propositions and Evidence for Mastery Results

Proposition	Procedural evidence	Empirical evidence	Type	Chapter(s)
Mastery status reflects students' knowledge, skills, and understandings	Documentation of mastery results, scoring method, and model fit procedures, combination of testlets,† student interaction with system,† aligned instruction†	Model fit, model parameters	Internal Structure	5
Linkage level mastery classifications are reliable	Description of reliability method	Reliability results	Internal Structure	8
Mastery results are consistent with other measures of student knowledge, skills, and understandings	Description of mastery results	Analyses on relationship of mastery results to First Contact ratings	Other Measures	4, 7

† Relies on evidence from Theory of Action input statements, as shown in Figure 10.9.

For mastery results to accurately indicate what students know and can do, linkage level mastery statuses must reflect students' knowledge, skills, and understandings. Accuracy of mastery results assumes students were assessed on the full breadth of content and at the appropriate complexity (Statement J) and that students were able to interact with the system to show their knowledge, skills, and understandings (Statement I). Accuracy also relies on earlier indirect inputs in the Theory of Action (e.g., received aligned instruction [Statement G], testlets appropriately measure linkage levels [Statement D]). Documentation describes the mastery results produced by the assessment, the diagnostic scoring method used to determine mastery, and the procedures for evaluating model fit. Overall, 98% of the estimated linkage level models showed acceptable levels of absolute model fit and/or classification accuracy. In instances of poor fit, items flagged for misfit are prioritized for retirement to improve the fit of the remaining items measuring the linkage level. At the parameter level, 93% of linkage levels have a conditional probability of masters providing a correct response greater than .6, 60% of linkage levels have a conditional probability of nonmasters providing a correct response less than .4, and 71% of linkage levels have a discrimination greater than .4. Together, the evidence indicates that mastery status reflects students' knowledge, skills, and understandings, with opportunity for continued data collection.

Linkage level mastery classifications should also be reliable. In 2022, 731 linkage levels (90%) met the recommended .7 cutoff for fair classification consistency (Johnson & Sinharay, 2018), indicating linkage level classifications are generally consistent and have low measurement error. These findings were persistent across linkage levels, which suggest there is precision of measurement across the continuum of knowledge, skills, and understandings, as defined by the five linkage levels.

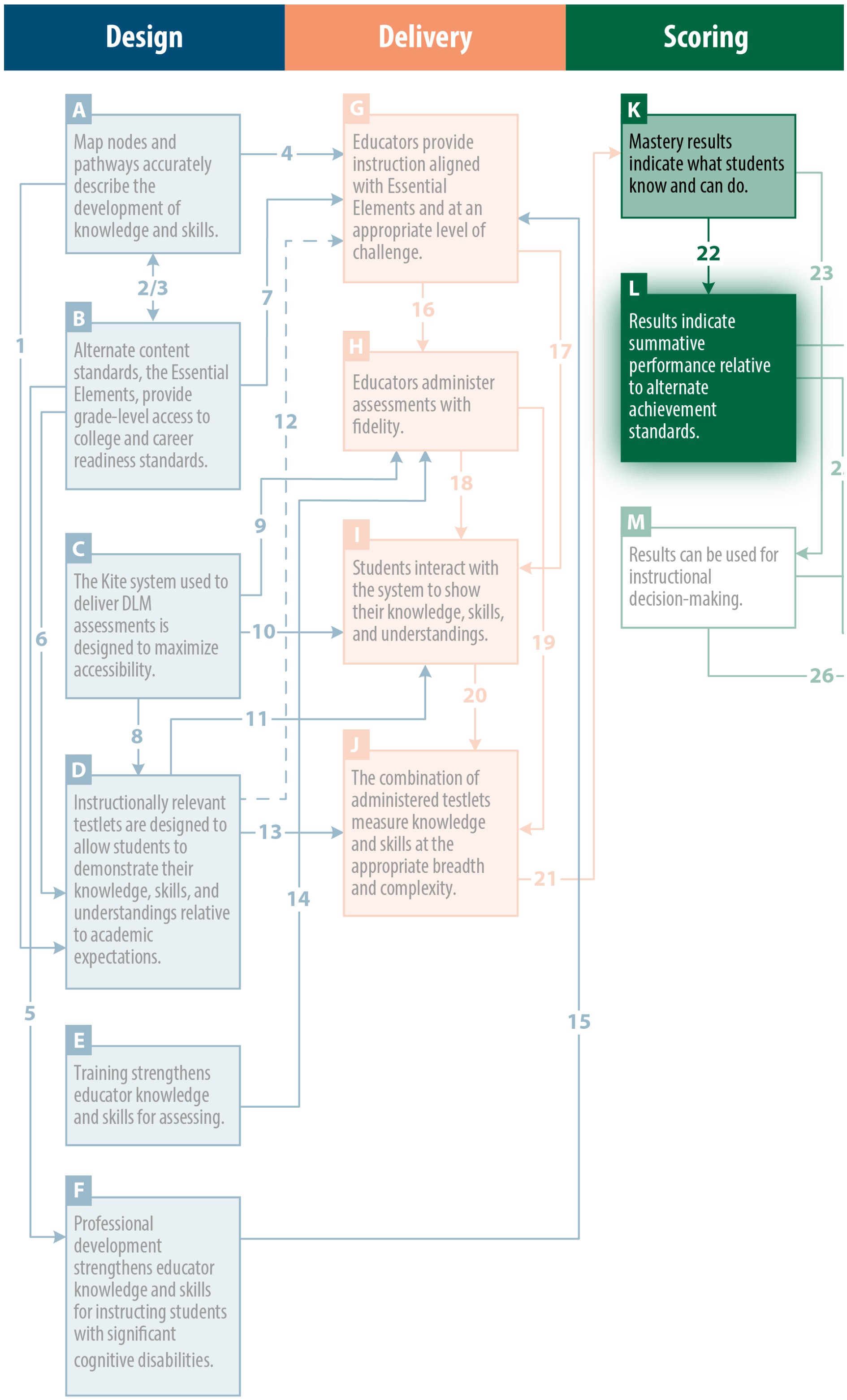
Finally, mastery results should be consistent with other measures of student knowledge, skills, and understandings. Documentation describes the grain size of the mastery results reported in the Learning Profile portion of student reports and, when students take optional instructionally embedded assessments, the mastery results displayed in the Instruction and Assessment Planner. Linkage level mastery had a moderately positive correlation with educator ratings of academic items on the First Contact survey. Together, the current evidence supports this proposition but is limited; there is opportunity for further study.

10.4.3.2. L: Results Indicate Summative Performance Relative to Alternate Achievement Standards

Mastery results are combined to summarize overall achievement in the subject, using four performance levels. Mastery results (Statement K) serve as a direct input into summative performance, as shown in Figure 10.10.

Figure 10.10

Theory of Action Inputs for Alternate Achievement Standards



Three propositions are related to the accuracy and reliability of summative performance relative to alternate achievement standards, as shown in Table 10.13.

Table 10.13

Propositions and Evidence for Alternate Achievement Standards

Proposition	Procedural evidence	Empirical evidence	Type	Chapter(s)
Performance levels provide meaningful differentiation of student achievement	Standard setting procedure, grade- and subject-specific performance level descriptors, accurate mastery results†	Standard setting survey data, performance distributions	Internal Structure, Content	6
Performance level determinations are reliable	Description of reliability method, accurate mastery results†	Reliability analyses	Internal Structure, Content	8
Performance level results are useful for communicating summative achievement in the subject to a variety of audiences	Intended uses documentation, Performance Profile development, General Research File, supplemental resources	Interviews, focus groups, governance board feedback	Consequences, Content	7

† Relies on evidence from Theory of Action input statements, as shown in Figure 10.10.

For results to indicate summative performance relative to alternate achievement standards, the performance levels must meaningfully differentiate student achievement. Performance levels are calculated from student mastery of linkage levels and rely on accurate mastery results (Statement K). Documentation describes the profile-based DLM standard setting method (Clark et al., 2017) and procedures for establishing grade- and subject-specific PLDs. Across all cut points (N = 414), panelists indicated they were comfortable with the group-recommended cut points in 94% of cases. Documentation also describes the adjustments made to the standards and the cut points for the grade- and subject-specific PLDs in 2022 because of the blueprint revisions, which reduced the total number of linkage levels available to be mastered. Using the adjusted cut points for the grade- and subject-specific PLDs, annual results for each subject demonstrate that student achievement is distributed across the four performance levels. Evidence shows that performance levels meaningfully differentiate student achievement.

Performance level determinations should also be reliable. Performance level reliability depends on accurate mastery determinations. Documentation describes the method for calculating performance level reliability (W. J. Thompson et al., 2019). In 2022, reliability results were high (e.g., polychoric correlations ranging from .789 to .928). These results indicate the DLM scoring procedure of assigning and reporting performance levels based on total linkage levels mastered produced reliable performance level determinations.

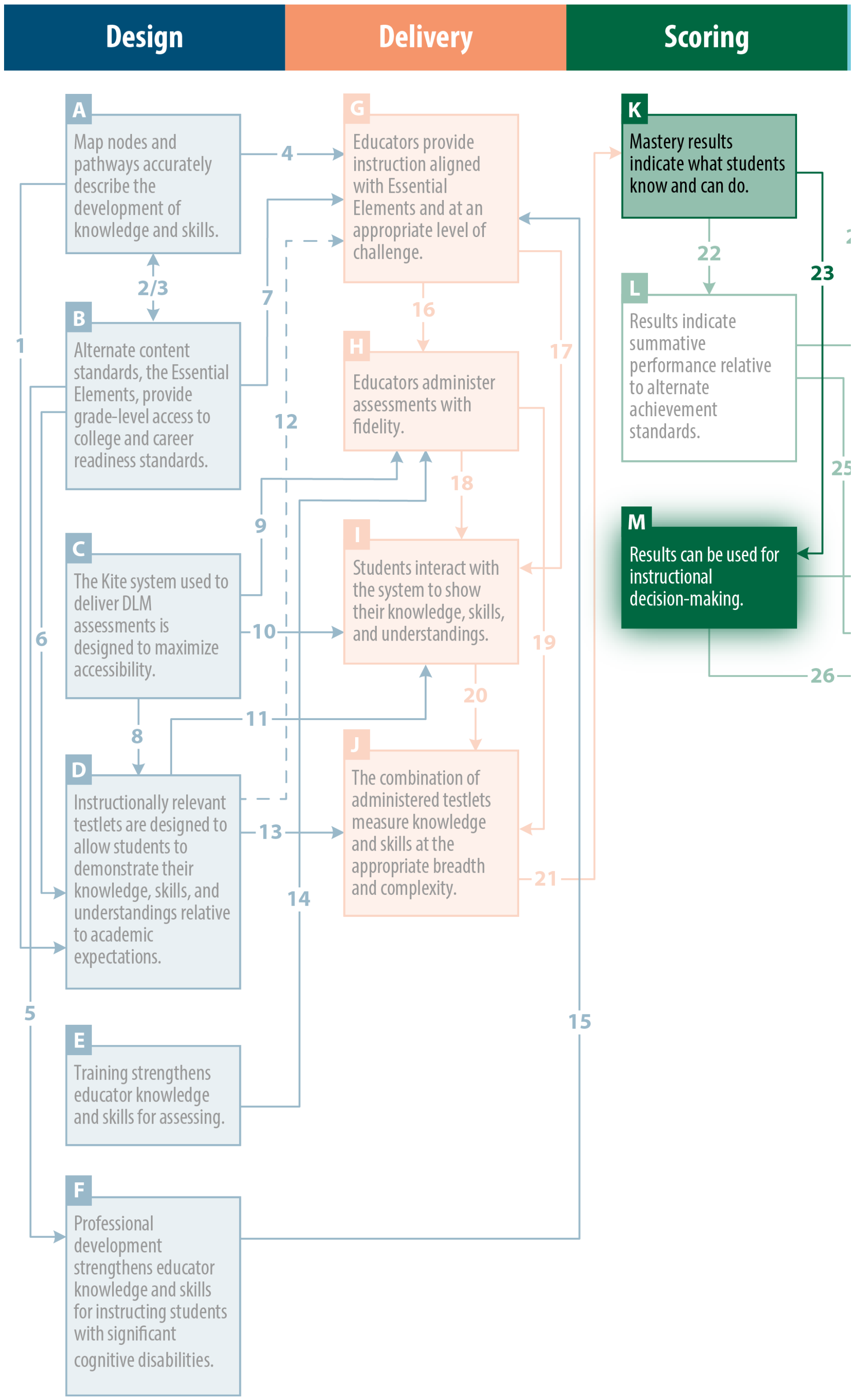
Performance level results should also be useful for communicating summative achievement in the subject to a variety of audiences. Consistent with the intended uses of DLM assessments, student performance levels were developed with the intention of communicating summative achievement to a variety of audiences. The Performance Profile portion of individual student score reports indicates the student's overall achievement in the subject, which is intended to communicate summative achievement to educators and parents. Aggregate reports, which are provided at the class, school, district, and state levels, also indicate students' performance levels. State education agencies additionally receive the General Research File, which includes each student's performance level. The General Research File can be input into state data warehouses and included in state accountability metrics. Supplemental resources support use of these files for various audiences, including parent, educator, district, and state interpretive guides and documentation. Interviews and focus groups with educators indicate the Performance Profile is useful for summarizing overall achievement and for communicating about overall achievement with parents (Clark et al., 2018; Karvonen et al., 2017). The governance board also provided feedback on the design of individual student score reports, aggregate reports, and the General Research File. Together, the evidence supports the proposition that performance level results are useful for communicating summative achievement in the subject to a variety of audiences.

10.4.3.3. M: Results Can Be Used for Instructional Decision-Making

DLM mastery results are intended to inform instructional decision-making. When educators use the optional instructionally embedded assessments, mastery results can inform instructional planning, monitoring, and adjustment. Mastery information provided on summative score reports can inform instructional planning in the subsequent academic year. The use of results for instructional decision-making is directly influenced by mastery results (Statement K). Indirect connections in the Theory of Action also influence the use of results (e.g., map structure, EEs, aligned instruction, administration fidelity, training and professional development, and students' interaction with the Kite Suite, and the combination of administered teslets), as shown in Figure 10.11.

Figure 10.11

Theory of Action Inputs for Results Being Instructionally Useful



Three propositions are related to instructional use of results. Table 10.14 summarizes the propositions and evidence evaluating the relevance of results for instruction.

Table 10.14

Propositions and Evidence for Results Being Instructionally Useful

Proposition	Procedural evidence	Empirical evidence	Type	Chapter(s)
Score reports are appropriately fine-grained	Report development process, Instruction and Assessment Planner structure	Interview data, educator cadre feedback	Consequences, Content	4, 7
Score reports are instructionally relevant and provide useful information for educators	Report development process, scoring and reporting ad hoc committee, accuracy of mastery results†	Interview data, test administrator survey	Consequences, Content	7
Educators can use results to communicate with parents about instructional plans and goal setting	Design of score reports, parent interpretive guide, Talking with Parents guide	Interview data, test administrator survey	Consequences, Content	7

† Relies on evidence from Theory of Action input statements, as shown in Figure 10.11.

For results to be instructionally useful, they should provide educators with appropriately fine-grained (i.e., specific and actionable) information. Documentation describes development and evaluation of the Learning Profile portion of score reports, which summarizes student linkage level mastery for each EE. The development process included rounds of interviews with educators providing feedback on their structure and content (Clark et al., 2018; Karvonen et al., 2017; Karvonen et al., 2016). For educators that use the optional instructionally embedded assessments, the Instruction and Assessment Planner adopts the structure of the Learning Profile to report mastery information that educators can use to inform instruction. Its design was informed by feedback collected during educator cadres. Collectively, the evidence demonstrates that score reports are appropriately fine-grained.

Score reports should also be instructionally relevant and provide useful information for educators. Documentation describes the process for designing and evaluating score report content for instructional utility, including annually convening a governance board scoring and reporting ad hoc committee to discuss score report improvements. Feedback from educator interviews generally indicates that the Learning Profile mastery results, along with the conceptual area mastery bar graphs in the Performance Profile, are instructionally relevant and provide useful information for instructional planning and goal setting (Clark et al., 2018). Evidence supports the proposition that score reports are instructionally relevant and provide useful information for educators.

Finally, educators should be able to use results to communicate with parents about instructional plans and goal setting. Resources to support communicating with parents include a parent interpretive guide and the Educator Guide for Talking With Parents. Focus group feedback indicates educators use score reports to discuss mastery results with parents, including ways parents can support student learning outside school (Clark et al., 2018). While there is opportunity for additional research, evidence to date supports the proposition that educators are able to use results to communicate with parents.

10.4.4. Long-Term Outcomes

The DLM program intends to achieve several long-term outcomes. According to the chain of reasoning in the Theory of Action, design, delivery, and scoring statements all combine to produce the following intended long-term outcomes for the DLM assessment system:

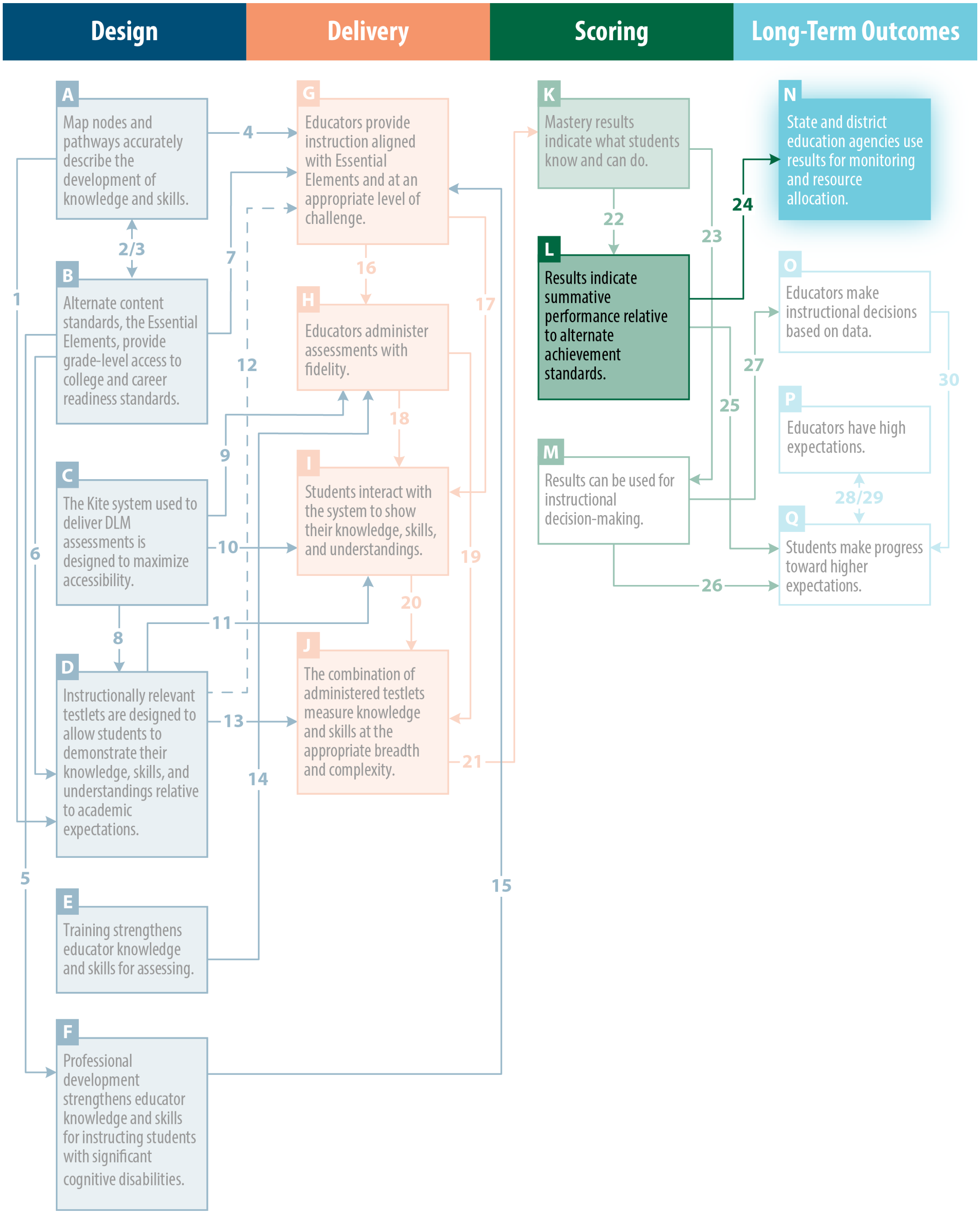
N. State and district education agencies use results for monitoring and resource allocation.
O. Educators make instructional decisions based on data.
P. Educators have high expectations.
Q. Students make progress toward higher expectations.

10.4.4.1. N: State and District Education Agencies Use Results for Monitoring and Resource Allocation

Students with significant cognitive disabilities were only recently included in accountability reporting (No Child Left Behind Act [NCLB], 2002). One of the long-term intended outcomes of the DLM program is that state and district education agencies use aggregated DLM assessment results in their decision-making processes. Use of DLM results at the district and state level relies on inputs from summative results (Statement L), as shown in Figure 10.12.

Figure 10.12

Theory of Action Inputs for State and District Use



There is one proposition for state and district education agencies' use of assessment results. Table 10.15 summarizes the evidence evaluating the proposition on state and district use of results.

Table 10.15

Propositions and Evidence for State and District Use

Proposition	Procedural evidence	Empirical evidence	Type	Chapter(s)
District and state education agency staff use aggregated information to evaluate programs and adjust resources	Available state and district reports, resources, state guidance		Consequences	7

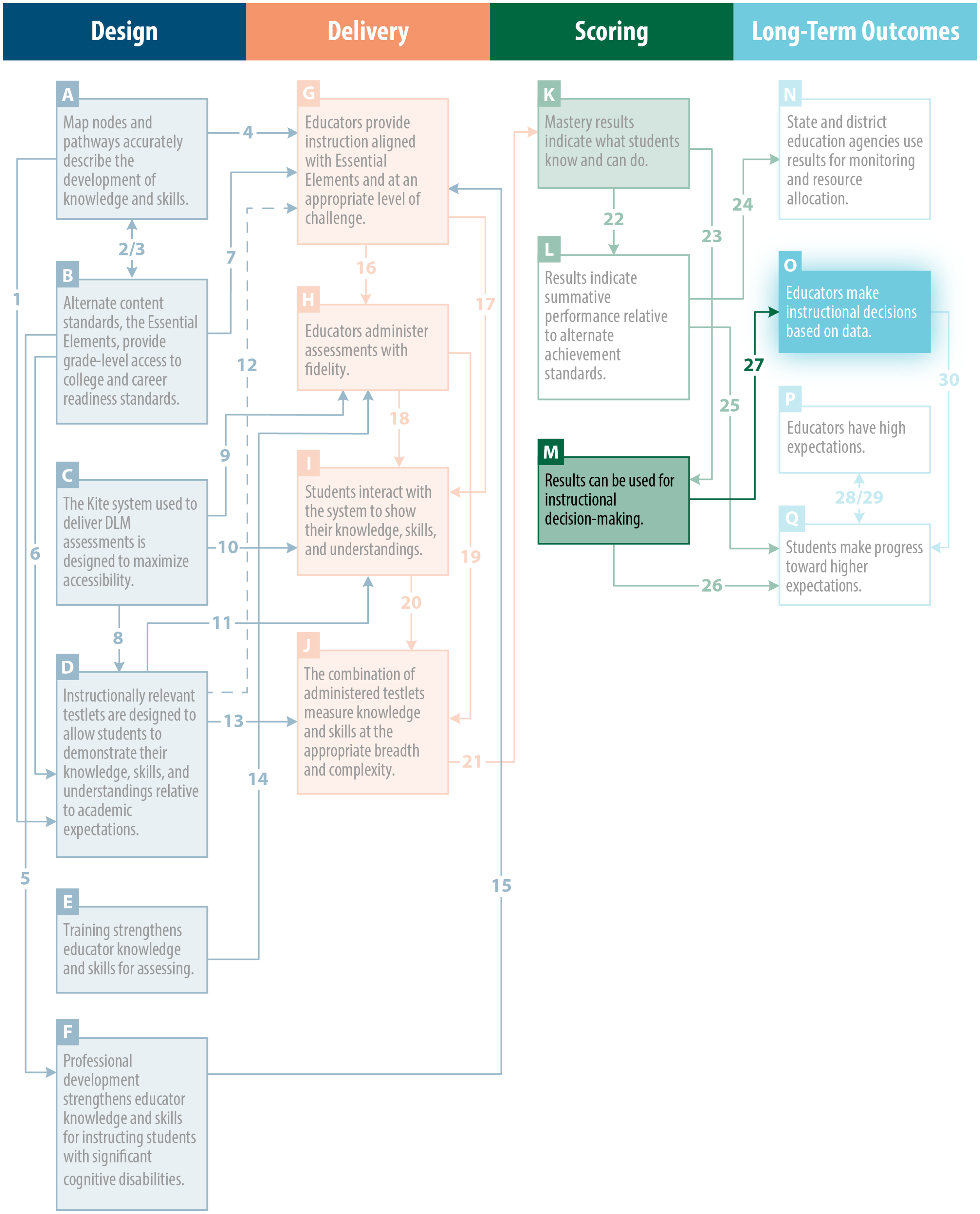
The DLM program intends for aggregated DLM results to inform educational policy decisions at the state and district level. Aggregated reports, which summarize student achievement at the class, school, district, and state levels, are available to state and district education agency staff. Supplemental resources describe the contents of reports and provide guidance on how they can be used. Individual states provide additional guidance on use of aggregated DLM assessment data for monitoring and resource allocation and are responsible for evaluating the effectiveness of the data for these purposes. Due to variation in policy and practice around use of DLM results for program evaluation and resource allocation within and across states, the DLM program presently relies on states to collect their own evidence for this proposition. DLM staff could collect data from states implementing the DLM System to strengthen the evidence for this proposition.

10.4.4.2. O: Educators Make Instructional Decisions Based on Data

Historically, score reports from alternate assessments largely classified all students as proficient and provided limited information that could inform instructional decision-making (Nitsch, 2013). One intended long-term outcome of the DLM System is that educators use DLM assessment results to make instructional decisions. The use of DLM results for making instructional decisions relies on inputs from the utility of the results (Statement M), as shown in Figure 10.13.

Figure 10.13

Theory of Action Inputs for Educators Making Instructional Decisions Based on Data



There are three propositions corresponding to educators making sound instructional decisions based on data from DLM assessments. Table 10.16 summarizes the propositions and evidence evaluating data-based instructional decision-making.

Table 10.16

Propositions and Evidence for Educators Making Instructional Decisions Based on Data

Proposition	Procedural evidence	Empirical evidence	Type	Chapter(s)
Educators are trained to use assessment results to inform instruction	*Test Administration Manual* (TAM), interpretation guides, state and local guidance, score report helplet videos	Video review feedback and use rates, focus groups	Consequences, Content	7
Educators use assessment results to inform subsequent instruction, including in subsequent academic year	Interpretation guides	Focus groups, test administrator survey	Consequences, Content	4, 7
Educators reflect on their instructional decisions	TAM, interpretation guides, score report helplet videos	Focus groups	Consequences, Content	3, 4

Educators should be trained to use DLM assessment results to inform instruction. The *Test Administration Manual* describes how to monitor student progress during the optional instructionally embedded window using mastery results in the Instruction and Assessment Planner and how to access summative individual student score reports. Documentation also describes the various resources to help educators interpret assessment results as intended, including an interpretive guide and a set of four short score report interpretation videos. State and local education agencies provide additional guidance on score report use. DLM staff conducted multiple rounds of internal review for helplet video content, followed by TAC review, and external review by educators from six states. Feedback from each group was incorporated into the final videos. However, there has been minimal use of the videos. Since 2018, the videos have had 4,851 plays and only 601 finishes. Additional information about educator training to use assessment results came from focus groups. Educators described variability in training they received on using assessment results to inform instruction, including some who received formal training from their building coordinator or district education agency, others who relied on DLM resources online, and others who received no training (Clark et al., 2018, 2022). There is opportunity for continuous improvement around educator training for using results to inform instruction.

Educators should use assessment results from the optional instructionally embedded window to inform subsequent instruction and summative assessment results to plan instructional priorities in the subsequent academic year. Interpretation guides describe using results from the optional instructionally embedded window to inform subsequent instruction. However, there is variability in the actual use of data to inform instruction. Focus groups and interviews provide some evidence that educators use DLM assessment results, including in the subsequent academic year. Educators described using the Learning Profile portion of score reports to plan instructional priorities; however, other educators reported not using the results (Clark et al., 2018). While there is some evidence to support the proposition that educators use results to inform instructional decision-making, given the variability in reported use, there is opportunity for continuous improvement in educator use of results to inform instruction.

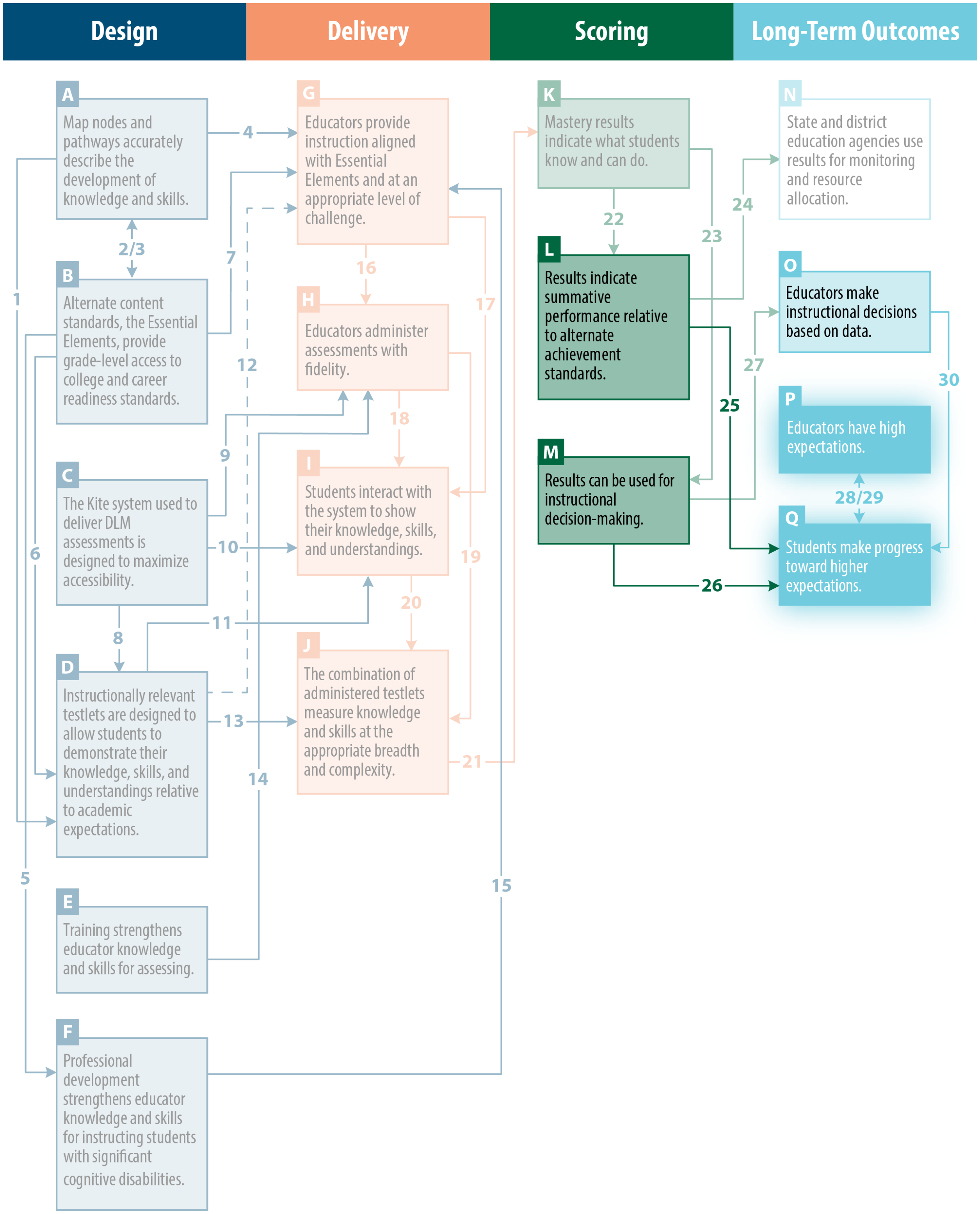
Finally, educators should use assessment results to reflect on their instructional decisions. Interpretation guides and other resources describe ways educators can reflect on their instructional decision-making. Focus group feedback indicated various ways educators reflect on their instructional decisions, including how DLM assessment data inform their decision-making (Clark et al., 2018, 2022). While current data provide evidence that some educators reflect on their instructional decisions, additional data should be collected for this proposition.

10.4.4.3. P: Educators Have High Expectations

Historically, students with significant cognitive disabilities have not been held to high academic expectations (e.g., Karvonen, Flowers, & Wakeman, 2013). As a result, the DLM program prioritizes a long-term outcome of educators having high expectations for all students who take DLM assessments. The Theory of Action demonstrates the reciprocal relationship between educators having high expectations and students making progress toward higher expectations (Statement Q) in Figure 10.14). As students demonstrate progress toward higher expectations, educators similarly increase their expectations, and increased expectations lead to students making greater progress.

Figure 10.14

Theory of Action Inputs for Educator Expectations and Student Progress



Two propositions correspond to educators' expectations for students taking DLM assessments. Table 10.17 summarizes the propositions and evidence evaluating educator expectations.

Table 10.17

Propositions and Evidence for Educators Having High Expectations

Proposition	Procedural evidence	Empirical evidence	Type	Chapter(s)
Educators believe students can attain high expectations	Map,† Essential Elements,† student interaction with system,† student progress†	Test administrator survey, focus groups	Consequences	3
Educators hold their students to high expectations	Aligned instruction†; testlet breadth and complexity†	Test administrator survey, focus groups, skill mastery survey	Consequences	4, 7

† Relies on evidence from Theory of Action input statements, as shown in Figure 10.14.

Theory of Action statements and propositions provide evidence that educators believe students with significant cognitive disabilities can attain high expectations. For instance, educators believing students can attain high expectations is supported by the map structure, rigorous grade-level expectations, allowing students to engage with the system as independently as they are able, and observing students' academic progress. Spring 2022 survey responses describe educators' perceptions of the academic content in the DLM assessments. In 2022, educators indicated that content reflected high expectations for 85% of students, which reflects a consistent annual increase since 2015, when it was 72%. Educators also indicated that testlet content measured important academic skills for 77% of students, which similarly reflects a consistent annual increase since 2015, when it was 50%. These data suggest that educator responses may reflect the awareness that the DLM assessments contain challenging content but show slightly more division on its importance. Focus group data similarly reveal variability in educator beliefs about students with significant disabilities achieving high expectations (Clark et al., 2018, 2022). The DLM program will continue to collect evidence regarding the extent that educators believe students can attain high expectations.

Educators should also hold their students to high expectations. Procedural evidence for this proposition relies heavily on Theory of Action inputs. For instance, when educators hold their students to high expectations, they provide rigorous academic instruction. A survey of 95 educators on their perceptions of their students' skill mastery revealed that many educators defined mastery as 75%–80% success on multiple trials. However, other educators perceived mastery as consistent skill demonstration over a longer period of time, generalization and transfer of skills, independence and speed of demonstration, and students' being able to explain the concept (Nehler & Clark, 2019). While these definitions differ from how mastery is determined on DLM assessments, they reflect some evidence of educators holding students to high expectations. Focus group findings reveal variability in how educators describe their expectations for students (Clark et al., 2022; Kobrin et al., 2022). While there is some evidence for this proposition, additional data collection is needed regarding educators holding all students to high expectations.

10.4.4.4. Q: Students Make Progress Toward Higher Expectations

A historic challenge for students who take alternative assessments is low expectations for their demonstration of academic skills (e.g., Timberlake, 2014). Historically, these students have been taught a largely functional curriculum intended to prepare them for independent living (Ryndak et al., 2014). Their achievement was often described as proficient, despite their instruction on relatively low-level academic skills (Altman et al., 2010; Nitsch, 2013). Therefore, an intended long-term outcome of the DLM assessment system is that students make progress toward higher expectations over time. In addition to the reciprocal relationship with educator expectations (Statement P), students making progress toward higher expectations has direct inputs from summative results (Statement L) and the utility of the results (Statement M) (Figure 10.14).

There are three propositions pertaining to students making progress toward higher expectations. The procedural and empirical evidence for these propositions, as well as evidence based on test content, constitutes evidence of the consequences of the DLM System. Table 10.18 summarizes the propositions and evidence evaluating student progress toward higher expectations.

Table 10.18

Propositions and Evidence for Student Progress Toward Higher Expectations

Proposition	Procedural evidence	Empirical evidence	Type	Chapter(s)
Alternate achievement standards are vertically articulated	Postsecondary opportunity (PSO) vertical articulation argument, grade- and subject-specific performance level descriptors (PLDs), description of PLD development process	Review of grade- and subject-specific PLDs	Content	6, 7
Students who meet alternate achievement standards are on track to pursue postsecondary opportunities	Description of PSO panel work, description of vertical articulation	PSO panel identification of academic skills, PSO panel alignment evidence	Consequences, Content	7
Students demonstrate academic progress toward higher expectations	Instruction and Assessment Planner, map,† Essential Elements,† aligned instruction,† show knowledge†	Test administrator survey, system data, focus groups	Consequences	4, 7

† Relies on evidence from Theory of Action input statements, as shown in Figure 10.14.

For students to make progress toward higher expectations, alternate achievement standards should be vertically articulated across grades. A postsecondary opportunities report (Karvonen et al., 2022) describes the vertical articulation of EEs across grades and map content from foundational to college- and career-ready expectations. The vertical articulation of DLM content structures informs vertical articulation of the achievement standards. Grade- and subject-specific PLDs similarly build across performance levels and grades, describing the types of skills students achieving at each level tend to master. Grade- and subject-specific PLDs were developed as part of the mastery profile-based standard setting method under advisement of the DLM TAC and included rounds of internal and governance board review. Collectively, the evidence indicates that the alternate achievement standards are vertically articulated.

Students who meet alternate achievement standards should be on track to pursue postsecondary opportunities. Documentation from the postsecondary opportunities study describes the procedures for convening panels to evaluate the extent to which PLDs for the DLM *At Target* achievement standards align with the skills necessary for pursuing postsecondary opportunities (postsecondary education or training, or competitive integrated employment). The panel ratings show that students who demonstrate proficiency on

DLM assessments demonstrate less complex versions of the skills necessary for postsecondary opportunities in the elementary grades, and those skills continue to develop through the middle and high school grades. Together, the evidence indicates that students who demonstrate proficiency on DLM assessments are on track to pursue postsecondary opportunities. However, as of 2022, there is a large percentage of students who do not yet demonstrate proficiency on DLM assessments.

Finally, students should demonstrate academic progress toward higher expectations. Procedural evidence for this proposition relies in part on Theory of Action inputs. For instance, students can demonstrate progress toward higher expectations due to the map structure and rigorous grade-level expectations in the EEs, receiving aligned instruction, and can demonstrate their progress by interacting with the system to show their knowledge, skills, and understandings. When educators use the optional instructionally embedded assessments, the Instruction and Assessment Planner summarizes student mastery, enabling educators to observe student progress over time. Focus group feedback provides some additional anecdotal evidence of students demonstrating progress toward higher expectations (Clark et al., 2018, 2022). Because of the challenges with reporting growth on alternate assessments (Nehler et al., 2019), under advisement of the DLM TAC, growth is not reported for DLM assessments. Additionally, comparisons of cross-year performance level distributions are cautioned, as these comparisons are based on changes in composition of independent samples for a grade and subject across years, and the overall definition of the population is changing over time (e.g., as states work toward the 1% threshold required under the Every Student Succeeds Act [ESSA], 2015). While there is some evidence of students making progress over time, additional evidence is needed to evaluate student progress toward higher expectations.

10.5. Evaluation Summary

In the three-tiered validity argument approach, we evaluate the extent to which statements in the Theory of Action are supported by the underlying propositions. Propositions are evaluated by the set of procedural and empirical evidence collected through 2021–2022. Table 10.19 summarizes the overall evaluation of the extent to which each statement in the Theory of Action is supported by the underlying propositions and associated evidence. We describe evidence according to its strength. We consider evidence for a proposition to be strong if the amount of evidence is sufficient, it includes both procedural and empirical evidence, and it is not likely to be explained by an alternative hypothesis. We consider the evidence for a proposition to be moderate if it includes only procedural evidence and/or if there is some likelihood that the evidence can be explained by an alternative hypothesis. We also note where current evidence is limited or additional evidence could be collected.

Table 10.19

Evaluation of Propositions for Each Theory of Action Statement

Statement	Overall evaluation
A. Map nodes and pathways accurately describe the development of knowledge and skills.	There is strong evidence that nodes are specified at the appropriate granularity, based on descriptions of the node development process and external review ratings. There is moderate evidence that nodes were correctly prioritized for linkage levels, based on description of the map development process, including rounds of internal review. There is strong evidence that nodes are correctly sequenced, based on the description of the procedures for specifying node connections and external review ratings, and indirect evidence from modeling analyses and alignment data for the correct ordering of linkage levels. Collectively, the propositions support the statement that map nodes and pathways accurately describe the development of knowledge and skills.
B. Alternate content standards, the Essential Elements, provide grade-level access to college and career readiness standards.	There is evidence that the grain size and description of Essential Elements (EEs) are sufficiently well-defined to communicate a clear understanding of the targeted knowledge, skills, and understandings, based on the development process and state and content expert review. There is strong evidence that EEs capture what students should know and be able to do at each grade to be prepared for postsecondary opportunities, including college, career, and citizenship, based on the development process, alignment studies, vertical articulation evidence, and indirect evidence from the postsecondary opportunities study. There is evidence that the EEs in each grade sufficiently sample the domain, based on the development process for EEs and blueprints prior to their adoption. There is strong evidence that EEs are accurately aligned to nodes in the learning maps, based on documentation of the simultaneous development process, external review ratings, and alignment study data. Collectively, the propositions support the statement that the EEs provide grade-level access to college and career readiness standards.

Table 10.19

Evaluation of Propositions for Each Theory of Action Statement (continued)

Statement	Overall evaluation
C. The Kite system used to deliver DLM assessments is designed to maximize accessibility.	There is evidence that system design is consistent with accessibility guidelines and contemporary code. There is strong evidence that supports needed by the student are available within and outside of the assessment system, based on system documentation, Personal Needs and Preferences Profile (PNP) data, test administrator survey responses, and focus groups. There is evidence that item types support the range of students in presentation and response, based on cognitive labs and test administrator survey responses. There is also evidence that the Kite Suite is accessible to educators, based on test administrator survey responses and focus group findings. Collectively, the propositions support the statement that the Kite Suite is designed to maximize accessibility.
D. Instructionally relevant testlets are designed to allow students to demonstrate their knowledge, skills, and understandings relative to academic expectations.	There is strong evidence that items within testlets are aligned to linkage levels, based on test development procedures, alignment data, and item analyses. There is generally strong evidence that testlets are designed to be accessible to students, based on test development procedures and evidence from external review, test administrator survey responses, and focus groups. There is evidence that testlets are designed to be engaging and instructionally relevant, with some opportunity for improvement, based on test development procedures and focus group feedback. There is strong evidence that testlets are written at appropriate cognitive complexity for the linkage level and that items are free of extraneous content. There is strong evidence from item analysis and differential item functioning analyses that items do not contain content that is biased against or insensitive to subgroups of the population. There is some evidence that items elicit consistent response patterns across different administration formats, with opportunity for additional data collection. Together, the propositions support the statement that instructionally relevant testlets are designed to allow students to demonstrate their knowledge, skills, and understandings relative to academic expectations.

Table 10.19

Evaluation of Propositions for Each Theory of Action Statement (continued)

Statement	Overall evaluation
E. Training strengthens educator knowledge and skills for assessing.	There is strong evidence that required training is designed to strengthen educator knowledge and skills for assessing, based on the documentation of the scope of training and passing requirements. There is some evidence that required training prepares educators to administer DLM assessments, based on survey responses, with opportunities for continuous improvement. There is strong evidence that required training is completed by all test administrators, based on Kite training status data files, state and local monitoring, and data extracts. Together, the propositions support the statement that training strengthens educator knowledge and skills for assessing.
F. Professional development strengthens educator knowledge and skills for instructing students with significant cognitive disabilities.	There is strong evidence that professional development covers topics relevant to instruction, based on the list of modules and educators' rating of the content. There is some evidence that educators access the professional development modules, based on the module completion data. There is some evidence that educators implement the practices on which they have been trained, but overall use of professional development modules is low. While there is opportunity for continuous improvement in the use of professional development modules and their application to instructional practice, when propositions are fulfilled (i.e., professional development is used), professional development strengthens educator knowledge and skills for instructing students with significant cognitive disabilities.
G. Educators provide instruction aligned with Essential Elements and at an appropriate level of challenge.	Evidence that educators provide students the opportunity to learn content aligned with the grade-level EEs shows variable results, based on responses to the test administrator survey and First Contact survey. There is some evidence that educators provide instruction at an appropriate level of challenge using their knowledge of the student, based on test administrator survey responses and focus groups. However, there is also some evidence from the opportunity to learn section of the test administrator survey that some students may not be receiving instruction aligned with the full breadth of academic content measured by the DLM assessment. There is a need for additional data collection on instructional practice. Together, the propositions provide some support that educators provide instruction aligned with EEs and at an appropriate level of challenge.

Table 10.19

Evaluation of Propositions for Each Theory of Action Statement (continued)

Statement	Overall evaluation
H. Educators administer assessments with fidelity.	There is strong evidence that educators are trained to administer testlets with fidelity, based on training documentation, test administrator survey responses, and test administration observations (TAOs). Documentation describes entry of accessibility supports, which is supported by PNP data from the system, but currently, there is no evidence available to evaluate the accuracy of accessibility supports enabled for students or their consistency with supports used during instruction. There is some evidence that educators allow students to engage with the system as independently as they are able, based on TAOs, test administrator survey responses, and PNP data, with the opportunity to collect additional data. There is evidence that educators enter student responses with fidelity, based on TAOs and writing interrater agreement studies. Overall, available evidence for the propositions indicates that educators administer assessments with fidelity, with opportunity for additional data collection.
I. Students interact with the system to show their knowledge, skills, and understandings.	There is evidence that students can respond to tasks regardless of sensory, mobility, health, communication, or behavioral constraints, based on test administrator survey responses, PNP selection data, TAOs, and focus groups. There is evidence from cognitive labs, test administrator survey responses, and TAOs that student responses to items reflect their knowledge, skills, and understandings and that students can interact with the system as intended. Evidence for the propositions collectively demonstrates that students interact with the system to show their knowledge, skills, and understandings.
J. The combination of administered testlets measure knowledge and skills at the appropriate breadth and complexity.	There is generally strong evidence that the First Contact survey correctly assigns students to complexity bands, based on pilot analyses, with the opportunity for additional research. There is strong evidence that administered testlets cover the full blueprint based on blueprint coverage data and Special Circumstance codes. There is also generally strong evidence that administered testlets are at the appropriate linkage level, with the opportunity to collect additional data about adaptation. Overall, the propositions moderately support the statement that the combination of administered testlets is at the appropriate breadth and complexity.

Table 10.19

Evaluation of Propositions for Each Theory of Action Statement (continued)

Statement	Overall evaluation
K. Mastery results indicate what students know and can do.	There is strong evidence that mastery status reflects students' knowledge, skills, and understandings, based on modeling evidence, and there is strong evidence that linkage level mastery classifications are reliable based on reliability analyses. There is some evidence that mastery results are consistent with other measures of student knowledge, skills, and understandings, but additional evidence is needed. Overall, the propositions support the statement that mastery results indicate what students know and can do.
L. Results indicate summative performance relative to alternate achievement standards.	There is strong evidence that performance levels meaningfully differentiate student achievement, based on the standard setting procedure, standard setting survey data, and performance distributions. There is strong evidence that performance level determinations are reliable, based on documentation and reliability analyses. The propositions support the statement that results indicate summative performance relative to alternate achievement standards.
M. Results can be used for instructional decision-making.	There is strong evidence that reports are fine-grained, based on documentation of score report development and on interview and focus group data. There is evidence that score reports are instructionally relevant and useful and that they provide relevant information for educators, based on documentation of the report development process, interview, and test administrator survey responses. There is evidence of variability in training on how educators can use results to inform instruction, based on training content and focus groups. The evidence suggests there is opportunity to collect additional data. There is evidence that educators can use results to inform instructional choices and goal setting, based on documentation of the report development, interview data, focus groups, and test administrator survey responses. There is evidence that educators can use results to communicate with parents, but there is variability in actual practice. Additional data should be collected. Overall, the propositions support results being useful for instructional decision-making.
N. State and district education agencies use results for monitoring and resource allocation.	There is some procedural evidence that district and state staff use aggregated information to evaluate programs and adjust resources. Because states and district policies vary regarding how results should be used for monitoring and resource allocation, states are responsible for collecting their own evidence for this proposition. Additional evidence could be collected from across state education agencies.

Table 10.19

Evaluation of Propositions for Each Theory of Action Statement (continued)

Statement	Overall evaluation
O. Educators make instructional decisions based on data.	There is some evidence that educators are trained to use assessment results to inform instruction, based on available resources, video use rates and feedback, and focus group feedback, with some variability in use. There is some evidence that educators use assessment results to inform instruction, based on feedback from focus groups and the test administrator survey. There is some evidence that educators reflect on their instructional decisions, based on focus group feedback. Evidence for each proposition provides some support for the statement that some educators make instructional decisions based on data, but variability indicates that not all educators receive training or use data to inform instruction. Additional evidence collection and continuous improvement in implementation would strengthen the propositions and provide greater support for this long-term outcome of the system.
P. Educators have high expectations.	There is some evidence that educators believe students can attain high expectations and hold their students to high expectations. Survey and focus group responses show variability in educator perspectives. There is opportunity for continued data collection of educators' understanding of high expectations as defined in the DLM System. Presently, the propositions provide some support for the statement that educators have high expectations for their students.
Q. Students make progress toward higher expectations.	There is strong evidence that the alternate achievement standards are vertically articulated, based on the performance level descriptor development process and postsecondary opportunities vertical articulation argument. There is strong evidence, based on the postsecondary opportunities alignment evidence, that students who meet alternate achievement standards are on track to pursue postsecondary opportunities. However, there is a large percentage of students who do not yet demonstrate proficiency on DLM assessments. There is some evidence showing students make progress toward higher expectations based on input from prior Theory of Action statements and focus group data. There are complexities in reporting growth for DLM assessments, and evaluating within-year progress relies on educator use of optional instructionally embedded assessments, which to date is low. Evidence collected for the propositions to date provides some support for the statement that students make progress toward higher expectations.

10.6. Continuous Improvement

The DLM program is committed to continuous improvement of assessments, educator and student experiences, and technological delivery of the assessment system. As described in Chapter 1, the DLM Theory of Action guides ongoing research, development, and continuous improvement. Through formal research and evaluation as well as informal feedback, guided by the DLM Governance Board, the TAC, and others, the DLM program has made many improvements since its launch in 2015, including the blueprint and test pool revision in 2019–2020. This section describes examples of improvements related to the design, delivery, and scoring of DLM assessments to support achieving the program's intended long-term outcomes.

10.6.1. Design Improvements

The Learning Maps and Test Development teams continually improve learning map development and item writing and testlet development processes. They use multiple sources of information from the field, research findings, and data collected throughout the school year. For example, in January 2018, the item writing process shifted from having all items written on site to a hybrid model involving both on-site and remote (online) activities. This hybrid approach is more efficient and maintains the high quality of items written. ATLAS staff have also refined procedures for retiring content from testlet pools over time. Model fit analyses, item-level differential item functioning evaluations, and content reviews are used to prioritize items and testlets for retirement. The aim of these retirement processes is to systematically refresh the operational pool with high-quality items.

10.6.2. Delivery Improvements

Improvements to test delivery and administration procedures focus on promoting accessibility, accurate delivery of testlet assignments, and a high-quality assessment experience for educators and students. Continuous improvements to the functionality of the Kite Suite have resulted in several changes in recent years. Staff annually prioritize updates to Educator Portal and Student Portal to improve user experience, drawing from input from the governance board and educator feedback from surveys and cadre groups where applicable.

10.6.3. Scoring and Reporting Improvements

DLM staff also make continuous improvements to scoring and reporting. The DLM TAC, including a modeling subcommittee of TAC members, regularly provides feedback on methodological psychometric topics. Examples of improvements based on this work include methods for evaluating model- and item-level fit and research to evaluate the ordering of linkage levels within EEs. Feedback from educator cadres and focus groups, along with a scoring and reporting ad hoc committee of DLM Governance Board members, has led to minor adjustments in score report contents (e.g., adding a link on score reports that directs users to additional resources on the DLM website) and to the creation of helplet videos to support score report interpretation and use.

10.7. Future Research

The evaluation of evidence for the statements in the Theory of Action identifies areas for future research. DLM staff will plan future research to collect additional evidence for propositions where current evidence is moderate or limited. We describe areas for future research throughout this technical manual. Longitudinal data collection is ongoing as part of the regular operations of the assessment system. As the DLM System continues to mature, additional evidence will be collected on a continuous basis. The DLM Governance Board will continue to collaborate on additional data collection as needed. Future studies will be guided by advice from the DLM TAC, using processes established over the life of the DLM System. Some examples of future research are described here.

In the area of design, DLM staff will conduct additional research on methods for evaluating the structure of the learning maps to provide a more comprehensive evaluation of the hierarchical ordering of linkage levels. In the area of delivery, DLM staff are examining new ways to gather information on students' opportunity to learn in order to further evaluate the extent to which educators provide aligned instruction. In the area of scoring, the spring 2023 test administrator survey will collect information on educator ratings of student mastery as additional evidence to evaluate the extent that mastery ratings are consistent with other measures of student knowledge, skills, and understandings. To evaluate long-term outcomes, the annual test administrator survey will continue to provide a source of data from which to investigate changes over time in the long-term effects of the assessment system for students and educators.

11. References

Accessible Teaching, Learning, and Assessment Systems. (2021). *2020–2021 DLM administration during COVID-19: Participation, performance, and educational experience* (Technical Report No. 21-02). University of Kansas. https://dynamiclearningmaps.org/sites/default/files/documents/publication/DLM-COVID.pdf

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian networks in educational assessment*. Springer. https://doi.org/10.1007/978-1-4939-2125-6

Altman, J. R., Lazarus, S. S., Quenemoen, R. F., Kearns, J., Quenemoen, M., & Thurlow, M. L. (2010). *2009 survey of states: Accomplishments and new issues at the end of a decade of change*. University of Minnesota, National Center on Educational Outcomes. https://files.eric.ed.gov/fulltext/ED511742.pdf

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

Babu, G. J. (2011). Resampling methods for model fitting and model selection. *Journal of Biopharmaceutical Statistics*, *21*(6), 1177–1186. https://doi.org/10.1080/10543406.2011.607749

Bechard, S., Clark, A. K., Swinburne Romine, R., Karvonen, M., Kingston, N., & Erickson, K. (2019). Use of evidence-centered design to develop learning maps-based assessments. *International Journal of Testing*, *19*(2), 188–205. https://doi.org/10.1080/15305058.2018.1543310

Bechard, S., Hess, K., Camacho, C., Russell, M., & Thomas, K. (2012). Why should cognitive learning models be used as the foundation for designing next generation assessment systems? In R. Cameto, S. Bechard, & P. Almond (Eds.), *Understanding learning progressions and learning maps to inform the development of assessment for students in special populations* [White paper]. SRI International Symposium, Menlo Park, CA.

Bechard, S., & Sheinker, A. (2012). *Basic framework for item writers using Evidence Centered Design (ECD)*. University of Kansas, Center for Educational Testing and Evaluation.

Betancourt, M. (2018, July 15). *A conceptual introduction to Hamiltonian Monte Carlo*. arXiv. http://arxiv.org/abs/1701.02434

Bradshaw, L. (2016). Diagnostic classification models. In A. A. Rupp & J. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (1st ed., pp. 297–327). John Wiley & Sons. https://doi.org/10.1002/9781118956588.ch13

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice*, *33*(1), 2–14. https://doi.org/10.1111/emip.12020

Bradshaw, L., & Levy, R. (2019). Interpreting probabilistic classifications from diagnostic psychometric models. *Educational Measurement: Issues and Practice*, *38*(2), 79–88. https://doi.org/10.1111/emip.12247

Camilli, G., & Shepard, L. A. (1994). *Method for Identifying Biased Test Items* (4th). SAGE Publications, Inc.

Carlin, B. P., & Louis, T. A. (2001). Empirical Bayes: Past, present and future. In A. E. Raftery, M. A. Tanner, & M. T. Wells (Eds.), *Statistics in the 21st century*. Chapman and Hall/CRC. https://doi.org/10.1201/9781420035391

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, *39*(2), 83–87. https://doi.org/10.2307/2682801

Center for Applied Special Technology. (2018, January 20). *Universal design for learning guidelines version 2.2*. https://udlguidelines.cast.org

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, *8*(4), 293–332. https://doi.org/10.1207/s1532690xci0804_2

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*(2), 123–140. https://doi.org/10.1111/j.1745-3984.2012.00185.x

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*, 551–558. https://doi.org/10.1016/0895-4356(90)90159-M

Cizek, G. J. (1996). Standard-setting guidelines. *Educational Measurement: Issues and Practice*, *15*(1), 13–21. https://doi.org/10.1111/j.1745-3992.1996.tb00802.x

Cizek, G. J., & Bunch, M. B. (2006). *Standard setting: A guide to establishing and evaluating performance standards on tests*. SAGE Publications, Inc. https://doi.org/10.4135/9781412985918

Clark, A. K., Karvonen, M., Kingston, N., Anderson, G., & Wells-Moreaux, S. (2015, April 15–19). *Designing alternate assessment score reports that maximize instructional impact*. National Council on Measurement in Education Annual Meeting, Chicago, IL. https://dynamiclearningmaps.org/sites/default/files/documents/presentations/NCME_2015_Score_Report_Utility_Paper.pdf

Clark, A. K., Karvonen, M., Swinburne Romine, R., & Kingston, N. (2018, April 12–16). *Teacher use of score reports for instructional decision-making: Preliminary findings*. National Council on Measurement in Education Annual Meeting, New York, NY. https://dynamiclearningmaps.org/sites/default/files/documents/presentations/NCME_2018_Score_Report_Use_Findings.pdf

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

Clark, A. K., Kobrin, J., & Hirt, A. (2022). *Educator perspectives on instructionally embedded assessment* (Research Synopsis No. 22-01). University of Kansas, Accessible Teaching, Learning, and Assessment Systems. https://dynamiclearningmaps.org/sites/default/files/documents/publication/IE_Focus_Groups_project_brief.pdf

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

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037//0033-2909.112.1.155

Confrey, J., Gianopulos, G., McGowan, W., Shah, M., & Belcher, M. (2017). Scaffolding learner-centered curricular coherence using learning maps and diagnostic assessments designed around mathematics learning trajectories. *ZDM – Mathematics Education*, *49*(5), 717–734. https://doi.org/10.1007/s11858-017-0869-1

Confrey, J., Maloney, A. P., & Corely, A. K. (2014). Learning trajectories: A framework for connecting standards with curriculum. *ZDM – Mathematics Education*, *46*(5), 719–733. https://doi.org/10.1007/s11858-014-0598-7

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton University Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1177/0146621612445470

DeBarger, A. H., Seeratan, K., Cameto, R., Haertel, G., Knokey, A.-M., & Morrison, K. (2011). *Alternate assessment design-mathematics* (Technical Report 9: Implementing evidence-centered design to develop assessments for students with significant cognitive disabilities: Guidelines for creating design patterns and development specifications and exemplar task templates for mathematics). SRI International. Washington D. C. http://alternateassessmentdesign.sri.com/techreports/AAD_M_TechRpt9_032911final.pdf

Dennis, A., Erickson, K., & Hatch, P. (2013). *The Dynamic Learning Maps core vocabulary: Overview* (Technical Review). https://www.dropbox.com/s/99ay2ypx37m6lps/DLM_Core_Vocabulary_Overview.pdf

DLM Consortium. (2021a). *Test Administration Manual 2021–2022*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems.

Dynamic Learning Maps Consortium. (2014). *Guide to external review of testlets*. University of Kansas, Center for Educational Testing and Evaluation. https://dynamiclearningmaps.org/sites/default/files/documents/External_Review/External_Review_Guide_IM.pdf

Dynamic Learning Maps Consortium. (2021b). *Accessibility Manual 2021–2022*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems.

Dynamic Learning Maps Consortium. (2021c). *Guide to DLM Required Test Administrator Training*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems.

Dynamic Learning Maps Consortium. (2022a). *2021–2022 Technical Manual Update—Science*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems.

Dynamic Learning Maps Consortium. (2022b). *Technology Specifications Manual 2021–2022*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems.

Efron, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statistical Science*, *29*(2), 285–301. https://doi.org/10.1214/13-STS455

Erickson, K. A., Hatch, P., & Clendon, S. (2010). Literacy, assistive technology, and students with significant disabilities. *Focus on Exceptional Children*, *42*(5), 1–17. https://doi.org/10.17161/foec.v42i5.6904

Erickson, K. A., & Karvonen, M. (2014, July 21–23). *College and career readiness instruction and assessment for pre-intentional and pre-symbolic communicators*. Office of Special Education Programs Project Directors' Meeting, Washington, D. C.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data (revised edition)*. MIT Press. https://doi.org/10.7551/mitpress/5657.001.0001

Every Student Succeeds Act, 20 U.S.C § 6301 (2015). https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

Falconer, J. R., Frank, E., Polaschek, D. L. L., & Joshi, C. (2022). Methods for eliciting informative prior distributions: A critical review. *Decision Analysis*. https://doi.org/10.1287/deca.2022.0451

Family Educational Rights and Privacy Act, 20 U.S.C § 1232g; 34 CFR Part 99 (1974). https://www.govinfo.gov/content/pkg/USCODE-2020-title20/pdf/USCODE-2020-title20-chap31-subchapIII-part4-sec1232g.pdf

Fang, G., Liu, J., & Ying, Z. (2019). On the identifiability of diagnostic classification models. *Psychometrika*, *84*(1), 19–40. https://doi.org/10.1007/s11336-018-09658-x

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*, 543–549. https://doi.org/10.1016/0895-4356(90)90158-L

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions* (2nd). John Wiley.

Flowers, C., Turner, C., Herrera, B., Towles-Reeves, L., Thurlow, M., Davidson, A., & Hagge, S. (2015). *Developing a large-scale assessment using components of evidence-centered design: Did it work?* [Paper presentation]. National Council on Measurement in Education.

Flowers, C., & Wakeman, S. (2016a). *Alignment of Dynamic Learning Maps operational items to grade-level content standards: Year-End models of testing*. ACERI Partners, LLC.

Flowers, C., & Wakeman, S. (2016b). *Alignment of Dynamic Learning Maps operational items to grade-level content standards: Integrated model for testing*. ACERI Partners, LLC.

Flowers, C., & Wakeman, S. (2020). *Alignment of Dynamic Learning Maps operational items to grade-level content standards: Year-End model of testing*. ACERI Partners, LLC.

Flowers, C., Wakeman, S., Browder, D. M., & Karvonen, M. (2009). Links for academic learning (LAL): A conceptual model for investigating alignment of alternate assessments based on alternate achievement standards. *Educational Measurement: Issues and Practice*, *28*(1), 25–37. https://doi.org/10.1111/j.1745-3992.2009.01134.x

Gentry, J. R. (1982). An analysis of developmental spelling in ”GNYS AT WRK”. *The Reading Teacher*, *36*(2), 192–200. https://www.jstor.org/stable/20198182

Goldstein, J., & Behuniak, P. (2010). Assumptions in alternate assessment: An argument-based approach to validation. *Assessment for Effective Intervention*, *36*(3), 179–191. https://doi.org/10.1177/1534508410392208

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282. https://doi.org/10.1007/BF02288892

Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 47–76). Routledge. https://doi.org/10.4324/9780203848203

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*(4), 262–277. https://doi.org/10.1177/0146621604272623

Henson, R., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210. https://doi.org/10.1007/s11336-008-9089-5

Hess, K. K. (2012). *Learning progressions in K–8 classrooms: How progress maps can influence classroom practice and perceptions and help teachers make more informed instructional decisions in support of struggling learners* (Synthesis Report No. 87). University of Minnesota, National Center on Educational Outcomes. https://nceo.umn.edu/docs/OnlinePubs/Synthesis87/SynthesisReport87.pdf

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. http://www.jstor.org/stable/4615733

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power raters using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329–349. https://doi.org/10.1207/S15324818AME1404_2

Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement*, *55*(4), 635–664. https://doi.org/10.1111/jedm.12196

Johnstone, C., Altman, J. R., & Moore, M. (2011). Universal design and the use of cognitive labs. In M. Russell & M. Kavanaugh (Eds.), *Assessing students in the margin: Challenges, strategies, and techniques* (pp. 425–442). Information Age Publishing.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Prager.

Karvonen, M., Bechard, S., & Wells-Moreaux, S. (2015, April 16–20). *Accessibility considerations for students with significant cognitive disabilities who take computer-based alternate assessments* [Paper presentation]. American Educational Research Association Annual Meeting, Chicago, IL.

Karvonen, M., Flowers, C., & Wakeman, S. (2009, April 13–17). *Predictors of access to the general curriculum for students with significant cognitive disabilities* [Paper presentation]. American Educational Research Association Annual Meeting, San Diego, CA.

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

Karvonen, M., Swinburne Romine, R., Clark, A. K., Brussow, J., & Kingston, N. M. (2017, April 26–30). *Promoting accurate score report interpretation and use for instructional planning*. National Council on Measurement in Education Annual Meeting, San Antonio, TX. https://dynamiclearningmaps.org/sites/default/files/documents/presentations/NCME_2017_Score_Reports.pdf

Karvonen, M., Wakeman, S., Flowers, C., & Moody, S. (2013). The relationship of teachers' instructional decisions and beliefs about alternate assessments of student achievement. *Exceptionality*, *21*(4), 238–252. https://doi.org/10.1080/09362835.2012.747184

Karvonen, M., Wakeman, S. Y., Browder, D. M., Rogers, M. A. S., & Flowers, C. (2011). *Academic curriculum for students with significant cognitive disabilities: Special education teacher perspectives a decade after IDEA 1997* (Research Report). National Alternate Assessment Center. https://files.eric.ed.gov/fulltext/ED521407.pdf

Karvonen, M., Clark, A. K., & Kavitsky, L. (2022). *Aligned academic achievement standards to support pursuit of postsecondary opportunities: Year-End model* (Technical Report No. 22-01). University of Kansas, Accessible Teaching, Learning, and Assessment Systems. https://dynamiclearningmaps.org/sites/default/files/documents/publication/YE_PSO_technical_report.pdf

Karvonen, M., Clark, A. K., & Kingston, N. M. (2016, April 1–11). *Alternate assessment score report interpretation and use: Implications for instructional planning*. National Council on Measurement in Education Annual Meeting, Washington, DC. https://dynamiclearningmaps.org/sites/default/files/documents/presentations/NCME_2016_DLM_Score_Report.pdf

Karvonen, M., Clark, A. K., & Nash, B. (2015). *2015 Year-End model standard setting: English language arts and mathematics* (Technical Report No. 15-03). University of Kansas; Center for Educational

Testing and Evaluation. https://dynamiclearningmaps.org/sites/default/files/documents/publication/Standard_Setting_Tech_Report_YE.pdf

Kearns, J. F., Towles-Reeves, E., Kleinert, H. L., Kleinert, J. O. R., & Thomas, M. K.-K. (2011). Characteristics of and implications for students participating in alternate assessments based on alternate academic achievement standards. *The Journal of Special Education*, *45*(1), 3–14. https://doi.org/10.1177/0022466909344223

Kingston, N., & Tiemann, G. C. (2012). Setting performance standards on complex assessments: The body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 201–224). Routledge. https://doi.org/10.4324/9780203848203

Kleinert, H. L., Browder, D. M., & Towles-Reeves, E. A. (2009). Models of cognition for students with significant cognitive disabilities: Implications for assessment. *Review of Educational Research*, *79*(1), 301–326. https://doi.org/10.3102/0034654308326160

Kobrin, J., Clark, A. K., & Kavitsky, E. (2022). *Exploring educator perspectives on potential accessibility gaps in the Dynamic Learning Maps alternate assessment* (Research Synopsis No. 22-02). University of Kansas, Accessible Teaching, Learning, and Assessment Systems. https://dynamiclearningmaps.org/sites/default/files/documents/publication/Accessibility_Focus_Groups_project_brief.pdf

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511611186

Linn, R. L. (2009). The concept of validity in the context of NCLB. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 195–212). Information Age Publishing, Inc.

Lissitz, R. W. (2009). Introduction. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 1–18). Information Age Publishing, Inc.

Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, *45*(1), 503–528. https://doi.org/10.1007/BF01589116

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187–212. https://doi.org/10.1007/BF02294535

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). American Council on Education and Macmillan.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 83–110). Information Age Publishing, Inc.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (Research Report No. RR-03-16). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Mislevy, R. J., & Riconscente, M. M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology* (PADI Technical Report No. 9). SRI International. https://padi.sri.com/downloads/TR9_ECD.pdf

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *Evidence-centered assessment design*. Educational Testing Service. https://learnlab.org/research/wiki/images/5/51/Evidence-Centered-Assessment-Design.pdf

Mislevy, R. J., & Gitomer, D. H. (1995). The role of probability-based inference in an intelligent tutoring system. *User Modeling and User-Adapted Interaction*, *5*(3–4), 253–282. https://doi.org/10.1007/BF01126112

Moinester, M., & Gottfried, R. (2014). Sample size estimation for correlations with pre-specified confidence interval. *The Quantitative Methods for Psychology*, *10*(2), 124–130. https://doi.org/10.20982/tqmp.10.2.p124

Nabi, S., Nassif, H., Hong, J., Mamani, H., & Imbens, G. (2022). Bayesian meta-prior learning using empirical Bayes. *Management Science*, *68*(3), 1737–1755. https://doi.org/10.1287/mnsc.2021.4136

Nash, B., Clark, A. K., & Karvonen, M. (2016). *First Contact: A census report on the characteristics of students eligible to take alternate assessments*. University of Kansas, Center for Educational Testing and Evaluation. https://dynamiclearningmaps.org/sites/default/files/documents/publication/First_Contact_Census_2016.pdf

National Council on Measurement in Education. (2018). *Position statement on theories of action for testing programs*. https://higherlogicdownload.s3.amazonaws.com/NCME/c53581e4-9882-4137-987b-4475f6cb502a/UploadedImages/Documents/NCME_Position_Paper_on_Theories_of_Action_-_Final_July__2018.pdf

National Governors Association Center for Best Practices and Council of Chief State School Officers. (2010). *Common core state standards*.

Neal, R. (2011, May 10). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (Vol. 20116022). Chapman and Hall/CRC. https://doi.org/10.1201/b10905-6

Nehler, C., & Clark, A. K. (2019). *Teacher rating of student mastery: Pilot* (Research Synopsis No. 19-02). University of Kansas, Accessible Teaching, Learning, and Assessment Systems. https://dynamiclearningmaps.org/sites/default/files/documents/publication/Teacher_Rating_of_Student_Mastery.pdf

Nehler, C., Clark, A. K., & Karvonen, M. (2019). *White paper: Considerations for measuring academic growth on Dynamic Learning Maps (DLM) alternate assessments*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems.

Nippold, M. A. (2007). *Later language development: School-age children, adolescents, and young adults* (3rd ed.). Pro Ed.

Nitsch, C. (2013). *Dynamic Learning Maps: The Arc parent focus groups*. The Arc. https://dynamiclearningmaps.org/sites/default/files/documents/publication/TheArcParentFocusGroups.pdf

No Child Left Behind Act, 20 U.S.C. § 6319 (2002). https://www.congress.gov/107/plaws/publ110/PLAW-107publ110.pdf

Nocedal, J., & Wright, S. J. (2006). *Numerical optimization*. Springer. https://doi.org/10.1007/978-0-387-40065-5

O'Leary, S., Lund, M., Ytre-Hauge, T. J., Holm, S. R., Naess, K., Dalland, L. N., & McPhail, S. M. (2014). Pitfalls in the use of kappa when interpreting agreement between multiple raters in reliability studies. *Physiotherapy*, *100*, 27–35. https://doi.org/10.1016/j.physio.2013.08.002

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann. https://doi.org/10.1016/C2009-0-27609-4

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice*, *27*(4), 15–29. https://doi.org/10.1111/j.1745-3992.2008.00135.x

Petrone, S., Rousseau, J., & Scricciolo, C. (2014). Bayes and empirical Bayes: Do they merge? *Biometrika*, *101*(2), 285–302. https://doi.org/10.1093/biomet/ast067

Pontius, R. G., Jr., & Millones, M. (2011). Death to kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, *32*, 4407–4429. https://doi.org/10.1080/01431161.2011.552923

Popham, W. J. (2011, April 8–12). *How to build learning progressions: Keep them simple, Simon!* American Educational Research Association Annual Meeting, New Orleans, LA.

Ravand, H., & Baghaei, P. (2020). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing*, *20*(1), 24–56. https://doi.org/10.1080/15305058.2019.1588278

Rupp, A. A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Ryndak, D. L., Taub, D., Jorgensen, C. M., Gonsier-Gerdin, J., Arndt, K., Sauer, J., Ruppar, A. L., Morningstar, M. E., & Allcock, H. (2014). Policy and the impact on placement, involvement, and progress in general education: Critical issues that require rectification. *Research and Practice for Persons with Severe Disabilities*, *39*(1), 65–74. https://doi.org/10.1177/1540796914533942

Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research: Learning trajectories for young children*. Routledge. https://doi.org/10.4324/9780203883785

Shepard, L. A. (2018). Learning progressions as tools for assessment and learning. *Applied Measurement in Education*, *31*, 165–174. https://doi.org/10.1080/08957347.2017.1408628

Sinharay, S., & Johnson, M. S. (2019). Measures of agreement: Reliability, classification accuracy, and classification consistency. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of Diagnostic Classification Models* (pp. 359–377). Springer International Publishing. https://doi.org/10.1007/978-3-030-05584-4_17

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 19–38). Information Age Publishing, Inc.

Stan Development Team. (2022). RStan: The R interface to Stan [R package version 2.21.5]. https://mc-stan.org/

Stefan, A. M., Evans, N. J., & Wagenmakers, E.-J. (2020). Practical challenges and methodological flexibility in prior elicitation. *Psychological Methods*, *27*(2), 177–197. https://doi.org/10.1037/met0000354

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370. https://www.jstor.org/stable/1434855

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*(2), 251–275. https://doi.org/10.1007/s00357-013-9129-4

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*(2), 317–339. https://doi.org/10.1007/s11336-013-9362-0

Templin, J., & Henson, R. (2008, March 24–28). *Understanding the impact of skill acquisition: Relating diagnostic assessments to measurable outcomes* [Paper presentation]. American Educational Research Association Annual Meeting, New York, NY.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments*. University of Minnesota, National Center on Educational Outcomes. Minneapolis, MN. https://nceo.umn.edu/docs/OnlinePubs/Synth44.pdf

Thompson, W. J. (2019). *Bayesian psychometrics for diagnostic assessments: A proof of concept* (Research Report No. 19-01). University of Kansas, Accessible Teaching, Learning, and Assessment Systems. https://doi.org/10.35542/osf.io/jzqs8

Thompson, W. J. (2020). *Reliability for the Dynamic Learning Maps assessments: A comparison of methods* (Technical Report No. 20-03). University of Kansas; Accessible Teaching, Learning, and Assessment Systems. https://dynamiclearningmaps.org/sites/default/files/documents/publication/Reliability_Comparison.pdf

Thompson, W. J., Clark, A. K., & Nash, B. (2019). Measuring the reliability of diagnostic mastery classifications at multiple levels of reporting. *Applied Measurement in Education*, *32*(4), 298–309. https://doi.org/10.1080/08957347.2019.1660345

Thompson, W. J., & Nash, B. (2019, April 4–8). Empirical methods for evaluating maps: Illustrations and results. In M. Karvonen (Chair), *Beyond learning progressions: Maps as assessment architecture* [Symposium]. National Council on Measurement in Education Annual Meeting, Toronto, Canada. https://dynamiclearningmaps.org/sites/default/files/documents/presentations/Thompson_Nash_Empirical_evaluation_of_learning_maps.pdf

Thompson, W. J., & Nash, B. (2022). A diagnostic framework for the empirical evaluation of learning maps. *Frontiers in Education*, *6*, 714736. https://doi.org/10.3389/feduc.2021.714736

Timberlake, M. T. (2014). Weighing costs and benefits: Teacher interpretation and implementation of access to the general education curriculum. *Research and Practice for Persons with Severe Disabilities*, *39*(2), 83–99. https://doi.org/10.1177/1540796914544547

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\widehat{\mathrm{R}}$ for assessing convergence of MCMC. *Bayesian Analysis*, *16*(2), 667–718. https://doi.org/10.1214/20-BA1221

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., & Gabry, J. (2022, August 4). *Pareto smoothed importance sampling*. arXiv. https://doi.org/https://doi.org/10.48550/arXiv.1507.02646

Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement*, *52*(4), 457–476. https://doi.org/10.1111/jedm.12096

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594. http://www.jmlr.org/papers/v11/watanabe10a.html

Xu, G. (2019). Identifiability and cognitive diagnosis models. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of Diagnostic Classification Models* (pp. 333–357). Springer International Publishing. https://doi.org/10.1007/978-3-030-05584-4_16

Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika*, *81*(3), 625–649. https://doi.org/10.1007/s11336-015-9471-z

Yorkston, K., Dowden, P., Honsinger, M., Marriner, N., & Smith, K. (1988). A comparison of standard and user vocabulary lists. *Augmentative and Alternative Communication*, *4*(4), 189–210. https://doi.org/10.1080/07434618812331274807

Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF* (Working Paper). University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science. Prince George, Canada.

A. Supplemental Information About the Overview

A.1. List of Terms

Table A.1

Glossary of Common Terms

Term	Definition
Accessible Teaching, Learning, and Assessment Systems	Part of the University of Kansas's Achievement and Assessment Institute. ATLAS develops and administers educational assessment programs including the Dynamic Learning Maps Alternate Assessment System.
Achievement and Assessment Institute	The unit within the University of Kansas which is comprised of several centers including Accessible Teaching, Learning, and Assessment Systems and Agile Technology Solutions, which develop and deliver Dynamic Learning Maps assessments.
Agile Technology Solutions	The organization that develops and maintains the Kite Suite and provides Dynamic Learning Maps Service Desk support to educators in the field.
Alternate achievement standards	Alternate or extended content standards that link to college and career readiness standards in general education that reflect the highest academic expectations for students with the most significant cognitive disabilities.
Alignment	The relationships between the content structures in the Dynamic Learning Maps assessment system and assessment items. The content of assessment items measure the student's knowledge, skills, and understandings reflected in the content standards which they are intended to measure.
Accessible pathway	A path of nodes in the learning map on the way to a learning target that a student can travel in order to demonstrate a type knowledge or skill, regardless of physical or sensory disability.
Answer options	Response choices in assessment items.
Claim	A broad statement about what the Dynamic Learning Maps Consortium expects students to learn and to be able to demonstrate within each content area. Each claim is subdivided into two or more conceptual areas.

Table A.1

Glossary of Common Terms (continued)

Term	Definition
College and career readiness standards	A set of grade-level academic content standards for grades K-12 that Dynamic Learning Maps Essential Elements link to at a reduced depth, breadth, and complexity.
Computer delivered testlet	A test designed to emphasize student interaction with the content of the testlet, regardless of the means of physical access to the computer. The contents of the testlets are presented directly to the student.
Conceptual area	A region within the learning map that contains nodes directly related to Essential Elements and nodes that represent concepts and skills that support the learning of the Essential Elements. Conceptual areas are comprised of clusters of connected concepts and skills and serve as models of how students may acquire and organize their content knowledge. Conceptual areas are considered subparts of the overall claims.
Connection	Directional relationship between two nodes. A connection is illustrated with arrows in the map. Synonym for pathway.
DLM learning maps	A learning map consisting of numerous nodes and connections representing the multiple learning progressions that cover the development of the cognitive and content-area skills from birth to high-school graduation. DLM learning maps also provide access to multiple and alternate routes to achieving the learning targets, making it more inclusive for learners with various disabilities.
Diagnostic classification model	Psychometric model used to estimate a student's mastery status for each assessed linkage level.
Dynamic Learning Maps Alternate Assessment	An assessment system designed to be accessible by students with the most significant cognitive disabilities, including those who also have hearing or visual disabilities, and/or neuromuscular, orthopedic, or other motor disabilities. The assessment includes computer-based assessments and a web-based dashboard for educators to manage student information. The assessment system also includes professional development to support instruction aligned to the Essential Elements and promote student progress in a learning maps-based environment.

Table A.1

Glossary of Common Terms (continued)

Term	Definition
DLM governance board	The goveranance board is made up of 2 representatives from each member state and advises on the administration, maintenance, and enhancement of the DLM System.
Educator Portal	An administrative application in the Kite Suite where educators manage student data, complete required test administration training, assign instructionally embedded assessments, retrieve resources needed for each assigned testlet, and retrieve reports.
Engagement activity	An activity that precedes a testlet that describes a scenario, taps prior knowledge or experience, and/or introduces the concept to be addressed. In English language arts, the text being read often serves as the engagement activity. In mathematics, the engagement activity provides context for the items.
Essential Element	Specific statements of knowledge and skills linked to the grade-level expectations identified in college and career readiness standards. Essential Elements build a bridge from the content in the grade-level standards to academic expectations for students with the most significant cognitive disabilities.
Essential Element Concept Map	A graphic organizer using principles of Evidence Centered Design and Universal Design for Learning to define content specifications for assessment. The EECM uses principles of evidence-centered design and provides information about evidence of Essential Element mastery, key vocabulary and concepts, and potential non-cognitive and accessibility barriers when assessing the target behaviors.
Linkage level	A small section of the learning map that contains one or more nodes that represent critical concepts or skills related to each Essential Element. Linkage levels are always related directly to grade level Essential Elements but extend back to foundational skills at the Initial Precursor level. There are five linkage levels for each Essential Element: Initial Precursor, Distal Precursor, Proximal Precursor, Target, and Successor. The nodes at the Target level are most closely related to the expectation in the Essential Element.

Table A.1

Glossary of Common Terms (continued)

Term	Definition
First Contact survey	A survey used to collect background information about students who are eligible for Dynamic Learning Maps assessments. The survey goes beyond basic demographic information and includes questions on topics such as communication, assistive technology devices, motor and sensory impairments, academic performance. Some questions from the First Contact survey are used to determine a student's entry point, or initialization, into the assessment.
Foundational skills	Common set nodes in the learning map that represent basic skills that precede academic knowledge, including attention, self-regulation, and organization, and provide an understructure for the academic skills.
Fungible	Exchangeable, able to be replaced by another identical item. In the Dynamic Learning Maps assessments, all items within an Essential Element and linkage level are assumed to be fungible, or exchangeable.
General Research File	The data file provided to states at the end of each year. It contains student demographic information and assessment results.
Initialization	The process by which existing information about a student is used to determine the point in the map where the student enters the assessment for the first time. Uses information from the First Contact survey.
Instructional plan	A plan, created through the Educator Portal Instruction and Assessment Planner, which includes a choice of Essential Element and linkage level and leads to assignment of an instructionally embedded assessment.
Instructionally embedded assessment	An assessment that occurs after instruction throughout the year so that testing informs teaching and benefits students' learning.
Instruction and Assessment Planner	An interface in Educator Portal which allows a test administrator to select an Essential Element and linkage level for a student with the goal of providing instruction and instructionally embedded assessment.
Kite System	The platform which includes Kite Student Portal and Kite Educator Portal.

Table A.1

Glossary of Common Terms (continued)

Term	Definition
Kite Student Portal	An online testing interface for students. The Kite Student Portal is available for use on PCs, Macs, Chromebooks, and iPads.
Learning Profile	Part of the individual student score report provided at the end of the year. Provides information about student mastery of linkage levels for every Essential Element assessed.
Mini-maps	A small section of the learning map representing a single Essential Element. Mini-maps consist of the nodes around which a set of testlets is developed.
Node	A node that specifies individual skills and understandings that were drawn from the research in English language arts and mathematics.
Performance Profile	Part of the individual student score report provided at the end of the year. Provides information about students' overall performance in the subject and for conceptual areas in each subject.
Personal Needs and Preferences profile (PNP)	Student-specific information that tells the Dynamic Learning Maps test delivery system what the needs are for individual users. The PNP includes information the system needs to make the student's user interface compatible with his or her accessibility needs. The PNP profile includes information about display enhancements, language and braille, assistive technology, and audio and environment supports. Educators who know the student provide the information in the profile.
Proposition	Propositions are claims that relate directly to the ultimate program goals and specific score purposes of the assessment system, providing the framework within which validity evidence can be judged.
Students with the most significant cognitive disabilities	A student who falls within one of the existing categories of disability under the Individuals with Disabilities Education Act (autism, deaf-blindness, hearing impairment, mental retardation, orthopedic impairment, deafness, emotional disturbance, multiple disability, traumatic brain injury, visual impairment, learning disability, speech and language impairment, other health impaired) whose cognitive impairments may prevent them from attaining grade-level achievement standards, even with supports.

Table A.1

Glossary of Common Terms (continued)

Term	Definition
Educator-administered testlet	A test designed to be administered directly by the test administrator outside of the Kite Suite. Student responses are entered into the testlet by the test administrator.
Testlet	A set of 3-9 items and an engagement activity. Combining multiple items and beginning with an engagement activity increases the instructional relevance of the assessment, and provides a better estimate of the students' knowledge, skills and abilities than can be achieved by a single test item.
Testlet Information Page (TIP)	A secure PDF document unique to each testlet that provides specific information to guide the test administrator in preparing for and administering the testlet.
Theory of action	Series of claims covering areas of design, administration, scoring, and intended long-term outcomes of the DLM assessment program.
Required training	Training modules are available in both self-directed and facilitated formats. Modules cover topics such as instruction, the use of assessment results, and required skills for test administrators.
Virtual community of practice	An online community for educators of students with the most significant cognitive disabilities. The community is self-moderated with oversight from faculty and staff at the Center for Literacy and Disability Studies. The community provides instructional resources and supports as well as discussion forums and groups.

A.2. List of Acronyms

Table A.2

Glossary of Common Abbreviations

Acronym	Definition
AERA	American Educational Research Association
ATLAS	Accessible Teaching, Learning, and Assessment Systems
AWS	Amazon Web Services

Table A.2

Glossary of Common Abbreviations (continued)

Acronym	Definition
BVI	Blind or visual impairment
CAST	Center for Applied Special Technology
CCSS	Common Core State Standards
CLDS	Center for Literacy and Disability Studies
CPD	Cognitive process dimension
DCM	Diagnostic classification model
DIF	Differential item functioning
DLM	Dynamic Learning Maps
ECD	Evidence-centered design
EE	Essential Element
EECM	Essential Element concept map
ELA	English language arts
EL	English learner
FERPA	Family Educational Rights and Privacy Act
GRF	General research file
HDCM	Hierarchical diagnostic classification model
ICC	Intraclass correlation
IEP	Individualized education program
Kite	Kansas Interactive Testing Engine
LCDM	Loglinear cognitive diagnostic model
LCS	Local caching server
OSX	Also macOS; the Apple operating system
PII	Personally-identifiable information

Table A.2

Glossary of Common Abbreviations (continued)

Acronym	Definition
PLD	Performance level descriptor
PNP	Personal Needs and Preferences
PSIS-LOO	Leave-one-out cross validation with Pareto smoothed importance sampling
SEA	State education agency
SWSCD	Students with significant cognitive disabilities
TAC	Technical Advisory Committee
TIP	Testlet information page
TTS	Text to speech
UDL	Universal design for learning
WAIC	Widely applicable information criterion

B. Supplemental Information About the Content Structures

B.1. Assessment Blueprints

Below are the grade- and subject-specific assessment blueprints for English language arts (ELA) and mathematics. All blueprints for ELA[74] and mathematics[75] are available on the DLM website.

B.1.1. Blueprints for English Language Arts

Grade 3: Essential Elements Assessed

Conceptual Area	Essential Element	Description*
ELA.C1.1		
	ELA.EE.RL.3.1	Answer who and what questions to demonstrate understanding of details in a text.
	ELA.EE.RL.3.3	Identify the feelings of characters in a story.
	ELA.EE.RI.3.2	Identify details in a text.
	ELA.EE.RI.3.3	Order two events from a text as "first" and "next."
ELA.C1.2		
	ELA.EE.RL.3.4	Determine words and phrases that complete literal sentences in a text.
	ELA.EE.RI.3.8	Identify two related points the author makes in an informational text.
	ELA.EE.L.3.5.c	Identify words that describe personal emotional states.
ELA.C1.3		
	ELA.EE.RL.3.9	Identify common elements in two stories in a series.
ELA.C2.1		
	ELA.EE.W.3.2.a	Select a topic and write about it including one fact or detail.
	ELA.EE.W.3.4	With guidance and support, produce writing that expresses more than one idea.

*Note: RL = Reading Literature, RI = Reading Informational, W = Writing, L = Language

[74] https://dynamiclearningmaps.org/sites/default/files/documents/Manuals_Blueprints/DLM_YE_ELA_Blueprint.pdf
[75] https://dynamiclearningmaps.org/sites/default/files/documents/Manuals_Blueprints/DLM_YE_Math_Blueprint.pdf

Grade 4: Essential Elements Assessed

Conceptual Area	Essential Element	Description*
ELA.C1.1		
	ELA.EE.RL.4.5	Identify elements that are characteristic of stories.
	ELA.EE.RI.4.1	Identify explicit details in an informational text.
	ELA.EE.RI.4.5	Identify elements that are characteristic of informational texts.
ELA.C1.2		
	ELA.EE.RL.4.2	Identify the theme or central idea of a familiar story, drama or poem.
	ELA.EE.RL.4.6	Identify the narrator of a story.
	ELA.EE.RI.4.4	Determine meaning of words in text.
	ELA.EE.L.4.5.c	Demonstrate understanding of opposites.
ELA.C1.3		
	ELA.EE.RI.4.9	Compare details presented in two texts on the same topic.
ELA.C2.1		
	ELA.EE.L.4.2.a	Capitalize the first word in a sentence.
	ELA.EE.L.4.2.d	Spell words phonetically, drawing on knowledge of letter-sound relationships, and/or common spelling patterns.
	ELA.EE.W.4.2.b	List words, facts, or details related to the topic.

*Note: RL = Reading Literature, RI = Reading Informational, W = Writing, L = Language

Grade 5: Essential Elements Assessed

Conceptual Area	Essential Element	Description*
ELA.C1.1		
	ELA.EE.RL.5.1	Identify words in the text to answer a question about explicit information.
ELA.C1.2		
	ELA.EE.RL.5.6	Determine the point of view of the narrator.
	ELA.EE.RI.5.2	Identify the main idea of a text when it is not explicitly stated.
	ELA.EE.RI.5.4	Determine the meanings of domain-specific words and phrases.
	ELA.EE.RI.5.8	Identify the relationship between a specific point and supporting reasons in an informational text.
	ELA.EE.L.5.4.a	Use sentence level context to determine which word is missing from a content area text.
ELA.C1.3		
	ELA.EE.RL.5.9	Compare stories, myths, or texts with similar topics or themes.
	ELA.EE.RI.5.3	Compare two individuals, events, or ideas in a text.
ELA.C2.1		
	ELA.EE.W.5.2.b	Provide facts, details, or other information related to the topic.
	ELA.EE.W.5.2.a	Introduce a topic and write to convey information about it including visual, tactual, or multimedia information as appropriate.

*Note: RL = Reading Literature, RI = Reading Informational, W = Writing, L = Language

Grade 6: Essential Elements Assessed

Conceptual Area	Essential Element	Description*
ELA.C1.1		
	ELA.EE.RI.6.5	Determine how the title fits the structure of the text.
ELA.C1.2		
	ELA.EE.RL.6.2	Identify details in a text that are related to the theme or central idea.
	ELA.EE.RL.6.4	Determine how word choice changes the meaning in a text.
	ELA.EE.RI.6.1	Analyze a text to determine what it says explicitly as well as what inferences should be drawn.
	ELA.EE.RI.6.6	Identify words or phrases in the text that describe or show the author's point of view.
	ELA.EE.RI.6.8	Distinguish claims in a text supported by reason.
	ELA.EE.L.6.5.b	Demonstrate understanding of words by identifying other words with similar and different meanings.
ELA.C1.3		
	ELA.EE.RL.6.3	Can identify how a character responds to a challenge in a story.
ELA.C2.1		
	ELA.EE.L.6.2.b	Spell untaught words phonetically, drawing on letter-sound relationships and common spelling patterns.
	ELA.EE.W.6.2.a	Introduce a topic and write to convey ideas and information about it including visual, tactual, or multimedia information as appropriate.
	ELA.EE.W.6.2.b	Provide facts, details, or other information related to the topic.

*Note: RL = Reading Literature, RI = Reading Informational, W = Writing, L = Language

Dynamic Learning Maps English Language Arts Year-End Assessment Model Blueprint Page 6 of 10

Grade 7: Essential Elements Assessed

Conceptual Area	Essential Element	Description*
ELA.C1.1		
	ELA.EE.RI.7.5	Determine how a fact, step, or event fits into the overall structure of the text.
ELA.C1.2		
	ELA.EE.RL.7.1	Analyze text to identify where information is explicitly stated and where inferences must be drawn.
	ELA.EE.RL.7.4	Determine the meaning of simple idioms and figures of speech as they are used in a text.
	ELA.EE.RI.7.2	Determine two or more central ideas in a text.
	ELA.EE.RI.7.8	Determine how a claim or reason fits into the overall structure of an informational text.
ELA.C1.3		
	ELA.EE.RL.7.5	Compare the structure of two or more texts (e.g., stories, poems, or dramas).
	ELA.EE.RI.7.3	Determine how two individuals, events or ideas in a text are related.
	ELA.EE.RI.7.9	Compare and contrast how different texts on the same topic present the details.
ELA.C2.1		
	ELA.EE.L.7.2.a	Use end punctuation when writing a sentence or question.
	ELA.EE.L.7.2.b	Spell words phonetically, drawing on knowledge of letter-sound relationships and/or common spelling patterns.
	ELA.EE.W.7.2.a	Introduce a topic and write to convey ideas and information about it including visual, tactual, or multimedia information as appropriate.
	ELA.EE.W.7.2.b	Provide facts, details, or other information related to the topic.
	ELA.EE.W.7.2.d	Select domain-specific vocabulary to use in writing about the topic.

*Note: RL = Reading Literature, RI = Reading Informational, W = Writing, L = Language

Grade 8: Essential Elements Assessed

Conceptual Area	Essential Element	Description*
ELA.C1.2		
	ELA.EE.RL.8.1	Cite text to support inferences from stories and poems.
	ELA.EE.RL.8.2	Recount an event related to the theme or central idea, including details about character and setting.
	ELA.EE.RI.8.2	Provide a summary of a familiar informational text.
	ELA.EE.RI.8.6	Determine an author's purpose or point of view and identify examples from text that describe or support it.
	ELA.EE.RI.8.8	Determine the argument made by an author in an informational text.
	ELA.EE.L.8.5.a	Demonstrate understanding of the use of multiple meaning words.
ELA.C1.3		
	ELA.EE.RL.8.3	Identify which incidents in a story or drama lead to subsequent action.
	ELA.EE.RL.8.5	Compare and contrast the structure of two or more texts.
ELA.C2.1		
	ELA.EE.W.8.2.b	Write one or more facts or details related to the topic.
	ELA.EE.W.8.2.c	Write complete thoughts as appropriate.
	ELA.EE.W.8.2.d	Use domain specific vocabulary related to the topic.
	ELA.EE.W.8.2.f	Provide a closing.
	ELA.EE.W.8.2.a	Introduce a topic clearly and write to convey ideas and information about it including visual, tactual, or multimedia information as appropriate.

*Note: RL = Reading Literature, RI = Reading Informational, W = Writing, L = Language

Grades 9-10: Essential Elements Assessed[1]

Conceptual Area	Essential Element	Description*
ELA.C1.2		
	ELA.EE.RL.9-10.4	Determine the meaning of words and phrases as they are used in a text, including idioms, analogies, and figures of speech.
	ELA.EE.RI.9-10.1	Determine which citations demonstrate what the text says explicitly as well as inferentially.
	ELA.EE.RI.9-10.2	Determine the central idea of the text and select details to support it.
	ELA.EE.RI.9-10.5	Locate sentences that support an author's central idea or claim.
	ELA.EE.RI.9-10.8	Determine how the specific claims support the argument made in an informational text.
ELA.C1.3		
	ELA.EE.RL.9-10.3	Determine how characters change or develop over the course of a text.
	ELA.EE.RL.9-10.5	Identify where a text deviates from a chronological presentation of events.
	ELA.EE.RI.9-10.3	Determine logical connections between individuals, ideas, or events in a text.
ELA.C2.1		
	ELA.EE.L.9-10.2.c	Spell most single-syllable words correctly and apply knowledge of word chunks in spelling longer words.
	ELA.EE.W.9-10.2.c	Use complete, simple sentences as appropriate.
	ELA.EE.W.9-10.2.d	Use domain specific vocabulary when writing claims related to a topic of study or text.
	ELA.EE.W.9-10.2.f	Provide a closing or concluding statement.
ELA.C2.2		
	ELA.EE.W.9-10.2.a	Introduce a topic clearly and use a clear organization to write about it including visual, tactual, or multimedia information as appropriate.
	ELA.EE.W.9-10.2.b	Develop the topic with facts or details.

*Note: RL = Reading Literature, RI = Reading Informational, W = Writing, L = Language

[1] The high school blueprint provides coverage options for students in grades 9, 10, and 11 to support the various testing requirements in different states in the consortium. Each state sets its own policy for which high school grade(s) are appropriate for DLM assessments.

Grades 11-12: Essential Elements Assessed[2]

Conceptual Area	Essential Element	Description*
ELA.C1.2		
	ELA.EE.RL.11-12.1	Analyze a text to determine its meaning and cite textual evidence to support explicit and implicit understandings.
	ELA.EE.RL.11-12.2	Recount the main events of the text which are related to the theme or central idea.
	ELA.EE.RI.11-12.4	Determine how words or phrases in a text, including words with multiple meanings and figurative language, impacts the meaning of the text.
	ELA.EE.RI.11-12.8	Determine whether the claims and reasoning enhance the author's argument in an informational text.
	ELA.EE.RI.11-12.5	Determine whether the structure of a text enhances an author's claim.
ELA.C1.3		
	ELA.EE.RL.11-12.3	Determine how characters, the setting or events change over the course of the story or drama.
	ELA.EE.RL.11-12.5	Determine how the author's choice of where to end the story contributes to the meaning.
	ELA.EE.RI.11-12.9	Compare and contrast arguments made by two different texts on the same topic.
ELA.C2.1		
	ELA.EE.W.11-12.2.c	Use complete, simple sentences, as well as compound and other complex sentences as appropriate.
	ELA.EE.W.11-12.2.d	Use domain specific vocabulary when writing claims related to a topic of study or text.
	ELA.EE.W.11-12.2.f	Provide a closing or concluding statement.
	ELA.EE.L.11-12.2.b	Spell most single-syllable words correctly and apply knowledge of word chunks in spelling longer words.
ELA.C2.2		
	ELA.EE.W.11-12.2.a	Introduce a topic clearly and write an informative or explanatory text that conveys ideas, concepts, and information including visual, tactual, or multimedia information as appropriate.
	ELA.EE.W.11-12.2.b	Develop the topic with relevant facts, details, or quotes.

*Note: RL = Reading Literature, RI = Reading Informational, W = Writing, L = Language

[2] The high school blueprint provides coverage options for students in grades 9, 10, and 11 to support the various testing requirements in different states in the consortium. Each state sets its own policy for which high school grade(s) are appropriate for DLM assessments.

B.1.2. Blueprints for Mathematics

Grade 3: Essential Elements Assessed

Conceptual Area	Essential Element	Description
M.C1.1		
	M.EE.3.NBT.2	Demonstrate understanding of place value to tens.
	M.EE.3.NBT.3	Count by tens using models such as objects, base ten blocks, or money.
M.C1.3		
	M.EE.3.OA.4	Solve addition and subtraction problems when result is unknown, limited to operands and results within 20.
M.C2.2		
	M.EE.3.G.2	Recognize that shapes can be partitioned into equal areas.
M.C3.1		
	M.EE.3.MD.4	Measure length of objects using standard tools, such as rulers, yardsticks, and meter sticks.
M.C3.2		
	M.EE.3.MD.3	Use picture or bar graph data to answer questions about data.
M.C4.1		
	M.EE.3.OA.1-2	Use repeated addition to find the total number of objects and determine the sum.
M.C4.2		
	M.EE.3.OA.9	Identify arithmetic patterns.

Grade 4: Essential Elements Assessed

Conceptual Area	Essential Element	Description
M.C1.1		
	M.EE.4.NF.1-2	Identify models of one half (1/2) and one fourth (1/4).
M.C1.3		
	M.EE.4.NBT.4	Add and subtract two-digit whole numbers.
M.C2.1		
	M.EE.4.MD.6	Identify angles as larger and smaller.
M.C2.2		
	M.EE.4.MD.3	Determine the area of a square or rectangle by counting units of measure (unit squares).
M.C3.1		
	M.EE.4.MD.2.a	Tell time using a digital clock. Tell time to the nearest hour using an analog clock.
	M.EE.4.MD.2.b	Measure mass or volume using standard tools.
	M.EE.4.MD.2.d	Identify coins (penny, nickel, dime, quarter) and their values.
M.C4.2		
	M.EE.4.OA.5	Use repeating patterns to make predictions.

Grade 5: Essential Elements Assessed

Conceptual Area	Essential Element	Description
M.C1.1		
	M.EE.5.NF.2	Identify models of thirds (1/3, 2/3, 3/3) and tenths (1/10, 2/10, 3/10, 4/10, 5/10, 6/10, 7/10, 8/10, 9/10, 10/10).
M.C1.2		
	M.EE.5.NBT.3	Compare whole numbers up to 100 using symbols (<, >, =).
	M.EE.5.NBT.4	Round two-digit whole numbers to the nearest 10 from 0—90.
M.C1.3		
	M.EE.5.NBT.5	Multiply whole numbers up to 5x5.
M.C2.1		
	M.EE.5.G.1-4	Sort two-dimensional figures and identify the attributes (angles, number of sides, corners, color) they have in common.
	M.EE.5.MD.3	Identify common three-dimensional shapes.
M.C2.2		
	M.EE.5.MD.4-5	Determine the volume of a rectangular prism by counting units of measure (unit cubes).
M.C3.2		
	M.EE.5.MD.2	Represent and interpret data on a picture graph, line plot, or bar graph.

Grade 6: Essential Elements Assessed

Conceptual Area	Essential Element	Description
M.C1.2		
	M.EE.6.NS.1	Compare the relationships between two unit fractions.
	M.EE.6.NS.5-8	Understand that positive and negative numbers are used together to describe quantities having opposite directions or values (e.g., temperature above/below zero).
M.C2.2		
	M.EE.6.G.1	Solve real-world and mathematical problems about area using unit squares.
	M.EE.6.G.2	Solve real-world and mathematical problems about volume using unit cubes.
M.C3.2		
	M.EE.6.SP.5	Summarize data distributions shown in graphs or tables.
M.C4.1		
	M.EE.6.EE.1-2	Identify equivalent number sentences.
	M.EE.6.EE.3	Apply the properties of addition to identify equivalent numerical expressions.

Grade 7: Essential Elements Assessed

Conceptual Area	Essential Element	Description
M.C1.1		
	M.EE.7.RP.1-3	Use a ratio to model or describe a relationship.
M.C1.3		
	M.EE.7.NS.1	Add fractions with like denominators (halves, thirds, fourths, and tenths) with sums less than or equal to one.
	M.EE.7.NS.2.a	Solve multiplication problems with products to 100.
	M.EE.7.NS.2.b	Solve division problems with divisors up to five and also with a divisor of 10 without remainders.
M.C2.1		
	M.EE.7.G.2	Recognize geometric shapes with given conditions.
M.C2.2		
	M.EE.7.G.4	Determine the perimeter of a rectangle by adding the measures of the sides.
M.C4.1		
	M.EE.7.EE.1	Use the properties of operations as strategies to demonstrate that expressions are equivalent.

Grade 8: Essential Elements Assessed

Conceptual Area	Essential Element	Description
M.C1.1		
	M.EE.8.NS.2.a	Express a fraction with a denominator of 100 as a decimal.
M.C1.3		
	M.EE.8.NS.1	Subtract fractions with like denominators (halves, thirds, fourths, and tenths) with minuends less than or equal to one.
M.C2.1		
	M.EE.8.G.5	Compare any angle to a right angle, and describe the angle as greater than, less than, or congruent to a right angle.
M.C2.2		
	M.EE.8.G.9	Use the formulas for perimeter, area, and volume to solve real-world and mathematical problems (limited to perimeter and area of rectangles and volume of rectangular prisms).
M.C3.2		
	M.EE.8.SP.4	Construct a graph or table from given categorical data, and compare data categorized in the graph or table.
M.C4.1		
	M.EE.8.EE.7	Solve simple algebraic equations with one variable using addition and subtraction.
M.C4.2		
	M.EE.8.EE.2	Identify a geometric sequence of whole numbers with a whole number common ratio.
	M.EE.8.F.1-3	Given a function table containing at least 2 complete ordered pairs, identify a missing number that completes another ordered pair (limited to linear functions).

Grade 9: Essential Elements Assessed[2]

Conceptual Area	Essential Element	Description
M.C1.3		
	M.EE.N.CN.2.a	Use the commutative, associative, and distributive properties to add, subtract, and multiply whole numbers.
	M.EE.N.CN.2.b	Solve real-world problems involving addition and subtraction of decimals, using models when needed.
	M.EE.N.CN.2.c	Solve real-world problems involving multiplication of decimals and whole numbers, using models when needed.
M.C2.1		
	M.EE.G.CO.1	Know the attributes of perpendicular lines, parallel lines, and line segments; angles; and circles.
	M.EE.G.MG.1-3	Use properties of geometric shapes to describe real-life objects.
M.C4.1		
	M.EE.A.SSE.1	Identify an algebraic expression involving one arithmetic operation to represent a real-world problem.
	M.EE.A.SSE.3	Solve simple algebraic equations with one variable using multiplication and division.

[2] The high school blueprint provides coverage options for students in grades 9, 10, and 11 to support the various testing requirements in different states in the consortium. Each state sets its own policy for which high school grade(s) are appropriate for DLM assessments.

Grade 10: Essential Elements Assessed[3]

Conceptual Area	Essential Element	Description
M.C1.3		
	M.EE.S.CP.1-5	Identify when events are independent or dependent.
M.C2.1		
	M.EE.G.CO.4-5	Given a geometric figure and a rotation, reflection, or translation of that figure, identify the components of the two figures that are congruent.
M.C3.1		
	M.EE.N.Q.1-3	Express quantities to the appropriate precision of measurement.
M.C3.2		
	M.EE.S.ID.1-2	Given data, construct a simple graph (line, pie, bar, or picture) or table, and interpret the data.
	M.EE.S.ID.4	Calculate the mean of a given data set (limit the number of data points to fewer than five).
M.C4.1		
	M.EE.A.CED.2-4	Solve one-step inequalities.
M.C4.2		
	M.EE.A.REI.10-12	Interpret the meaning of a point on the graph of a line. *For example, on a graph of pizza purchases, trace the graph to a point and tell the number of pizzas purchased and the total cost of the pizzas.*
	M.EE.F.BF.1	Select the appropriate graphical representation (first quadrant) given a situation involving constant rate of change.

[3] The high school blueprint provides coverage options for students in grades 9, 10, and 11 to support the various testing requirements in different states in the consortium. Each state sets its own policy for which high school grade(s) are appropriate for DLM assessments.

Grade 11: Essential Elements Assessed[4]

Conceptual Area	Essential Element	Description
M.C1.3		
	M.EE.N.RN.1	Determine the value of a quantity that is squared or cubed.
	M.EE.S.IC.1-2	Determine the likelihood of an event occurring when the outcomes are equally likely to occur.
M.C2.1		
	M.EE.G.CO.6-8	Identify corresponding congruent and similar parts of shapes.
M.C3.2		
	M.EE.S.ID.3	Interpret general trends on a graph or chart.
M.C4.2		
	M.EE.F.BF.2	Determine an arithmetic sequence with whole numbers when provided a recursive rule.
	M.EE.F.IF.4-6	Construct graphs that represent linear functions with different rates of change and interpret which is faster/slower, higher/lower, etc.

[4] The high school blueprint provides coverage options for students in grades 9, 10, and 11 to support the various testing requirements in different states in the consortium. Each state sets its own policy for which high school grade(s) are appropriate for DLM assessments.

C. Supplemental Information About Assessment Delivery

C.1. First Contact Survey Items Used for Determining Complexity Bands

A subset of questions from the First Contact survey are used to determine the expressive communication, English language arts, mathematics, and writing complexity bands. The questions from the survey used to determine each band are shown in the following sections.

C.1.1. Expressive Communication

Does the student use speech to meet expressive communication needs?

- Yes
- No

Choose the highest statement that describes the student's expressive communication with speech.

- Regularly combines 3 or more spoken words according to grammatical rules to accomplish a variety of communicative purposes (e.g., sharing complex information, asking/answering longer questions, giving directions to another person)
- Usually uses 2 spoken words at a time to meet a variety of more complex communicative purposes (e.g., obtaining things including absent objects, social expressions beyond greetings, sharing information, directing another person's attention, asking/answering questions, and commenting)
- Usually uses only 1 spoken word at a time to meet a limited number of simple communicative purposes (e.g., refusing/rejecting things, making choices, requesting attention, greeting, and labeling)

Does the student use sign language in addition to or in place of speech to meet expressive communication needs?

- Yes
- No

Choose the highest statement that describes the student's expressive communication with sign language.

- Regularly combines 3 or more signed words according to grammatical rules to accomplish a variety of communicative purposes (e.g., sharing complex information, asking/answering longer questions, giving directions to another person)
- Usually uses 2 signed words at a time to meet a variety of more complex communicative purposes (e.g., obtaining things including absent objects, social expressions beyond greetings, sharing information, directing another person's attention, asking/answering brief questions, and commenting)
- Usually uses only 1 signed word at a time to meet a limited number of simple communicative purposes (e.g., refusing/rejecting things, making choices, requesting attention, greeting, and labeling)

Does the student use augmentative or alternative communication in addition to or in place of speech or sign language to meet expressive communication needs?

- Yes
- No

C.1.2. English Language Arts

Student's approximate instructional level of reading text with comprehension (print or Braille): Mark the highest one that applies.

- Above third grade level
- Above second grade level to third grade level
- Above first grade level to second grade level
- Primer to first grade level
- Reads only a few words or up to pre-primer level
- Does not read any words when presented in print or Braille (not including environmental signs or logos)

Reading skills: MARK EACH ONE to show the approximate percent of time that the student uses each skill—0% (student does not exhibit this skill), none to 20% of the time, 21% to 50% of the time, 51% to 80% of the time, More than 80% of the time.

A) Recognizes single symbols presented visually or tactually (e.g., letters, numerals, environmental signs such as restroom symbols, logos, trademarks, or business signs such as fast food restaurants)

C.1.3. Mathematics

Math skills: MARK EACH ONE to show the approximate percent of time that the student uses each skill—0% (student does not exhibit this skill), none to 20% of the time, 21% to 50% of the time, 51% to 80% of the time, More than 80% of the time.

C) Sorts objects by common properties (e.g., color, size, shape)
D) Adds or subtracts by joining or separating groups of objects
E) Forms groups of objects for multiplication or division
F) Multiplies and/or divides using numerals

C.1.4. Writing

Writing skills: Indicate the highest level that describes the student's writing skills. Choose the highest level that the student has demonstrated even once during instruction, not the highest level demonstrated consistently. Writing includes any method the student uses to write using any writing tool that includes access to all 26 letters of the alphabet. Examples of these tools include paper and pencil, traditional keyboards, alternate keyboards and eye-gaze displays of letters.

- Writes paragraph length text without copying using spelling (with or without word prediction)
- Writes sentences or complete ideas without copying using spelling (with or without word prediction)
- Writes words or simple phrases without copying using spelling (with or without word prediction)
- Writes words using letters to accurately reflect some of the sounds
- Writes using word banks or picture symbols
- Writes by copying words or letters

- Scribbles or randomly writes/selects letters or symbols

D. Supplemental Information About Standard Setting

D.1. Technical Advisory Committee Recommendation

On May 11, 2022, the DLM Technical Advisory Committee (TAC) prepared and accepted a resolution endorsing the adequacy, quality of judgments, and extent to which the Year-End standards adjustment process met professional standards and recommended the adoption of the adjusted cut points. The full TAC resolution is presented below.

DLM TAC Recommendation on Standards Adjustment
May 11, 2022

Introduction DLM staff presented a topic guide titled "Year-End Standards Adjustment: Options for Statistical Adjustment," dated April 2022, to the DLM TAC for discussion at its May 11, 2022 DLM TAC meeting. DLM state partners, in collaboration with DLM staff and the DLM TAC, elected to conduct a standards adjustment in lieu of a full-scale standard setting process when the updated Year-End model blueprints were adopted in 2019-20. This blueprint update reduced the number of Essential Elements (EEs) and total linkage levels available to be mastered for each grade and subject. These changes required the adoption of adjusted cut points used to determine performance levels from the total linkage levels mastered in order to account for the changes in the assessment system.

Background In the topic guide, DLM staff presented the original plan for standards adjustment, which was to consist of two steps:

1. Reduce each of the original cut points according to the proportion that the blueprint was reduced in each grade level and subject, and
2. Using spring 2020 impact data, adjust the proportion-reduced cut points +/- 2 total linkage levels mastered to match the 2019 student performance level distributions more closely.

Due to COVID, it was not possible to complete step 2 in spring 2020 or spring 2021. However, a proof-of-concept of the proposed adjustment method was used for the spring 2021 scoring, which adjusted the proportion-reduced cuts an estimated distribution of linkage levels mastered from 2019 (by removing the EEs that were removed from the revised blueprints). However, this method did not account for other changes made to the Year-End test design, such as a switch to single-EE testlets, adding more items per EE, or changes to adaptive routing that resulted from the use of single-EE testlets.

Issue to be Addressed A decision needs to be made on the adjustment method to be used in 2022 and beyond. Two options are under consideration:

1. Adjust the proportionally-reduced cut points +/- 2 points, based on observed impact data from 2022, or
2. Continue to use cut points that were adjusted based on the estimated impact data from 2019 (and which were used for scoring in 2021).

TAC Discussion Following the presentation of a more detailed description of each option and the methods that might be used to implement it, a list of advantages and disadvantages of each approach, and achievement data from 2022 (as of May 6) for each option, there was considerable discussion among members of the TAC.

Although COVID may have affected the sample composition, demographic statistics did not show any unexpected large differences from pre-COVID administrations. Changes in state participation were duly noted.

TAC Recommendation After presentation of the performance-level results using both the cut points based on the observed 2022 data and the cut points based on the estimated 2019 data that were used operationally in 2021, the DLM TAC unanimously adopted a motion that DLM recommend to states the use of Option 1 in reporting DLM results.

The TAC also unanimously recommended DLM provide appropriate caveats to states on the reporting and use of state results from the past several school years.

D.2. Example Grade and Subject Performance Level Descriptors

Below are example grade- and subject-specific performance level descriptors for grade 3 English language arts (ELA) and mathematics. All performance level descriptors for ELA[76] and mathematics[77] are available on the DLM website.

DYNAMIC® LEARNING MAPS

DLM® Performance Level Descriptors–ELA: Grade 3

Year-End Model

Emerging	A student who achieves at the **emerging** performance level typically can identify familiar people, objects, or places; identify feeling words; identify sequences; and identify text structure when reading literature and informational text. The student indicates and identifies familiar people, objects, or places associated with a text by • recognizing similar and different physical characteristics of objects • understanding words for absent objects or people • attending to object characteristics when verbally cued • seeking objects that are absent or are of interest to the student The student identifies feeling words by • identifying personal feelings The student identifies sequences and text structure by • noticing new objects • identifying forward sequences from familiar routines When writing, the student • attends to objects, people, or pictures • makes a choice between two objects
Approaching the Target	A student who achieves at the **approaching the target** performance level typically can identify details and facts, demonstrate an understanding of language, identify feeling words, and identify text structure when reading literature and informational text. The student identifies details and facts by • recognizing similar or different physical characteristics of objects • identifying and understanding relationships between concrete details • answering who or what questions about texts The student demonstrates an understanding of language by • identifying words with the same, similar, or different meanings • identifying real-world uses of words The student identifies and understands feeling words by • identifying the feelings of characters The student identifies text structure by • recognizing pictures from familiar texts When writing, the student • selects a familiar topic • connects two or more words

DLM Performance Level Descriptors: ELA 1 of 30

© 2022 Accessible Teaching, Learning, and Assessment Systems (ATLAS), University of Kansas

76 https://dynamiclearningmaps.org/sites/default/files/documents/ELA_PLDs_YE.pdf
77 https://dynamiclearningmaps.org/sites/default/files/documents/Math_PLDs_YE.pdf

At Target	A student who achieves at the **at target** performance level typically can identify details and facts, demonstrate an understanding of language, identify feelings, and recognize text structure when reading literature and informational text. The student identifies details and facts by • identifying concrete details • answering who, what, when, where, or why questions The student demonstrates an understanding of language by • determining words or phrases that complete literal sentences from texts The student identifies feelings by • identifying personal feelings • identifying character feelings The student identifies text structure by • determining which event comes first in a text • using text features to locate information • identifying common elements in two texts When writing, the student • uses facts and details to write about a topic • expresses more than one idea
Advanced	A student who achieves at the **advanced** performance level typically can recognize details, facts, and supporting points and reasons made by the author; demonstrate an understanding of language; identify feelings; and recognize text structure when reading literature and informational text. The student recognizes details, facts, and supporting points and reasons made by the author by • associating concrete details with events The student demonstrates an understanding of language by • understanding definitions for unambiguous words in a text • identifying words or phrases to complete literal sentences The student identifies feelings by • relating characters' feelings to their actions The student identifies text structure by • identifying the temporal order of information or events in a text • using text features to locate information • comparing elements of two texts When writing, the student • selects an informational topic • includes information from resources to support the topic • expresses complete thoughts

DLM Performance Level Descriptors: ELA 2 of 30

DYNAMIC® LEARNING MAPS

DLM® Performance Level Descriptors–Math: Grade 3

Year-End Model

Emerging	A student who achieves at the **emerging** performance level typically uses attributes or characteristics to identify and sort familiar objects into sets. The student uses attributes or characteristics to identify and sort familiar objects into sets by • recognizing sets of objects and determining if the objects in a set are the same or different based on a given attribute (for example, size, shape, or texture) • understanding the combining and dividing of objects by attending to a particular set of objects and then moving the objects either to create a group or to create separate sets
Approaching the Target	A student who achieves at the **approaching the target** performance level typically represents and solves problems using an understanding of abstract math concepts and symbols. The student represents and solves problems using an understanding of abstract math concepts and symbols by • recognizing how numbers appear in a sequence (for example, 5, 6, 7) and counting to 30 • communicating basic place-value knowledge by recognizing ten objects as a tens unit • comparing length when shown two similar objects • classifying shapes based on a given attribute (for example, number of sides)

DLM Performance Level Descriptors: Mathematics 1 of 18
© 2022 Accessible Teaching, Learning, and Assessment Systems (ATLAS), University of Kansas

At Target	A student who achieves at the **at target** performance level typically uses attributes or characteristics to identify and sort familiar objects into sets, makes sense of problems and perseveres in solving them, identifies repeating calculations or patterns, and uses mathematical terms and identifies connections between mathematical concepts. The student uses attributes or characteristics to identify and sort familiar objects into sets by • understanding the difference between parts of objects and whole objects The student makes sense of problems and solves them by • identifying the place value of two-digit numbers to the tens place • selecting appropriate tools for measuring • calculating the length of objects using informal units of measurement • identifying shapes divided into fractional parts • recognizing the structure of a picture or bar graph • identifying symbols used in equations (for example, =, −, +) The student identifies repeating calculations or patterns by • classifying data based on given attributes (for example, number of objects) • skip counting by tens (for example, 10, 20, 30) The student uses mathematical terms and identifies connections between mathematical concepts by • communicating length in inches and feet
Advanced	A student who achieves at the **advanced** performance level typically calculates accurately and makes sense of problems and perseveres in solving them. The student calculates accurately by • solving repeated addition problems (for example, 2 + 2 + 2 or 3 + 3 + 3) • solving basic addition and subtraction problems with solutions up to 20 • multiplying numbers 1 through 5 The student makes sense of problems and solves them by • answering questions about the data displayed in a graph