DYNAMIC®
LEARNING MAPS

2014-2015 Technical Manual

Year-End

June 2016

Dynamic Learning Maps Consortium. (2016, June). *2014-2015 Technical Manual – Year-End*. Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation.

Acknowledgements

The publication of this technical manual 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 manual possible. We are especially grateful for the contributions of Annie Davidson and to the members of the Dynamic Learning Maps (DLM) Technical Advisory Committee who graciously provided their expertise and feedback. Members of the Technical Advisory Committee include:

Jamal Abedi, Ph.D., *University of California-Davis*
Russell Almond, Ph.D., *Florida State University*
Greg Camilli, Ph.D., *Rutgers University*
Karla Egan, Ph.D., *Independent Consultant*
George Engelhard Jr., Ph.D., *University of Georgia*
James Pellegrino, Ph.D., *University of Illinois-Chicago*
Edward Roeber, Ph.D., *Assessment Solutions Group/Michigan Assessment Consortium*
David Williamson, Ph.D. , *Educational Testing Service*
Phoebe Winter, Ph.D. , *Independent Consultant*

DLM project staff who made significant writing contributions to this technical manual are listed below with gratitude.

Sue Bechard, Ph.D., *Senior Advisor*
Amy Clark, Ph.D., *Psychometrician*
Meagan Karvonen, Ph.D., *Director*
Jonathan Schuster, Ph.D., *Research Associate*
Russell Swinburne Romine, Ph.D., *Associate Director for Test Development and Production*
Jonathan Templin, Ph.D., *Associate Professor of Educational Psychology, University of Kansas*

Project staff who supported the development of this manual through key contributions to design, development, or implementation of the Dynamic Learning Maps Alternate Assessment System are listed below with gratitude.

Lori Andersen
Brianna Beitling
Jennifer Brussow
Karen Erickson
Nora Fairchild
Claire Greer
Lisa Harkrader
Neal Kingston
Allison Lawrence
Caroline Mark
Lee Ann Mills
Michael Muenks
Brooke Nash
Marianne Perie
Lindsay Ruhter
Michelle Shipman
Susan Thomas
Jacob Thompson
Lisa Weeks
Sheila Wells-Moreaux

Table of Contents

List of Tables

List of Figures

I. INTRODUCTION

The Dynamic Learning Maps® (DLM®) Alternate Assessment System assesses student achievement in mathematics and English Language Arts (ELA) 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 support inferences about student achievement, progress, and growth in the given content area. Results provide information that can be used to guide instructional decisions as well as information appropriate for use with state accountability programs.

The DLM Alternate Assessment 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 multiple-choice assessments cannot. The DLM alternate assessment system provides optional, instructionally-embedded testlets that are available for use in day-to-day instruction. A year-end assessment is administered in the spring and results from that assessment are reported for state accountability purposes and programs. This design is referred to as the year-end model and is one of two models for the DLM Alternate Assessment System.[1]

This chapter describes the foundations of the DLM Alternate Assessment System, including the background, history, purpose, and key characteristics of the program. This chapter lays the groundwork for subsequent chapters on the DLM map, assessment design, test development and administration, psychometric modeling, standard setting, reporting, reliability and validity, professional development, and evaluation processes and procedures. An overview of subsequent chapters is included at the end of this chapter. While these chapters describe the essential components of the assessment system separately, several key topics will be addressed throughout this manual, including the DLM maps, accessibility, and validity.

I.1. Background

In 2010, the U.S. Department of Education's Office of Special Education Programs (OSEP), awarded a General Supervision Enhancement Grant to the DLM consortium, which is overseen by the Center for Educational Testing and Evaluation (CETE) in the Achievement and Assessment Institute (AAI) at the University of Kansas.

The DLM project was developed by a consortium of state education agencies (SEAs). In 2010, 13 SEAs were involved: Iowa, Kansas, Michigan, Mississippi, Missouri, New Jersey, North

[1] See Assessments later in this chapter for an overview of both models.

Carolina, Oklahoma, Utah, Virginia, Washington, West Virginia, and Wisconsin. By the end of the fifth year (2015), there were 18 member states with Alaska, Colorado, Illinois, North Dakota, and Vermont joining in 2013 and New Hampshire and Pennsylvania joining in 2014. In the 2014–2015 academic year, all current[2] partner states except Pennsylvania and North Carolina delivered DLM operational assessments in ELA and math.

In addition to CETE and partner states, other key partners during the project included the Center for Literacy and Disability Studies at the University of North Carolina at Chapel Hill, Edvantia (which merged with McREL during the project), The Arc, and the Center for Research Methods and Data Analysis at the University of Kansas. The project was also supported by a technical advisory committee (TAC) and a special education advisory committee.

There were four goals for the OSEP-funded project.

- **Goal 1:** To link the assessment content with the Common Core State Standards (CCSS) by drafting Essential Elements (EEs) and to develop achievement level descriptors that describe what students with the most significant cognitive disabilities should know and be able to do.
- **Goal 2:** To develop ELA and mathematics maps with content appropriate for each grade level. To develop multiple learning tasks for nodes in the ELA and mathematics maps at the appropriate grade level.
- **Goal 3:** To develop a comprehensive computerized system that includes test development, test delivery, test administration, and results reporting.
- **Goal 4:** To develop and implement a professional development program for educators of students with the most significant cognitive disabilities that includes three modes of delivery.

Overall, the four goals were met (Good & Davis, 2015).

- Essential Elements were drafted in the second year of the project, refined in the third year, and approved by the DLM consortium states in July 2013. Alternate achievement standards were developed during a standards-setting meeting in June 2015 and adopted by the consortium in August 2015 after the TAC reviewed and approved the methodology and panel process.
- Primary development of the ELA and mathematics maps occurred in the first three years of the grant, and assessment content was developed in years four and five.
- The four applications that comprise the Kansas Interactive Testing Engine (KITE) system (the DLM maps, Content Builder, Test Delivery Engine, and Educator Portal) were developed by year five to support test development, delivery, administration, and reporting.

[2] Michigan, Virginia, and Washington left the DLM Consortium by 2014–2015 and did not test operationally.

- A series of professional development modules was created. All of the modules are available in two primary formats: self-directed and facilitated. In addition, some of the consortium members used materials from the modules to build customized, state-specific versions, which constituted a third mode of delivery.

The goals of the grant were exceeded in several areas. For example, significantly more assessment items and professional development modules were developed and delivered than what was originally set forth in the cooperative agreement with OSEP. Figure 1 summarizes major milestones across the lifespan of the program.

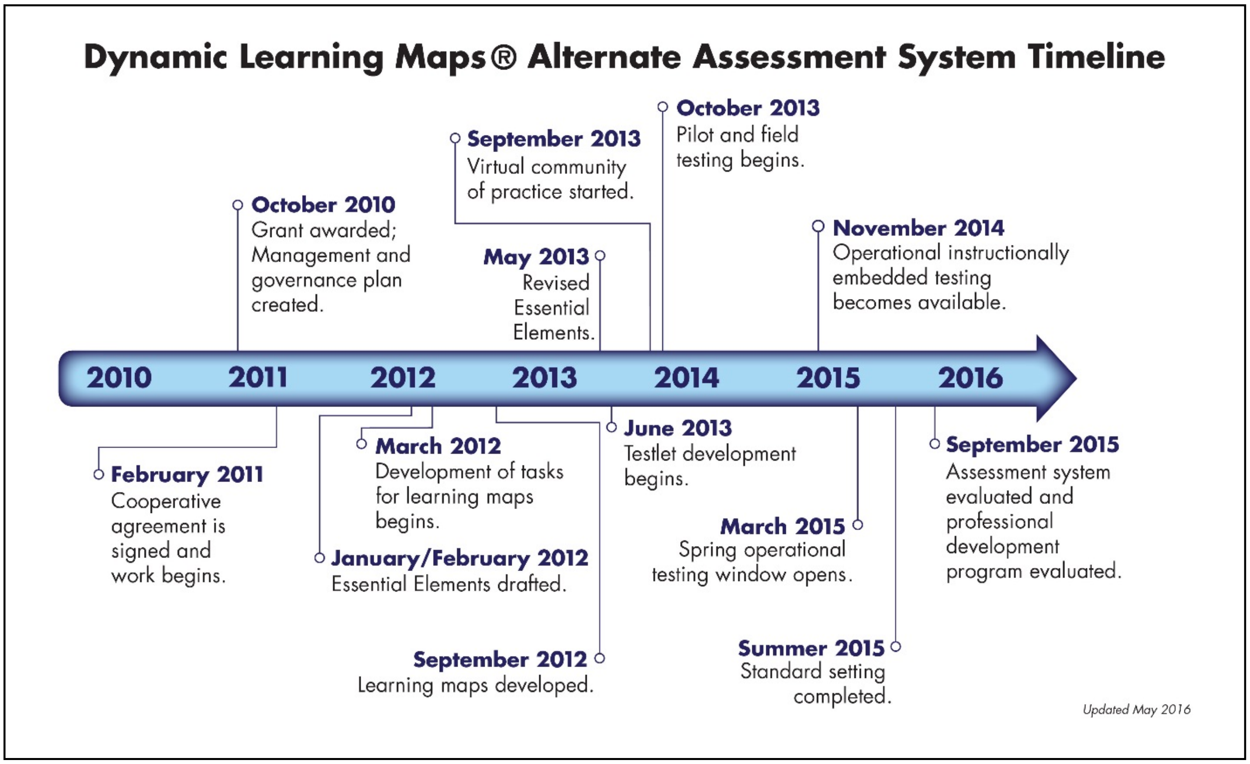

Dynamic Learning Maps® Alternate Assessment System Timeline
2010
2011
2012
2013
2014
2015
2016
October 2010 Grant awarded; Management and governance plan created.
September 2013 Virtual community of practice started.
October 2013 Pilot and field testing begins.
May 2013 Revised Essential Elements.
November 2014 Operational instructionally embedded testing becomes available.
February 2011 Cooperative agreement is signed and work begins.
March 2012 Development of tasks for learning maps begins.
January/February 2012 Essential Elements drafted.
September 2012 Learning maps developed.
June 2013 Testlet development begins.
March 2015 Spring operational testing window opens.
Summer 2015 Standard setting completed.
September 2015 Assessment system evaluated and professional development program evaluated.
Updated May 2016

Figure 1. Five-year timeline for OSEP-funded project, 2010–2015.

I.1.A. Student Population

The Dynamic Learning Maps Alternate Assessment System serves students with the most significant cognitive disabilities, 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, the most significant cognitive disabilities impact students in and out of the classroom and across life domains, not just in academic settings. The DLM Alternate Assessment System is designed for students with these significant instruction and support needs.

The DLM Alternate Assessment System provides the opportunity for students with the most significant cognitive disabilities to show what they know instead of documenting only what they do not know. 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, derived from the CCSS (often referred to in this manual as college and career readiness standards), are called Essential Elements and are the learning targets for the DLM assessments for grades 3-12 in ELA and mathematics.

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 Essential Elements 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 Essential Elements 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 measureable 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 Alternate Assessment System eligibility criteria also include 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

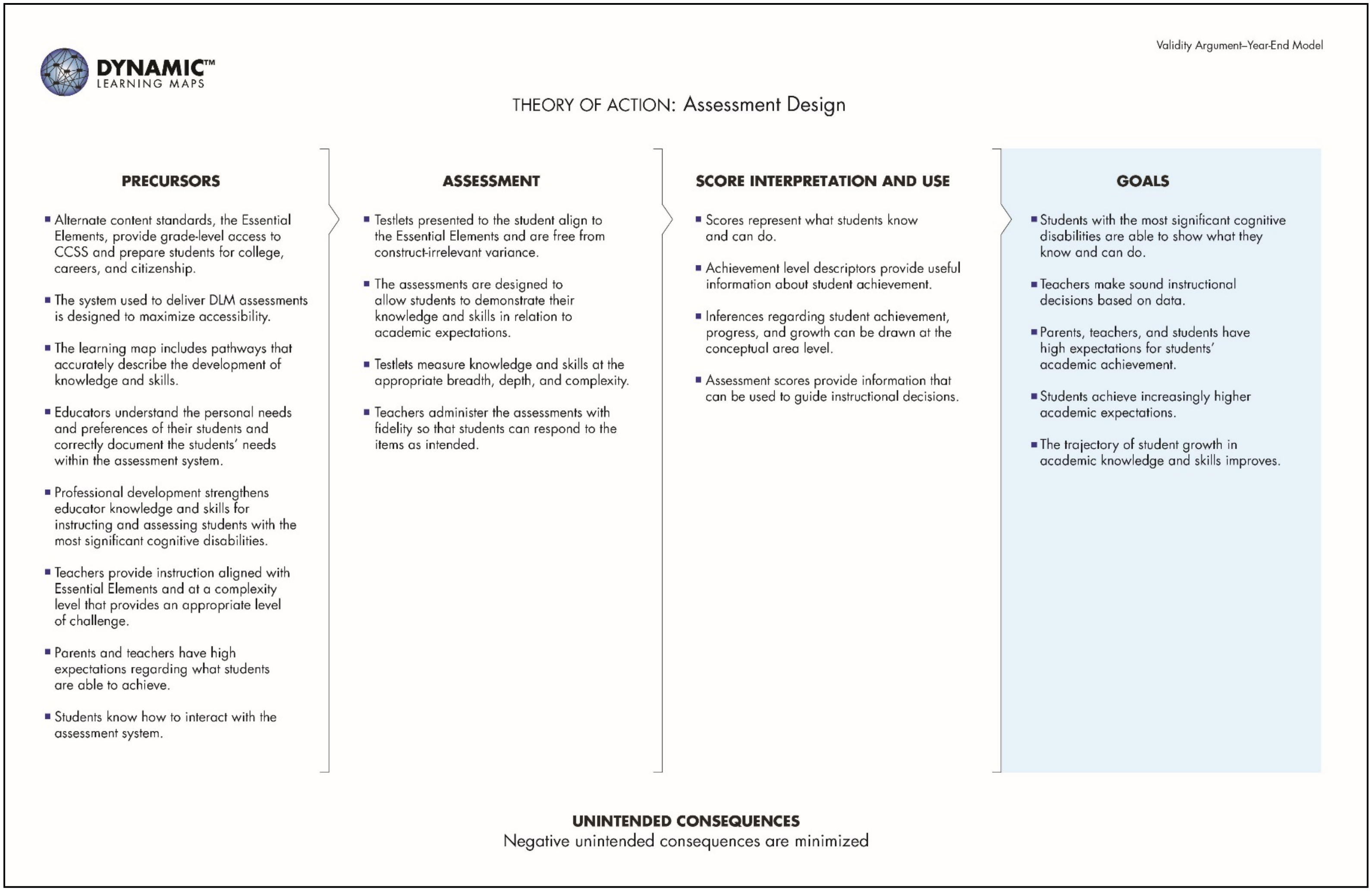
I.1.B. Theory of Action

The theory of action that guided the design of the DLM Alternate Assessment System was formulated in 2011 and revised and finalized in December 2013. 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, educators, and parents/guardians.

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 Alternate Assessment 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 to 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 ratchet up stakes for students and 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 progress and achievement. By using diagnostic information to inform instruction, educators will understand how to build the depth and breadth of conceptual understanding and will think differently about how to educate students in the context of DLM maps. Ultimately, educators, parents/guardians, and others will hold higher expectations of students, and the educational experiences and growth of students will continually improve.

After identifying these overall guiding principles and anticipated outcomes, specific elements of the DLM Alternate Assessment System theory of action were articulated to inform assessment design and to highlight the associated validity arguments. The theory elements were organized around four main topics: precursors to assessment development and implementation, assessment features, score interpretation and use, and goals of the assessment system (see Figure 2).



DYNAMIC™ LEARNING MAPS

Validity Argument–Year-End Model

THEORY OF ACTION: Assessment Design

PRECURSORS

- Alternate content standards, the Essential Elements, provide grade-level access to CCSS and prepare students for college, careers, and citizenship.
- The system used to deliver DLM assessments is designed to maximize accessibility.
- The learning map includes pathways that accurately describe the development of knowledge and skills.
- Educators understand the personal needs and preferences of their students and correctly document the students' needs within the assessment system.
- Professional development strengthens educator knowledge and skills for instructing and assessing students with the most significant cognitive disabilities.
- Teachers provide instruction aligned with Essential Elements and at a complexity level that provides an appropriate level of challenge.
- Parents and teachers have high expectations regarding what students are able to achieve.
- Students know how to interact with the assessment system.

ASSESSMENT

- Testlets presented to the student align to the Essential Elements and are free from construct-irrelevant variance.
- The assessments are designed to allow students to demonstrate their knowledge and skills in relation to academic expectations.
- Testlets measure knowledge and skills at the appropriate breadth, depth, and complexity.
- Teachers administer the assessments with fidelity so that students can respond to the items as intended.

SCORE INTERPRETATION AND USE

- Scores represent what students know and can do.
- Achievement level descriptors provide useful information about student achievement.
- Inferences regarding student achievement, progress, and growth can be drawn at the conceptual area level.
- Assessment scores provide information that can be used to guide instructional decisions.

GOALS

- Students with the most significant cognitive disabilities are able to show what they know and can do.
- Teachers make sound instructional decisions based on data.
- Parents, teachers, and students have high expectations for students' academic achievement.
- Students achieve increasingly higher academic expectations.
- The trajectory of student growth in academic knowledge and skills improves.

UNINTENDED CONSEQUENCES

Negative unintended consequences are minimized

Figure 2. *Dynamic Learning Maps theory of action for the year-end model.*

I.1.C. KEY FEATURES

Consistent with the theory of action, key elements were identified to guide the design of the DLM Alternate Assessment System. The list below mirrors the organization of this manual and provides chapter references. Terms are defined in the glossary (Appendix A.1).

1. **Fine-grained learning map models that guide instruction and assessment**

   Learning map models are a unique key feature of the DLM Alternate Assessment System and drive the development of all other components. While the DLM 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 and alternate 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, Hess, Camacho, Russell, & Thomas, 2012). A fine-grained learning map model provides a great advantage in measuring growth, especially growth within short periods of time or for students who learn more slowly or idiosyncratically than the typical learner. The use of DLM maps helps to realize a vision of a cohesive, comprehensive system of assessment. DLM map development is described in Chapter II.

2. **A subset of particularly important nodes that serve as grade-level content standards and provide an organizational structure for educators**

   Crucial to the use of fine-grained learning map models for instruction and test development is the selection of nodes that serve as learning targets accompanied by the selection of nodes that build the knowledge, skills, and abilities required to achieve the content standard expectations for each grade and content area. This neighborhood of nodes forms a local learning progression toward a specific learning target. The development of EEs and the selection of nodes for assessment are described in Chapter III.

3. **Instructionally relevant testlets that model good instruction and reinforce learning**

   Instructionally relevant assessments consist of activities an educator would want to do for purely instructional purposes, combined with the systematic gathering and analysis of data. These assessments necessarily take different forms depending on the population of students and the concepts being taught. 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 in DLM. Item and testlet design are described in Chapter III.

4. **Instructionally embedded assessments that reinforce the primacy of instruction**

   The DLM alternate assessment is designed as an adaptive, computer-delivered, instructionally embedded assessment that is intended to be relaxed, constant, and integrated with classroom instruction. It also includes an end-of-year assessment that, either separately or in combination with the instructionally embedded assessment, is used to meet the requirements of accountability systems. Embedded assessments must be sensitive to the access needs of the student and the curricular needs of the educator. The DLM assessments provide flexibility in the selection and delivery of testlets so that educators can customize the assessment experience for each student. Test administration is described in Chapter IV.

5. Accessibility by design and alternate testlets

Accessibility is a prerequisite to validity, the degree to which a test score interpretation is justifiable for a particular purpose and supported by evidence and theory (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Therefore, throughout all phases of development, the DLM Alternate Assessment 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 adaptive dynamic routing. Since all students taking an alternate assessment based on alternate academic achievement standards are students with the most significant disabilities, accessibility supports are universally available. The emphasis is on selecting the appropriate accessibility features and tools for each individual student. Accessibility considerations are described in Chapter II (alternate pathways), Chapter III (testlet development), and Chapter IV (accessibility during test administration).

6. Status and growth reporting that is readily actionable

Due to the unique characteristics of a map–based system, DLM requires 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 map. Each EE has related nodes at five associated levels of complexity, called linkage levels. Diagnostic classification modeling is used to determine a student's likelihood of mastering each linkage level associated with each EE. A student's overall performance level in the subject is determined by aggregating linkage level mastery information across EEs. This scoring model supports reports that can be immediately used to guide instruction and describe levels of mastery. The DLM modeling approach is described in Chapter V and score report design is described in Chapter VII.

I.2. System Components

I.2.A. Learning Map Models

The DLM Alternate Assessment System is based on large, fine-grained learning map models. These learning map models are highly connected representations of how academic skills are acquired, as reflected in research literature. The DLM maps consist of nodes that represent discrete knowledge, skills, and understandings in either ELA or mathematics, as well as important foundational skills that support student learning of the targets associated with grade-level content standards. Connections between nodes represent the development of skills and understandings. With approximately 1,900 nodes in the ELA map, 2,400 nodes in the mathematics map, and over 140 foundational nodes[3] that are associated with both content areas, the maps go beyond traditional learning progressions to include multiple and alternate pathways by which students may develop content knowledge.

Seen in its entirety, the DLM map is highly complex, as shown in, Figure 3 which displays a large section of the mathematics map, with the nodes in red boxes and the connecting lines in black.

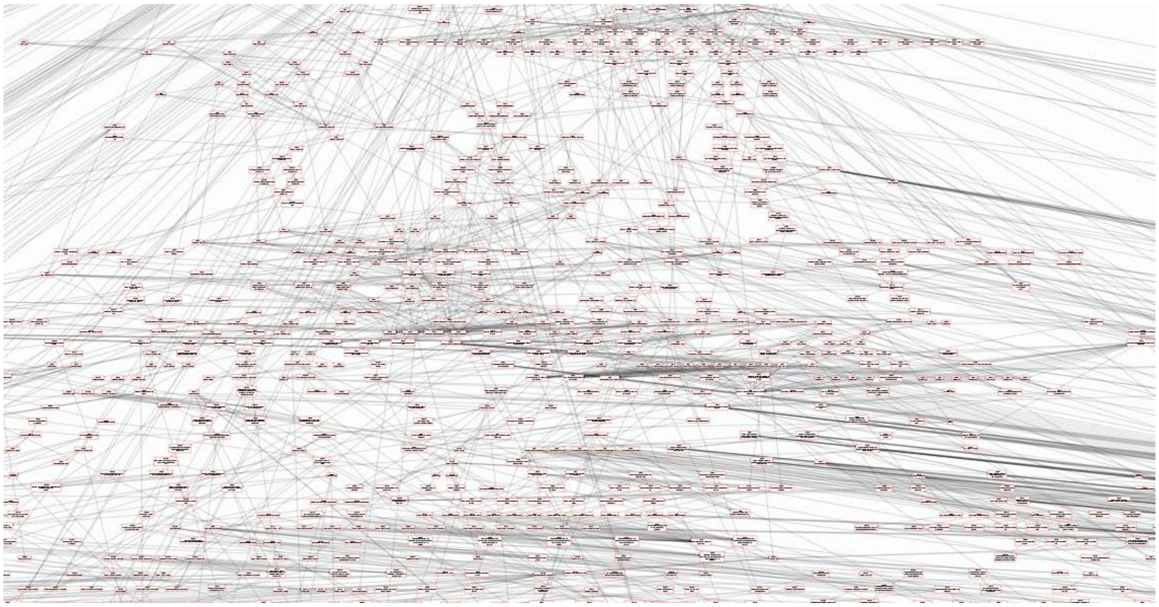


Figure 3. Section of the mathematics map.

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

A closer look at smaller sections of the map reveals how the discrete nodes are described and connected. Figure 4 provides an illustration of a small segment of the ELA map. DLM maps are read from the top down, moving from the least to most complex concepts.

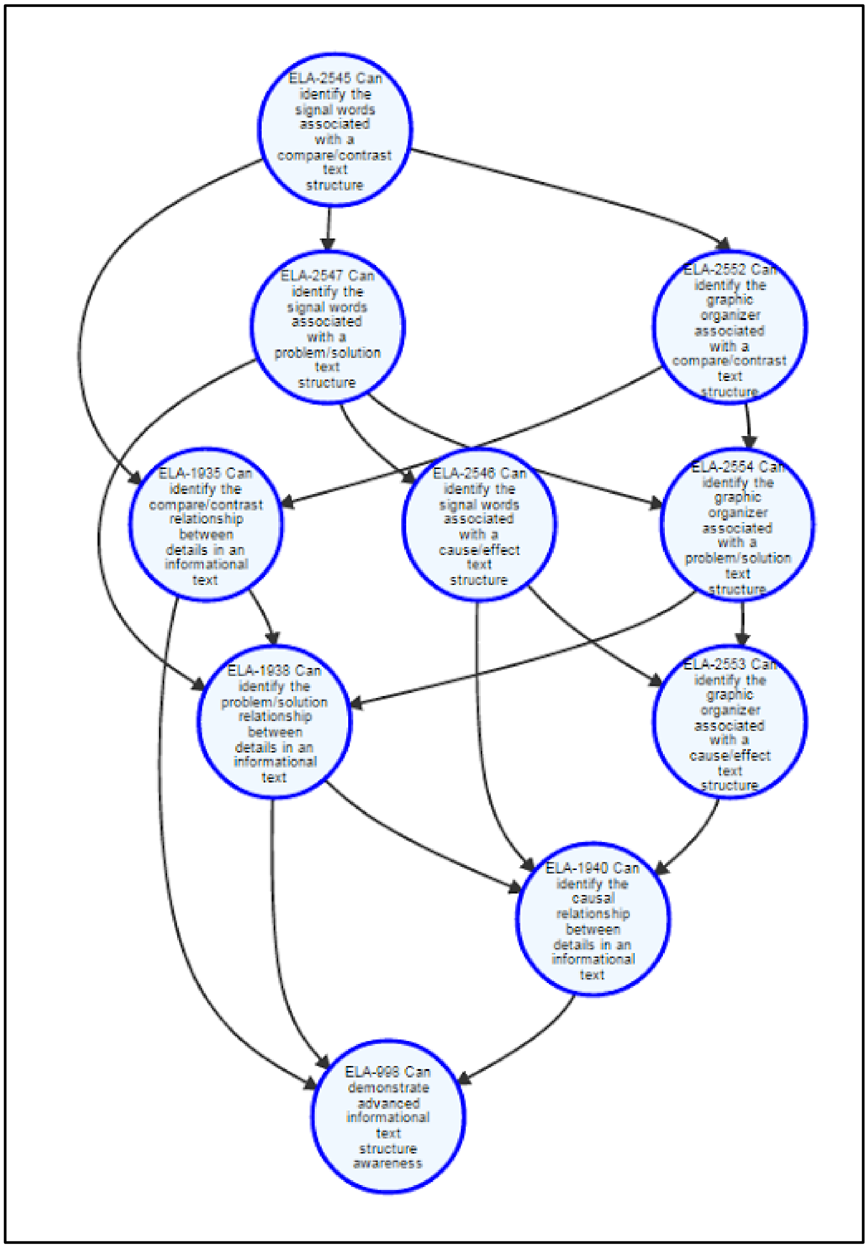

ELA-2545 Can identify the signal words associated with a compare/contrast text structure
ELA-2547 Can identify the signal words associated with a problem/solution text structure
ELA-2552 Can identify the graphic organizer associated with a compare/contrast text structure
ELA-1935 Can identify the compare/contrast relationship between details in an informational text
ELA-2548 Can identify the signal words associated with a cause/effect text structure
ELA-2554 Can identify the graphic organizer associated with a problem/solution text structure
ELA-1938 Can identify the problem/solution relationship between details in an informational text
ELA-2553 Can identify the graphic organizer associated with a cause/effect text structure
ELA-1940 Can identify the causal relationship between details in an informational text
ELA-998 Can demonstrate advanced informational text structure awareness

Figure 4. Sample excerpt from the DLM ELA map.

Given the large amount of information contained in the maps, an organizational structure was designed to articulate where the content standards are located and their relationships to important cognitive concepts. This organization of the academic content in the DLM Alternate

Assessment System is illustrated conceptually in three layers (claims, conceptual areas, and EEs), as shown in Figure 5. In brief, claims are broad statements about what the DLM Consortium expects students to learn and be able to demonstrate within each content area. 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. Essential Elements are based on the general education grade-level content standards, but are at reduced depth, breadth, and complexity. They link the general education content standards to grade-level expectations that are at an appropriate level of rigor and challenge for students with the most significant cognitive disabilities. This organization is discussed in more detail below.

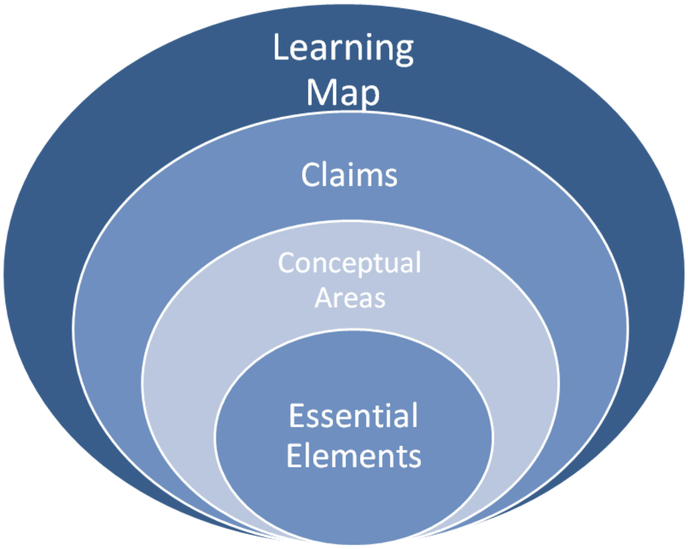


Figure 5. Layers of content in the DLM Alternate Assessment System.

The EEs specify academic targets, while the DLM map clarifies how students can reach those targets. For each EE, neighborhoods of nodes, called linkage levels, are identified as assessment targets. Assessment items are based on nodes at the five linkage levels: Initial Precursor (IP), Distal Precursor (DP), Proximal Precursor (PP), Target (T), and Successor (S).

The overall structure of the DLM Alternate Assessment System had four key relationships between system elements (see Figure 6):

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

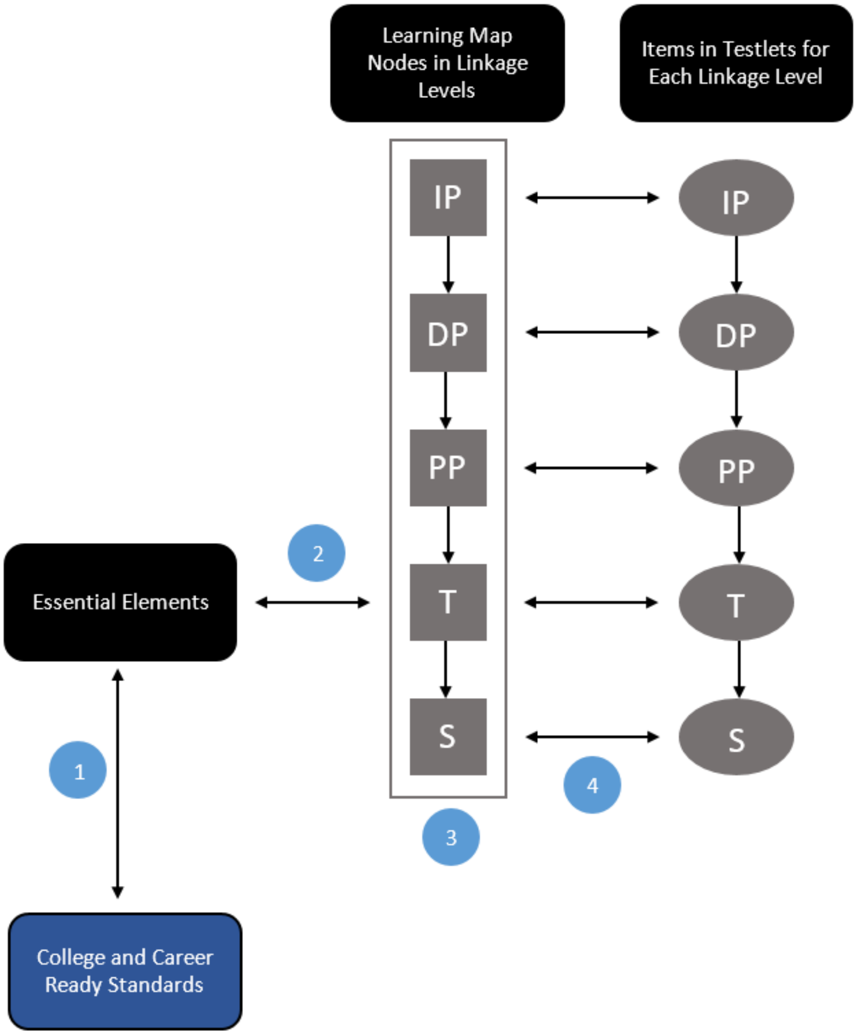

Learning Map Nodes in Linkage Levels
Items in Testlets for Each Linkage Level
IP
DP
PP
T
S
IP
DP
PP
T
S
Essential Elements
College and Career Ready Standards
1
2
3
4

Figure 6. *Relationships in the DLM Alternate Assessment System.*

Note: Linkage levels are Initial Precursor (IP), Distal Precursor (DP), Proximal Precursor (PP), Target (T), and Successor (S).

I.2.B. CLAIMS AND CONCEPTUAL AREAS

Modern test development approaches, such as evidence-centered design (Mislevy, Steinberg, & Almond, 1999), are founded on the idea that test design starts with specific claims about what students know and are able to do and the evidence needed to support such claims. While evidence-centered design is multifaceted, it starts with a set of claims regarding significant knowledge in the domains of interest (e.g., mathematics and ELA) as well as an understanding of how that knowledge is acquired.

Regions of the DLM maps that reflect single EEs can be displayed in mini-maps, which detail the nodes that constitute the EE's linkage levels. Larger sections of the map 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.

The DLM Alternate Assessment System divides both ELA and mathematics content into four broad claims, which are subdivided into nine conceptual areas for each content area (Table 1 and Table 2). The claims and conceptual areas apply to all grades in the DLM Alternate Assessment System. Claims are overt statements of what students are intended to learn as a result of mastering skills within a broad section of the map. Conceptual areas are nested within claims and are made up of multiple conceptually related content standards and nodes that support and extend beyond those standards. This system of claims and conceptual areas organizes the map, which is otherwise too complex to use effectively.

The claims that have been developed for the DLM Alternate Assessment System identify the major domains of interest within ELA (Table 1) and mathematics (Table 2) for students with the most significant cognitive disabilities. As broad statements about expected student learning, claims focus the scope of the assessment. Because the DLM map identifies possible paths by which students may acquire academic skills, the claims also help organize the structures of related knowledge, skills and abilities represented in the DLM maps for this population of students. Thus, the claims serve as a foundation for evaluating the validity of inferences made from test scores.

Conceptual areas further define the knowledge and skills required to meet the broader claims. Each claim includes two or three conceptual areas. Conceptual areas are regions of the DLM map organized around common cognitive processes and content.

Table 1. ELA Claims and Conceptual Areas

Claim	Conceptual Area	
Students can comprehend text in increasingly complex ways.	1.1	Determine critical elements of text.
	1.2	Construct understandings of text.
	1.3	Integrate ideas and information from text.
Students can produce writing for a range of purposes and audiences.	2.1	Use writing to communicate.
	2.2	Integrate ideas and information in writing.
Students can communicate for a range of purposes and audiences.	3.1	Use language to communicate with others.
	3.2	Clarify and contribute in discussion.
Students can investigate topics and present information.	4.1	Use sources and information.
	4.2	Collaborate and present ideas.

Table 2. Mathematics Claims and Conceptual Areas

Claim	Conceptual Area	
Students demonstrate increasingly complex understanding of number sense.	1.1	Understand number structures (counting, place value, fractions, etc.).
	1.2	Compare, compose, and decompose numbers and sets.
	1.3	Calculate accurately and efficiently using simple arithmetic operations.
Students demonstrate increasingly complex spatial reasoning and understanding of geometric principles.	2.1	Understand and use geometric properties of two- and three-dimensional shapes.
	2.2	Solve problems involving area, perimeter, and volume.
Students demonstrate increasingly complex understanding of measurement, data, and analytic procedures.	3.1	Understand and use measurement principles and units of measure.
	3.2	Represent and interpret data displays.
Students solve increasingly complex mathematical problems, making productive use of algebra and functions.	4.1	Use operations and models to solve problems.
	4.2	Understand patterns and functional thinking.

Figure 7 provides an example of a conceptual area.

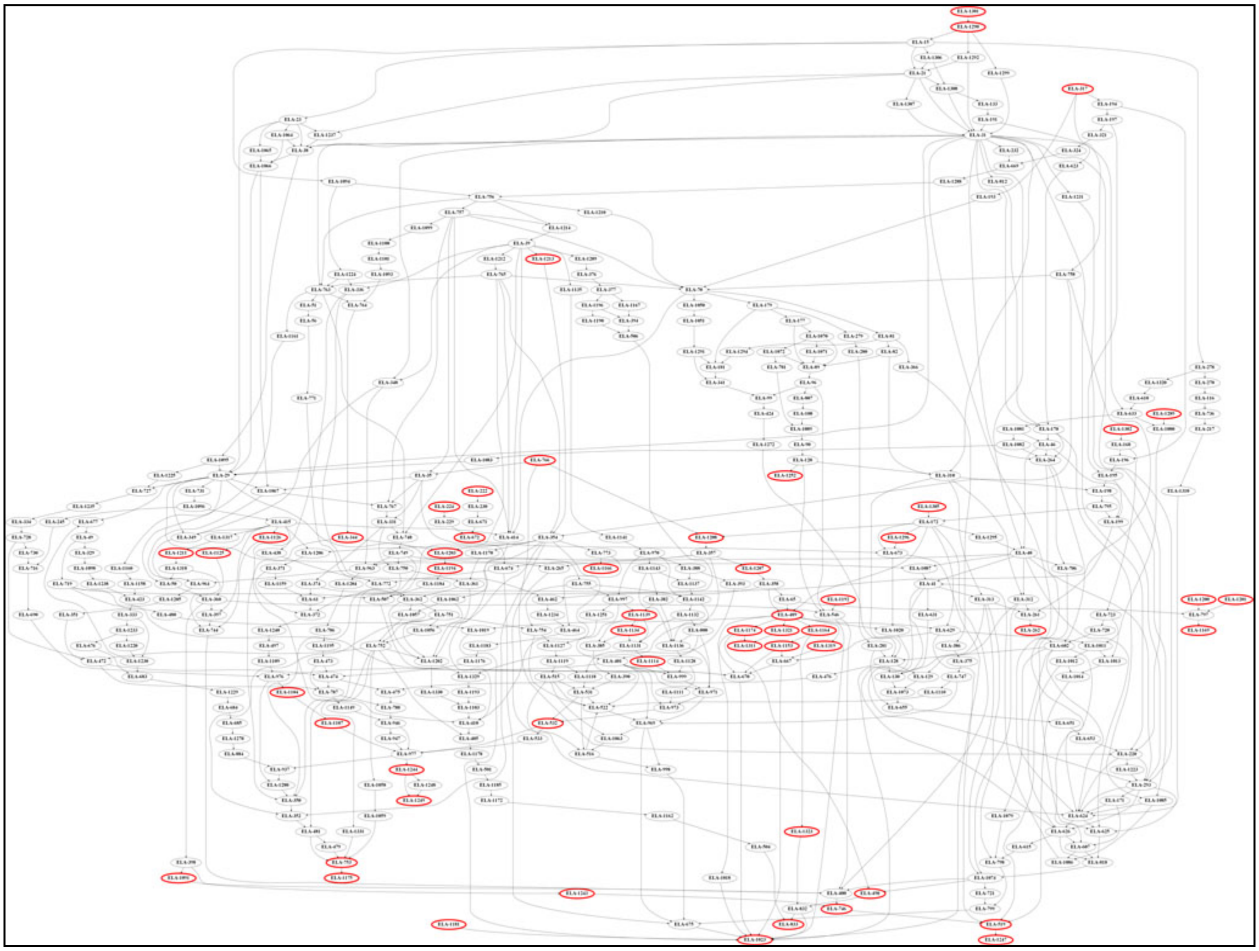


Figure 7. Section of the DLM ELA map for the conceptual area CA 1.2: Construct understandings of text. The red circles mark nodes aligned to EEs.

The DLM claims and conceptual areas provide a framework for organizing nodes on the DLM maps and, accordingly, the EEs.

I.2.C. Essential Elements

The Dynamic Learning Maps EEs are specific statements of knowledge and skills. 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. In other words, EEs are the alternate content standards of the college and career readiness content standards used in general education assessments. The DLM EEs within a particular claim or conceptual area link to one another, and the DLM map reflects the paths a student may take to acquire the knowledge and skills within a claim or conceptual area. An EE is located within a conceptual area based on the cognitive processes and skills required to meet the learning target described by the EE.

The progression of content and skills 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 skills that are not of obvious increasing rigor compared to the grade-to-grade differences found in the general education college and career readiness standards. However, to the degree possible, the skills 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 college and career readiness standards.

The EEs specify academic targets, while the DLM map clarifies how students can reach those targets. The DLM assessments are aligned to grade-level content standards at reduced depth, breadth, and complexity in order to be appropriate for the student population. For each EE, small collections of nodes are identified earlier in the map that represent critical junctures on the path toward the standard. Nodes are also identified past the standard, in order to give students an opportunity to grow toward the grade-level targets for students without significant cognitive disabilities.

The small collections of related nodes are called linkage levels. The Target linkage level reflects the grade-level expectation in the EE. There are three linkage levels below the Target (Initial Precursor, Distal Precursor, and Proximal Precursor) and one linkage level beyond the Target (Successor). Table 3 and Table 4 show examples of related system elements in ELA and mathematics. Both tables provide specific examples of the layers illustrated in Figure 5 with the addition of linkage levels and DLM map nodes. Elements are shown on the left, from broadest to most specific, and descriptions are provided on the right.

Table 3. Assessment System Elements with Examples for ELA

Element	Description
ELA Claim 1 (C1)	Student can comprehend text in increasingly complex ways.
ELA Conceptual Area 1 (C1.1)	Determine critical elements of text.
Essential Element RL.3.1	Can produce responses to questions seeking information on specific characters and what each of them did in a narrative by providing details on them.
Target Linkage Level	Answer who and what questions to demonstrate understanding of details in a text.
DLM Map Node	ELA-1678: Can answer who and what questions about details in a narrative.

Table 4. Assessment System Elements with Examples for Mathematics

Element	Description
Math Claim 1 (C1)	Students demonstrate increasingly complex understanding of number sense.
Math Conceptual Area 3 (C1.3)	Calculate accurately and efficiently using simple arithmetic operations.
Essential Element 6.NS.2	Apply the concept of fair share and equal shares to divide.
Target Linkage Level	Demonstrate the concept of division.
DLM Map Nodes	M-549: Divide by 1. M-550: Divide by 2. M-551: Divide by 3. M-552: Divide by 4. M-553: Divide by 5. M-558: Divide by 10.

While Table 3 and Table 4 show only the target linkage level for the example EE, the other linkage levels are also included in the overall structure of the system design, with different nodes assigned to each linkage level. Nodes in these five linkage levels are the basis for developing assessment items as shown above in Figure 6. Additionally, the nodes and their relationships are described in mini-maps that item writers use during test development. Examples of nodes associated with each linkage level are provided in Chapter III.

I.2.D. ASSESSMENTS

The DLM assessments are delivered as a series of testlets, each of which contains an unscored engagement activity and three to eight items. Assessment items are written to align to nodes at one of the five linkage levels and are clustered into testlets (see Figure 8). Therefore, each linkage level is specifically assessed. Students are placed in the assessment at the appropriate linkage level based on information collected about their expressive communication and academic skills. Suggestions for the next appropriate testlet are provided by the system, based on the student's performance.

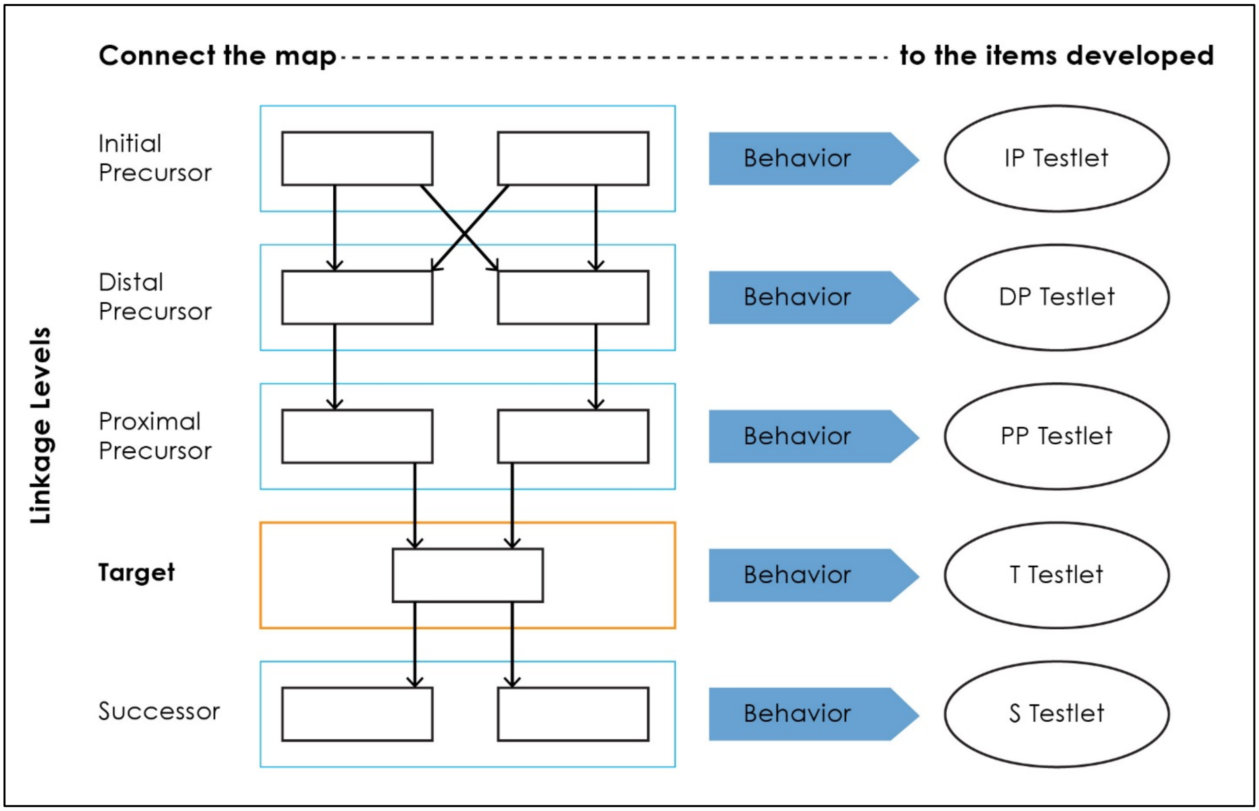


Figure 8. *Relationship between DLM map nodes in five linkage levels and items in testlets. Small black boxes represent nodes in the DLM map. Blue and orange boxes represent collections of nodes in linkage levels. The orange box denotes the Target linkage level for the EE. There may be more than one node at any linkage level.*

Assessment blueprints consist of EEs prioritized for assessment by the DLM Consortium. To achieve blueprint coverage, each student is administered a series of testlets. Each testlet is delivered through an online platform, the Kansas Interactive Testing Engine (KITE). Student results are based on evidence of mastery of the linkage levels for every assessed EE.

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

- **Integrated model.** In the first of two general testing windows, instructionally embedded assessments occur throughout the fall, winter, and early spring. Educators have some choice of which EEs to assess, within constraints. For each EE, the system recommends a linkage level for assessment and the educator may accept the recommendation or choose another linkage level. During the second testing window in the spring, all students are re-assessed on several EEs on which they were taught and assessed earlier in the year. During the spring window the system assigns the linkage level based on student performance on previous testlets; the linkage level for each EE may be the same as or different from what was assessed during the instructionally embedded window. At the end of the year, scores used for summative purposes are based on mastery estimates for linkage levels for each EE (including performance on all instructionally embedded and spring testlets). The pools of operational assessments for the instructionally embedded and spring windows are separate.
- **Year-end model.** In a single operational testing window in the spring, all students take testlets that cover the whole blueprint. Each student is assessed at one linkage level per EE. The linkage level for each testlet varies based on student performance on the previous testlet. The assessment results reflect the student's performance and are used for accountability purposes each school year. The instructionally embedded assessments are available during the school year but are optional and do not count toward summative results. In two states, the high school blueprints are based on End-of-Instruction courses rather than specific grades.

Information in this manual is common to both models wherever possible and is specific to the year-end model where appropriate. A separate version of the TECHNICAL MANUAL exists for the integrated model.

I.3. TECHNICAL MANUAL OVERVIEW

This manual provides evidence to support the DLM Consortium's assertion of technical quality and the validity of assessment claims.

Chapter I provides the theoretical underpinnings of the DLM Alternate Assessment System, including the background, purpose, rationale, target student population, problems addressed, and design. The chapter describes how assessment claims and conceptual areas were identified in the DLM map and how EEs, linked to the conceptual areas, 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 II describes the process by which the DLM maps were developed. Extensive, detailed work was necessary to create the DLM 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 maps are the conceptual and content basis for the DLM Alternate Assessment System.

Chapter III outlines procedural evidence related to test content and response process propositions[4]. It 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 the DLM map nodes 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 III includes summaries of external reviews for content, bias, and accessibility. The final portion of the chapter describes the pilot and field tests.

Chapter IV provides an overview of the fundamental design elements that characterize test administration and how each element supports the DLM theory of action. The chapter relates how students are assigned their first testlet using the First Contact survey results and describes the assessment delivery modes (computer delivery and teacher delivery) and assessment windows (instructionally embedded and spring). The following sections briefly describe test administration protocols, accessibility tools and features, test security, and system usability.

Chapter V 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.

Chapter VI 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 operational assessment administration. This chapter also describes the process of developing grade- and subject-specific performance level descriptors in ELA and mathematics.

Chapter VII reports the 2014–2015 operational results, including student participation data. The chapter details the percent of students at each performance level (impact); subgroup performance by gender, race, ethnicity, and English language learner status; and the percent of students who showed mastery at each linkage level. Finally, the chapter provides descriptions of all types of score reports, data files, and interpretive guidance.

Chapter VIII 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, and subject (overall performance).

[4] The term "proposition" 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 III).

Chapter IX describes additional validation evidence not covered in previous chapters. It looks back at the intended score uses and interpretations as stated in the theory of action, and it details the evaluation of test content through review and alignment study results. The chapter relates how response processes were evaluated through cognitive lab results and review of test score integrity and how the internal structure of the assessment was evaluated through dimensionality and differential item functioning studies as well as a review of the DLM map and external alignment studies. Finally, the chapter discusses the consequences of assessment in terms of intended and potentially unintended consequences.

Chapter X describes the training and professional development that was offered across the DLM Consortium, including the 2014–2015 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 2014–2015 instructional professional development are included.

Chapter XI synthesizes the evidence provided in the previous chapters. It evaluates how the evidence supports the intended interpretations and uses of results from the 2014–2015 DLM assessments.

II. MAP DEVELOPMENT

Chapter I provided an introductory description and illustration of the DLM maps[5] in light of the Dynamic Learning Maps Alternate Assessment System purpose and program goals. In Chapter II, the development process for the DLM maps is described. Extensive, detailed work was necessary to establish and flesh out the DLM maps in light of the Common Core State Standards (CCSS) 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 maps are the conceptual and content basis for the DLM® Alternate Assessment System.

II.1. Design of the Dynamic Learning Maps Assessment System

Learning map models are a type of cognitive model composed of multiple interconnected learning targets and other critical knowledge and skills. The development of the DLM assessment system's learning map models began with a review of the existing literature on learning progressions, a widely accepted and similar approach to assessing student growth (Daro, Mosher, & Corcoran, 2011; Heritage, 2008). Learning progressions identify an academic target and the sequenced building blocks that precede the mastery of this skill (Popham, 2011). Progressions have been used most widely in formative assessment, assisting educators in understanding the gap between current performance and a grade-level standard. However, because learning progressions might only depict a single, linear pathway toward the academic target, they often represent only the most commonly used route that an average student follows.

Despite their utility for typical learners, linear learning progressions have had limited relevance to students with the most significant cognitive disabilities (Kearns, Towles-Reeves, Kleinert, Kleinert, & Thomas, 2011). Because learning progressions have been developed for the general education population and frequently contain only a single, linear pathway toward an academic target, they are unable to represent significant variations in learning (e.g., acquiring writing skills with limited mobility or learning to read with hearing impairments). Students with the most significant cognitive disabilities have sensory differences that require pathways circumventing the potentially problematic skills located in learning progressions. To overcome this issue, the DLM project expanded upon existing notions of learning progressions by including additional building blocks on the way to learning targets and by showing the hypothesized connections and interactions between different learning targets. These changes to the typical learning progression formed a learning map model, which is a web-like network of connected learning targets (Bechard, Hess, Camacho, Russell, & Thomas, 2012). To complement the progression of grade-level learning targets, the DLM maps also depict the skills and knowledge acquired between birth and school entry, which provides the foundation for their development. Additionally, the DLM maps provide access to multiple and alternate routes to

[5] In this chapter "learning map models" and "DLM maps" refer to the specific learning map models developed to support the DLM assessment system.

achieving the learning targets, making it more inclusive for learners with various disabilities (Erickson & Karvonen, 2014).

In summary, designing the DLM maps was an attempt to overcome some potential weaknesses of simple, linear learning progressions by combining multiple learning progressions that cover related topics together in a single representation. Thus, the DLM maps consist of numerous connections between the multiple learning progressions that cover the development of the cognitive and content-area skills from birth to high-school graduation. The numerous connections between the multiple learning progressions provide pathways for all students to acquire skills that are critical for mastering grade-level learning targets. The current versions of the DLM maps contain hypothesized representations of potential learning pathways described in research literature and were designed to be as inclusive as possible for students with the most significant cognitive disabilities. Assessments created using DLM maps will provide frequent opportunities for revising and improving this representation as student responses are collected and a better understanding of student learning is achieved.

II.1.A. DEVELOPMENT PROCESS

The foremost goal of using the DLM maps is to support the process of making inferences about student mastery in the context of a large-scale assessment. As a result, choosing an analysis method was central to early decisions about the structural design of the maps. Bayesian Network analysis is a type of probabilistic model used to make inferences where interconnected factors are present (Pearl, 1988; Koller & Friedman, 2009). In particular, a Bayes Net allows for explicit description of the relationships between connected skills, to facilitate in making inferences of skill proficiencies from indirect data (e.g., inferring the mastery of precursor nodes when a subsequent node has already been mastered), and to allow for efficient storage and computation regardless of the size of the learning map model.

Like Bayesian Networks, DLM maps consist of two basic elements: nodes and connections. The nodes are essential, unique, observable, and testable knowledge and skills[6]. There are two basic types of nodes: those that represent learning targets and nodes that represent the significant knowledge and skills supporting the development of the learning targets. The second element, connections (called "edges" in Bayesian Network analysis), forms the relationship between nodes.

The DLM Consortium developed maps in English language arts (ELA) and mathematics, both of which begin with a common set of basic skills called foundational nodes. To create these interconnected maps, the DLM Consortium followed a four-step process.

1. Identification and Representation of Learning Targets
2. Identification and Representation of Additional Supporting Skills
3. Linking the DLM Maps to the Essential Elements

[6] For analysis purposes, nodes are latent, dichotomous variables. See Chapter V for additional information.

4. Development of Connections between Nodes and Building Alternate Pathways

Once developed, the first evaluation of the DLM maps consisted of educator and expert review of map sections. Empirical analyses of the structure of the DLM maps as data becomes available is also planned.

II.1.A.i. Learning Targets: CCSS and Essential Elements

The first step was to identify learning targets, which provide a basic frame for the DLM maps. Because the DLM assessment measures student achievement of Essential Elements aligned to college and career-readiness content standards, the Common Core State Standards (CCSS) documents served as a starting place for node development.[7] Specifically, grade-level CCSS standards became individual nodes within the DLM maps. When a standard contained multiple skills unsuitable to be combined into a single node, the incompatible skills were represented as distinct nodes in the DLM maps.

Once the nodes representing the learning targets had been created, they were arranged in the DLM maps according to grade-level(s). From this frame, the supporting knowledge and skill nodes were identified to fill in the gaps between the learning target nodes. This process is further described in the next section.

Because the primary goal of the DLM Consortium is to assess what students with the most significant cognitive disabilities know and can do, alternate grade-level expectations called Essential Elements (EEs) were created to reflect more accurately the knowledge, skills, and understandings that are appropriately challenging grade-level targets for students with the most significant cognitive disabilities. Within each content area and strand/cluster, the EEs were derived from the CCSSs to represent similar skill development sequences as the CCSSs.

EEs were first written based on the CCSS, independent of the map development process in 2012. (See chapter III for a description of EE development.) At the same time that the EEs were being developed, DLM was actively engaged in building the maps in mathematics and ELA. Because the development of the EEs and the maps happened simultaneously, alignment between the EEs and the maps was not possible until the fall of 2012. The process of evaluating the alignment between the EEs and the maps involved reconciling the content of the EEs to the content represented in the nodes and connections of the maps in ELA and mathematics. This process resulted in a revision to the EEs in 2013 and significant revisions to the DLM maps to insure 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 map were aligned to the EE. If no existing node(s) corresponded to the content of the EE, nodes were created and placed in the map models at appropriate locations according to their content. New nodes were placed by analyzing the existing map structure to identify precursor and

[7] The CCSS were initially used in early map development. The Essential Elements were later integrated into the map as an additional set of targets (largely preceding the CCSS targets).

successor skills to the new node. Once identified, content teams proposed placements of new nodes and connections based on literature reviews and expert judgment.

II.1.A.ii. Supporting Knowledge and Skills

After identifying the learning targets, an extensive literature review was conducted to create nodes reflecting the knowledge and skills surrounding the development of these targets. Given that the CCSS for kindergarten begins at a relatively complex cognitive and language level, the content teams employed bottom-up methods in the literature search, looking initially for research concerning early cognitive development (e.g., can attend to object characteristics due to language cues) and then building toward the more advanced grade-level learning targets (e.g., can answer *wh-* questions about details in a narrative). Wherever possible, the content teams used empirical research to drive the development of nodes.

As an example, Table 5 depicts the procedure used by the ELA content team to create the supporting nodes of academic targets. After reviewing the CCSS in a domain area, the team conducted a literature review of articles, books, and book chapters summarizing the developmental research in that domain area. This literature review was the primary source of the supporting skills and knowledge depicted in the learning map models.

Table 5. Node creation procedures using a literature review as source material

Standard	Identified Handbook/ Chapter Book	Identified Author	Key Article	Nodes
Writing Anchor Standards (CCSS.ELA-Literacy.CCRA.W. 4-6 **Production and distribution of writing***	**MacArthur, C. A., Graham, S., & Fitzgerald, J. (Eds.). (2008). *Handbook of writing research.* Guilford Press.**	**V. W. Berninger**	**Berninger, V. W.** **"Development of language by hand and its connections with language by ear, mouth, and eye." *Topics in Language Disorders* 20.4 (2000): 65-84.**	**F-133 Can produce undifferentiated scribbles;** **F-132 Can produce linear scribbles; Can produce scribbles left-to-right, top-to-bottom**

*Note: *This example considers the precursor nodes to production of writing.*

In addition to empirical literature, common instructional practices and other curricular information were used to represent skill development in the gaps between the learning target nodes. Despite the DLM project's focus on students with the most significant cognitive disabilities, the empirical literature on the acquisition of academic skills used in developing the DLM maps in ELA and mathematics was based largely on typical learners.[8] As a result, the content teams focused on first building a "super highway" to represent typical development with multiple pathways to learning targets. The map was then adapted by adding alternate routes for student populations requiring additional or different cognitive skills (see Alternate Pathways heading below).

II.1.A.ii.a Critical Sources

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

II.1.A.ii.b Nodes Reflect the Products of Learning and Cognitive Growth

As previously stated, to be included in the DLM maps, a node must represent essential, unique, observable, and testable aspects of knowledge and skill. Furthermore, the knowledge and skills surrounding and supporting the learning targets in the DLM maps may vary significantly in kind, especially since the DLM maps represent skills and knowledge developed and acquired between birth and school entry. For the supporting knowledge and skills developing between birth and school entry, the individual nodes reflect the learning and cognitive growth that occurs during this period by representing how the skills become increasingly more complex. For example, early skills, such as seeking the attention of others, provides the basis for more complex skills, such as using words to request, comment, and command.

The supporting knowledge and skills that develop after school entry provide the individual stepping stones between academic targets. Because academic targets represent benchmarks students achieve across grades, additional critical skills not mentioned in the CCSS are required

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

to help the students achieve these targets. The aforementioned supporting knowledge and skills constitutes these critical skills. Along with the academic targets, these intermediary skills reflect and result from increased cognitive resources (e.g., increases in working memory allow children to produce longer and more complex sentences) and/or instruction (e.g., students learn basic abstract symbols following exposure and explicit instruction). Regardless of the source, these shifts in thinking and skills form the framework of nodes surrounding grade-level targets.

II.1.A.ii.c Foundation Nodes

Even in the early grades, learning targets associated with grade-level standards require the application of basic skills. These basic skills are required across content domains and include such things as attention, self-regulation, and language, as well as cognitive skills such as categorization. The nodes representing these basic skills form the base of the DLM map and are called foundational nodes. Some students with the most significant cognitive disabilities must be taught learning targets associated with foundational nodes in order to work toward learning targets associated with grade-level standards (Kleinert, Browder, & Towles-Reeves, 2009).

II.1.A.ii.d Node Development Criteria

The nodes representing the learning targets, and the knowledge and skills supporting them, had to meet certain requirements to be included in the DLM map. The first requirement determined whether **the new node was essential for progressing to later learning targets**. Only those nodes that contribute to the development of a learning target were entered into the DLM maps. The decisions for node entry based on this requirement were the result of expert judgment based on the research synthesis conducted by project staff. Nodes could be essential for progressing to single or multiple learning targets. The next requirement focused on whether **the new node was a component of a learning target or was the learning target itself**. As mentioned previously, nodes that were themselves the learning targets were automatically included in the DLM maps. More complex learning targets containing multiple skills were broken down into their component parts, and each part was aligned to one or more nodes. For example, the following fifth-grade ELA learning target includes multiple skills: “Determine a theme of a story, drama, or poem from details in the text, including how characters in a story or drama respond to challenges or how the speaker in a poem reflects upon a topic; summarize the text.” Individual elements of the learning targets, “Determine the theme of a story, drama, or poem” and “Can summarize a narrative,” are separate nodes since the identification of theme and the ability to summarize a text are distinct cognitive skills. In contrast, if a potential new node was a component of the learning target, it was only added to a section of DLM map as a new, separate node if it was not a restatement of the existing node. In other cases, nodes were edited to combine multiple applications of a skill seen in different learning targets. In an example of a combination, the nodes, “Can determine the meaning of words alluding to other narratives” and “Can determine the meaning of phrases alluding to other narratives,” were combined into a single DLM map node, since the cognitive skill (determining meaning in allusion) is essentially the same but is used in different contexts (i.e., single words or phrases).

The third requirement was that the **new node could be assessed**. To be included in the DLM maps, the new node had to be assessable, because all nodes needed to be observed and measured to provide information on a student's ability level. Lastly, **the new node had to be distinct** by providing information that extends the skill(s) acquired in the preceding node(s) but is less complex than the skill(s) acquired in the succeeding node(s). For example, consider this sequence:

Recognize parts of a whole or given unit → Recognize fraction → Recognize denominator

If the new node was not distinct from the preceding or succeeding node(s), it was combined with a node already in the map. In summary, the DLM maps contain only nodes meeting the requirements described, and each node contains important skills and knowledge toward the development of the learning targets.

II.1.A.iii. Development of Connections Between Nodes

After the learning target and supporting (foundational and grade-level) 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 forms a relationship between two nodes—the origin node and the destination node. Origin nodes precede, and are hypothesized to develop before, the acquisition of the destination nodes. Once the most common pathways in a given domain were created and identified, connections between the nodes in different domains were created when appropriate, giving the DLM maps their interconnectedness between all domain areas.

As an example, a small section of the map is provided in Figure 9 and Figure 10. This section of the DLM map covers the node for asking and answering questions when reading a narrative text. Figure 9 illustrates the structure of the map, including multiple pathways, while Figure 10 displays the nodes used for assessment. Both figures highlight a specific pathway to demonstrate the interconnected nature of the DLM maps. The pathway depicts how a student would progress from being able to pay attention to object characteristics as a result of language cues to being able to answer *wh-* questions concerning the details provided in the narrative. 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 DLM map section. The colored nodes in both figures represent linkage nodes, which are nodes that have been identified as making a significant contribution to the development of the learning target by the DLM map developers and content experts. These nodes typically precede or directly follow the development of the target node, but they are not the only nodes contributing to the learning target, nor do they prescribe the only route that can be taken toward acquiring it.

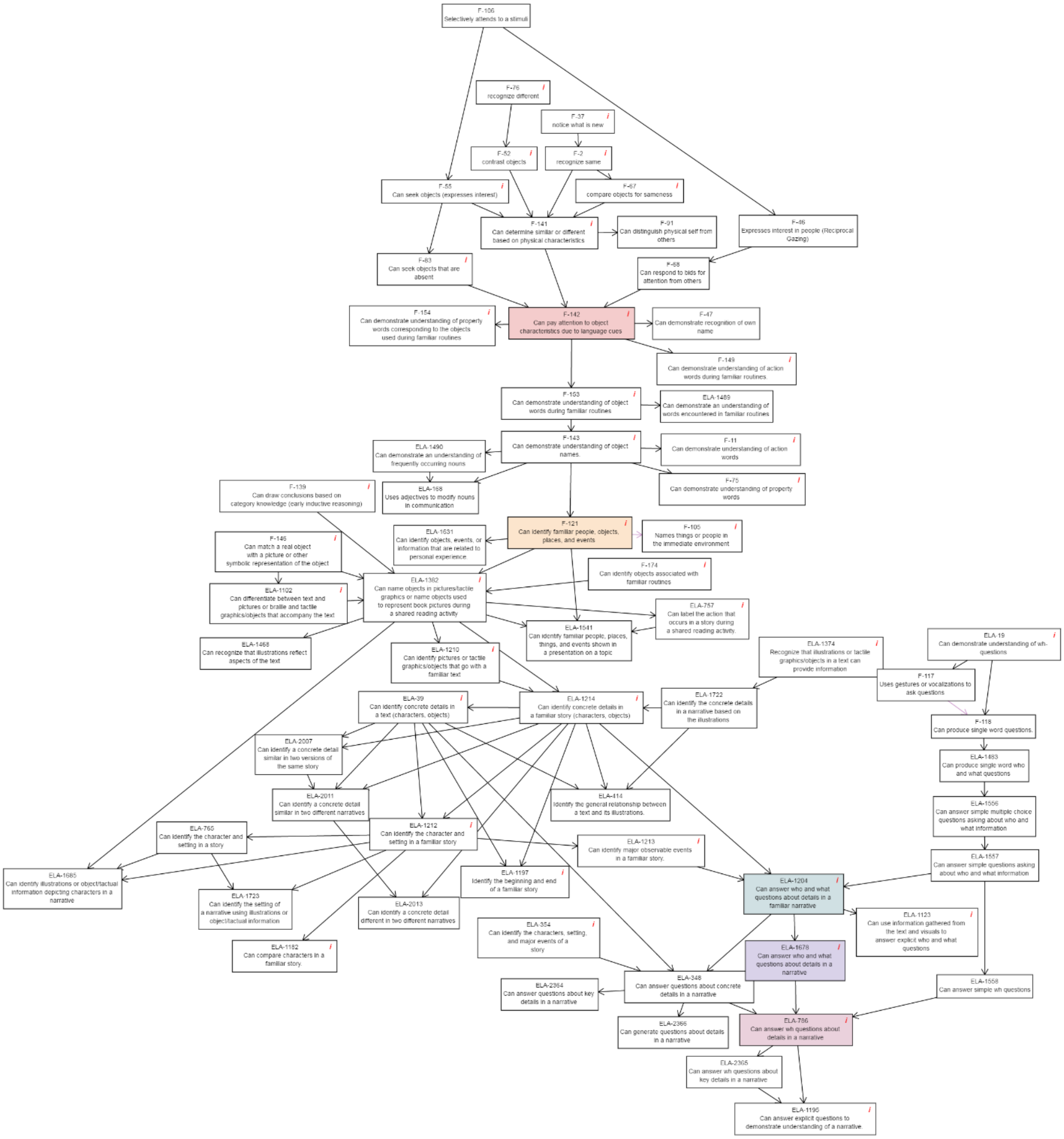

F-106 Selectively attends to a stimuli
F-76 recognize different
F-37 notice what is new
F-52 contrast objects
F-2 recognize same
F-55 Can seek objects (expresses interest)
F-67 compare objects for sameness
F-141 Can determine similar or different based on physical characteristics
F-91 Can distinguish physical self from others
F-46 Expresses interest in people (Reciprocal Gazing)
F-83 Can seek objects that are absent
F-68 Can respond to bids for attention from others
F-154 Can demonstrate understanding of property words corresponding to the objects used during familiar routines
F-142 Can pay attention to object characteristics due to language cues
F-47 Can demonstrate recognition of own name
F-149 Can demonstrate understanding of action words during familiar routines.
F-153 Can demonstrate understanding of object words during familiar routines
ELA-1489 Can demonstrate an understanding of words encountered in familiar routines
F-143 Can demonstrate understanding of object names.
ELA-1490 Can demonstrate an understanding of frequently occurring nouns
F-11 Can demonstrate understanding of action words
F-139 Can draw conclusions based on category knowledge (early inductive reasoning)
ELA-168 Uses adjectives to modify nouns in communication
F-75 Can demonstrate understanding of property words
F-121 Can identify familiar people, objects, places, and events
F-105 Names things or people in the immediate environment
F-146 Can match a real object with a picture or other symbolic representation of the object
ELA-1631 Can identify objects, events, or information that are related to personal experience.
F-174 Can identify objects associated with familiar routines
ELA-1102 Can differentiate between text and pictures or braille and tactile graphics/objects that accompany the text
ELA-1382 Can name objects in pictures/tactile graphics or name objects used to represent book pictures during a shared reading activity
ELA-757 Can label the action that occurs in a story during a shared reading activity.
ELA-1541 Can identify familiar people, places, things, and events shown in a presentation on a topic
ELA-1468 Can recognize that illustrations reflect aspects of the text
ELA-1210 Can identify pictures or tactile graphics/objects that go with a familiar text
ELA-1374 Recognize that illustrations or tactile graphics/objects in a text can provide information
ELA-19 Can demonstrate understanding of wh-questions
F-117 Uses gestures or vocalizations to ask questions
ELA-39 Can identify concrete details in a text (characters, objects)
ELA-1214 Can identify concrete details in a familiar story (characters, objects)
ELA-1722 Can identify the concrete details in a narrative based on the illustrations
F-118 Can produce single word questions.
ELA-2007 Can identify a concrete detail similar in two versions of the same story
ELA-1483 Can produce single word who and what questions
ELA-2011 Can identify a concrete detail similar in two different narratives
ELA-414 Identify the general relationship between a text and its illustrations.
ELA-1556 Can answer simple multiple choice questions asking about who and what information
ELA-765 Can identify the character and setting in a story
ELA-1212 Can identify the character and setting in a familiar story
ELA-1213 Can identify major observable events in a familiar story.
ELA-1557 Can answer simple questions asking about who and what information
ELA-1685 Can identify illustrations or object/tactual information depicting characters in a narrative
ELA-1197 Identify the beginning and end of a familiar story
ELA-1204 Can answer who and what questions about details in a familiar narrative
ELA-1723 Can identify the setting of a narrative using illustrations or object/tactual information
ELA-2013 Can identify a concrete detail different in two different narratives
ELA-1123 Can use information gathered from the text and visuals to answer explicit who and what questions
ELA-354 Can identify the characters, setting, and major events of a story
ELA-1182 Can compare characters in a familiar story.
ELA-1678 Can answer who and what questions about details in a narrative
ELA-348 Can answer questions about concrete details in a narrative
ELA-2364 Can answer questions about key details in a narrative
ELA-1558 Can answer simple wh questions
ELA-2366 Can generate questions about details in a narrative
ELA-786 Can answer wh questions about details in a narrative
ELA-2365 Can answer wh questions about key details in a narrative
ELA-1195 Can answer explicit questions to demonstrate understanding of a narrative.

Figure 9. A small section of DLM map that represents skills related to question development, ranging from initially paying attention to other people and objects to answering questions about a narrative. The color-coded nodes in the map section represent a pathway of tested nodes covering a third grade Reading Literature Essential Element focused on answering questions in narratives.

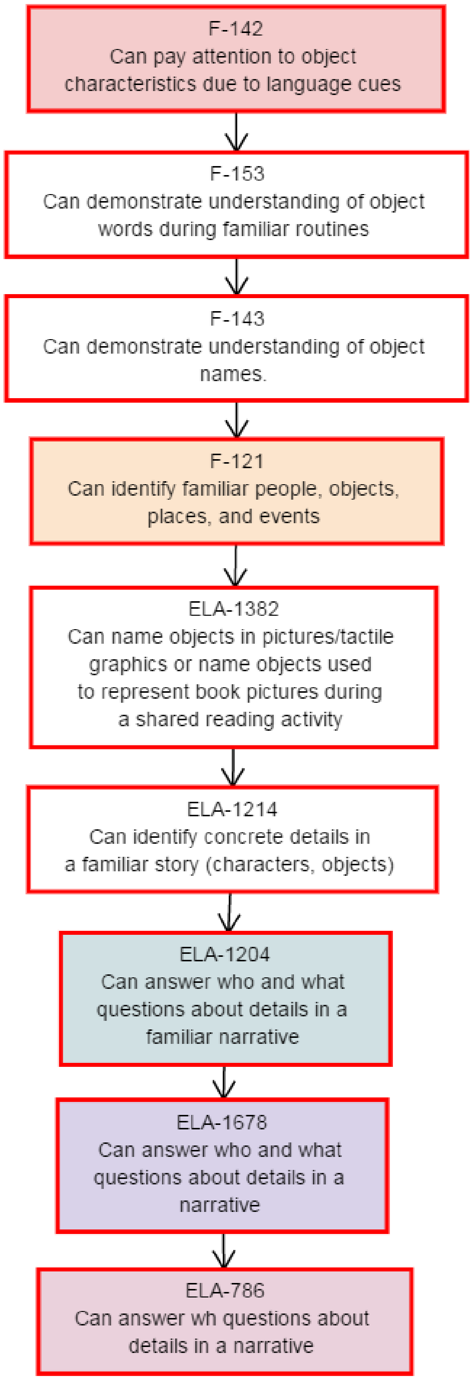

F-142
Can pay attention to object characteristics due to language cues
F-153
Can demonstrate understanding of object words during familiar routines
F-143
Can demonstrate understanding of object names.
F-121
Can identify familiar people, objects, places, and events
ELA-1382
Can name objects in pictures/tactile graphics or name objects used to represent book pictures during a shared reading activity
ELA-1214
Can identify concrete details in a familiar story (characters, objects)
ELA-1204
Can answer who and what questions about details in a familiar narrative
ELA-1678
Can answer who and what questions about details in a narrative
ELA-786
Can answer wh questions about details in a narrative

Figure 10. *The pathway of nodes covering a third grade Reading Literature Essential Element focused on answering questions in narratives as represented in the DLM map section in Figure 11. The colored nodes represent the different levels at which students are tested for this EE. The learning target for this*

EE is represented as the penultimate (purple) node. Additional information about assessment design is provided in Chapter III.

II.1.A.iv. Alternate Pathways

Creating learning targets for students with the most significant cognitive disabilities does not sufficiently provide all students access to the content. Some students with the most significant cognitive disabilities exhibit other disabilities that make it difficult for them to provide evidence of mastery for some nodes in the DLM maps. A critical step in making the DLM maps accessible to all students included the creation of alternate paths. An alternate path contains nodes and connections that are overtly modeled to account for a specific set of skills that students with learning differences must acquire en route to a learning target.

The DLM project staff, in partnership with the Center for Literacy and Disability Studies at the University of North Carolina (UNC), enhanced the maps for students with the most significant cognitive disabilities. The UNC team reviewed each node and considered whether the node was accessible to individuals with differences across four primary areas: vision, hearing, mobility, and communication (e.g., students with autism). Nodes that were flagged during this process were deemed to be probably inaccessible even when potential accommodations were considered. As an example, many of the early writing nodes involve skills like scribbling before students eventually are able to produce letters and numerals. For individuals with mobility differences, the writing acquisition process will involve learning to use assistive technology to select letters and numbers. In this example an accommodation allowing the student to select scribbles would be inappropriate. As a result, the early writing nodes related to scribbling were flagged as inaccessible since the cognitive process of learning to write involves some fundamental differences for student using assistive technology to communicate. These flagged nodes were often clustered together and represented regions within the map that posed challenges for learners with specific types of disabilities.

As an example, in Figure 11, students with mobility impairments would not learn to write through the set of nodes identified for mobility-typical learners (e.g., drawing scribbles, diagonal lines, circles) as depicted by the green nodes. Rather, learners would need to learn to select letters using an alternate system (e.g., assistive technology) as depicted by the orange nodes. This set of nodes represents the cognitive steps involved with learning to use alternative writing methods and are not necessary for students without mobility impairments. These nodes and connections are referred to as alternate paths. Most alternate paths occur early in the DLM maps and, once acquired, allow the student to achieve academic targets if provided appropriate access via assessments and instruction based on principles of universal design.

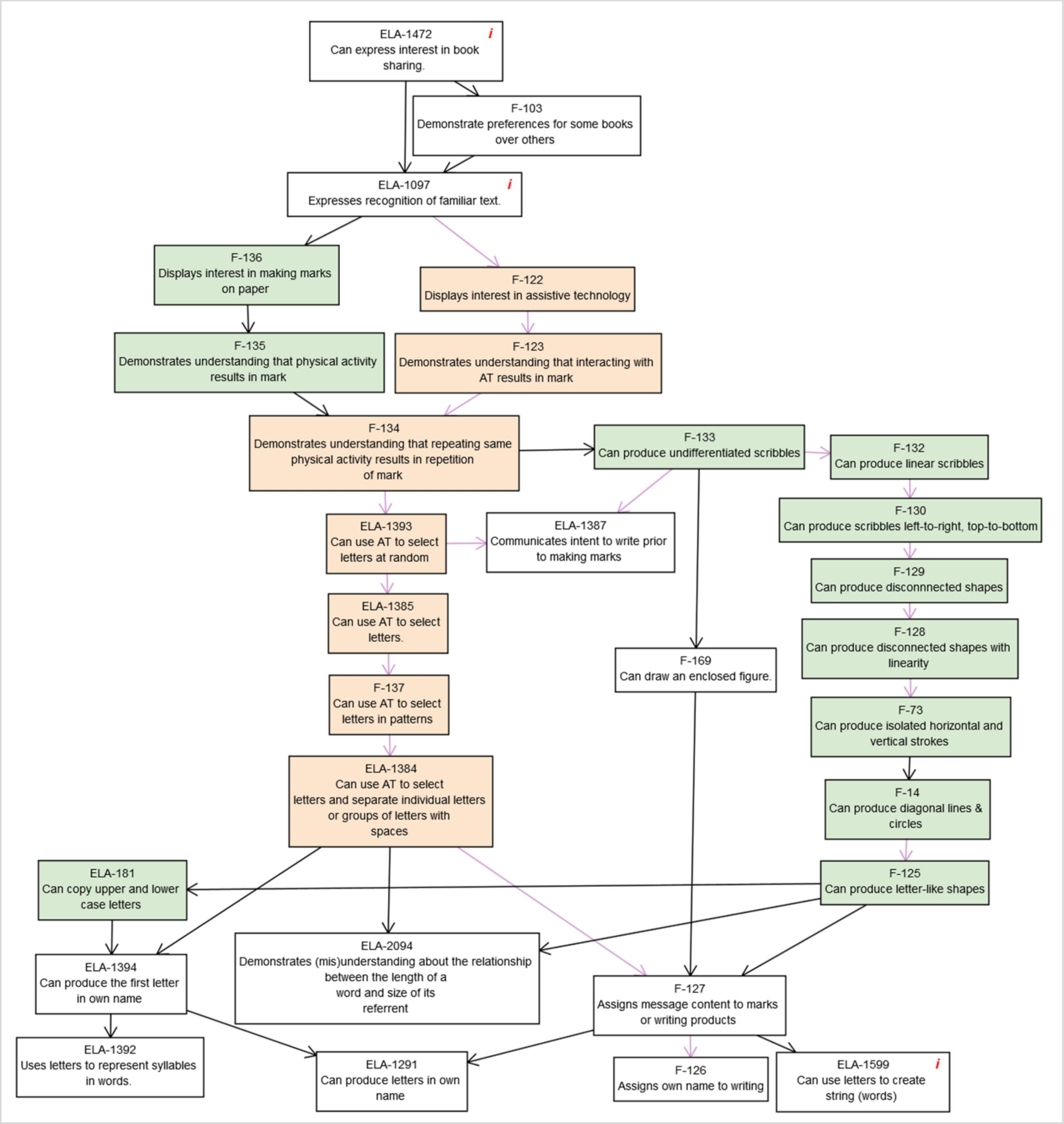

ELA-1472 i
Can express interest in book sharing.
F-103
Demonstrate preferences for some books over others
ELA-1097 i
Expresses recognition of familiar text.
F-136
Displays interest in making marks on paper
F-122
Displays interest in assistive technology
F-135
Demonstrates understanding that physical activity results in mark
F-123
Demonstrates understanding that interacting with AT results in mark
F-134
Demonstrates understanding that repeating same physical activity results in repetition of mark
F-133
Can produce undifferentiated scribbles
F-132
Can produce linear scribbles
F-130
Can produce scribbles left-to-right, top-to-bottom
ELA-1393
Can use AT to select letters at random
ELA-1387
Communicates intent to write prior to making marks
F-129
Can produce disconnnected shapes
ELA-1385
Can use AT to select letters.
F-128
Can produce disconnected shapes with linearity
F-169
Can draw an enclosed figure.
F-137
Can use AT to select letters in patterns
F-73
Can produce isolated horizontal and vertical strokes
ELA-1384
Can use AT to select letters and separate individual letters or groups of letters with spaces
F-14
Can produce diagonal lines & circles
ELA-181
Can copy upper and lower case letters
F-125
Can produce letter-like shapes
ELA-1394
Can produce the first letter in own name
ELA-2094
Demonstrates (mis)understanding about the relationship between the length of a word and size of its referrent
F-127
Assigns message content to marks or writing products
ELA-1392
Uses letters to represent syllables in words.
ELA-1291
Can produce letters in own name
F-126
Assigns own name to writing
ELA-1599 i
Can use letters to create string (words)

Figure 11. An alternative path around writing for students with a mobility impairment. The green nodes indicate the writing development for mobility-typical students, while the orange nodes suggest an alternate path students with mobility impairments can follow in writing development using assistive technology.

II.1.B. EDUCATOR AND EXPERT REVIEWS

By 2014 the DLM maps underwent three major external reviews by educators: K-5, 6-12, and special education. The purpose of the first two reviews was to leverage the expertise of general educators, identified by State Education Agency (SEA) personnel from the states included in the DLM Consortium, to examine both nodes and connections by grade level. For each node, the team was to consider: (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). Teams then reviewed individual origin-to-destination connections for appropriateness (e.g., is the connection from skill A to skill B logical?). If the educators found a node or connection they disagreed with, found illogical, or contained a gap, they stated the reasons for their disagreement and attempted to provide evidence for their reasons. When possible, the educators provided potential solutions for the problematic node or connection by suggesting how the node could be fixed or what node (new or old) should come in between the connected nodes. As they were reviewing the DLM maps, the educators were reminded that they were to focus on only the typical progression of the average student in acquiring the grade-level learning targets. Following the K-5 and 6-12 reviews, a round of internal edits were conducted to incorporate the educator feedback.

Similar to the K-5 and 6-12 reviews, a review of specific map sections was also conducted by special educators and related service providers to make the content of the map accessible to students with the most significant cognitive disabilities. Participants were experts across a range of disabilities identified by their SEA. Prior to the special educator review, collaborators from the University of North Carolina at Chapel Hill who had deep expertise in education for students with the most significant cognitive disabilities identified multiple areas in the DLM maps in which students with specific types of disabilities (e.g., vision, hearing, mobility, and communication) might have difficulty performing. To gather feedback on these potential problems, reviewers were asked to evaluate these flagged areas, and based on their expert judgment, make recommendations for pathways that would be more accessible. In some cases, universal design (UD) principles could be implemented to make the node content accessible by changing how the skill would be assessed (i.e., allowing for multiple ways to demonstrate skills). The application of these principles ensured that nodes (where possible) represented skills and understandings that were not dependent on information exclusively available through one sense. These decisions were largely guided by UD principles of flexibility of use and equitability of use. In other cases, it was clear that some students needed to acquire cognitive skills different than the general education population in order to achieve a learning target (see the writing example provided above). If alternate nodes were required, participants attempted to identify an alternate path around the problematic node(s) by describing the specific instructional method or the cognitive skills required to circumvent the node(s) and achieve the learning target. In summary, the educators in the special education review proposed edits to increase the accessibility of the content of existing nodes and connections to this student population and created alternate paths with new nodes and connections appropriate to meet the students' needs.

II.1.C. Empirical Analysis of the DLM Maps

Once the diagnostic classification models (DCM; see Chapter V) are refined and there is sufficient information to support empirical analysis of the DLM maps, several aspects of the structure of the DLM maps will be evaluated, such as

- quality of model fit;
- the uniqueness of the hypothesized nodes (i.e., are nodes distinguishable from one another); and
- directionality of relationships among nodes (i.e., does mastery of nodes go in the anticipated order or are there reversals, where a student has mastered a later node without mastering an earlier node).

As DCM results become available, the DLM staff will use a systematic approach to evaluate findings in regard to the structure of the DLM maps. The content development teams will review the DCM results and compare them to the DLM maps' structure. Based on the criteria listed above, the DLM staff will identify any potential areas that require editing and will consult the relevant research on ELA and mathematics skill acquisition/development to compare with both the DCM results and the structure of the DLM maps. Findings will be discussed with the state partners and the DLM Technical Advisory Committee. After that, DLM staff may refine any parts of the DLM maps to account for the DCM results. Empirical analyses and the DLM maps refinement are expected to continue to be a part of the ongoing work of improving the accuracy and representativeness of the DLM maps. Sufficient data are expected to be available to begin the evaluation process after the 2015-16 operational assessment.

II.1.D. DLM Maps for the 2014-15 Operational Assessment System

Table 6 includes the overall statistics describing the DLM maps as of August, 2015. This version of the interconnected set of ELA, mathematics and foundational DLM maps was the basis for the operational assessments delivered in 2014-15. Foundational nodes support both ELA and math maps.

Table 6. Number of nodes and connections in the DLM maps by node category

Node Category	Number of Nodes	Number of Connections
English Language Arts	1919	5045
Foundational	150	277
Mathematics	2399	5200
Total	4468	10522

III. ITEM AND TEST DEVELOPMENT

Chapter III provides procedural evidence as part of the overall validity argument with emphasis on support to test content and response process claims. Chapter contents include how Evidence-Centered Design (ECD) was used to develop testlets, the basic unit of test delivery for the DLM system. Further, the chapter describes how the learning map model nodes and Essential Elements (EEs) were used to develop concept maps to specify item and testlet development. By applying principles of Universal Design for Learning (UDL), the student population characteristics were factored into the entire development process, including an emphasis on accessibility and bias considerations. Chapter III includes summaries of external reviews for content, bias, and accessibility. The final portions of the chapter describe pilot test, field tests, and the final pool of operational assessments for 2014–2015.

III.1. Review of Assessment Structure

As discussed in Chapters I and II, the DLM Alternate Assessment System uses learning map models that 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 English language arts (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 and alternate pathways by which students may develop content knowledge and skills.

The DLM Alternate Assessment System uses a variant of evidence-centered design (ECD) to develop processes for item and test development. 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, Steinberg, & Almond, 1999, p. 1.). Four broad claims were developed for each content area of ELA and mathematics, which were then subdivided into nine conceptual areas, in order to work within the highly complex learning map models (Chapter I). Claims are overt statements of what students are intended to learn as a result of mastering skills within a very large neighborhood of the map. Conceptual areas are nested within claims and are comprised of multiple conceptually related content standards and nodes that support and extend beyond them. The claims and conceptual areas apply to all grades in the DLM Alternate Assessment System.

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

For each EE, linkage levels—small collections of nodes which represent critical junctures on the path toward and beyond the learning target—were identified in the map. A linkage level is a location of a node or nodes in the map where an assessment was developed for that particular EE.

The EEs specify academic targets, while the map clarifies how students can reach those targets. Assessment items were developed based on nodes at five linkage levels. 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 standard. Nodes are also identified beyond the standard, in order 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 (IP)
2. Distal Precursor (DP)
3. Proximal Precursor (PP)
4. Target (T)
5. Successor (S)

The nodes and their relationships are described in mini-maps that item writers used during test development (see Chapter I for a discussion of the relationship of the system elements).

III.1.A. Development of the Essential Elements

The DLM EEs are alternate or extended content standards that link to college and career readiness standards. The development of the EEs began in February 2011, when initial planning meetings were held between DLM project staff; Edvantia, Inc., a DLM subcontractor; state partners; and state educational agency content experts. These meetings were held to ensure that state partners were in agreement with the process designed by Edvantia and the goals of the EEs. Throughout the process of developing the EEs, staff and stakeholders were encouraged to ensure that the content of the EEs increased in complexity from grade to grade. This approach was key to ensuring that the EEs represented the highest possible expectations for students with significant cognitive disabilities (SWSCD).

During development of the EEs, important emphasis was placed on ensuring that the expectations reflected increasing academic rigor across grades. An example of three related EEs from the “Key Ideas and Details” strand is shown in Table 7. 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.

Table 7. Example of Increasing Complexity in Related EEs across Grades

Grade Level EE	RI.3.2	RI.7.2	RI.9–10.2
EE Descriptions	Identify details in a text	Determine two or more central ideas in a text	Determine the central idea of the text and select details to support it

Table 8 shows the increasing complexity from linkage level to linkage level for the same EEs shown in Table 7. This example provides an illustration of how complexity increases both across linkage levels and across grade levels.

Table 8. Example of Increasing Complexity of Skills in Related Linkage Levels for Three EEs Across Grades

Linkage Level	RI.3.2	RI.7.2	RI.9–10.2
Initial Precursor	Can correctly look at the scene demonstrating a possible event and ignore the scene demonstrating an impossible event based on an understanding that objects still exist despite not being seen	Can pair an object with a picture, tactile graphic, or other symbolic representation of the object	Can identify the concrete details, such as individuals, events, or ideas in familiar informational texts
Distal Precursor	Can pay attention to either the entire object, a characteristic of the object, or an action in which the object can perform after some verbal label has been attached to it	Can identify the concrete details mentioned in informational texts	Can identify the details in an informational text that relate to the topic of the text based on their similarities

Linkage Level	RI.3.2	RI.7.2	RI.9–10.2
Proximal Precursor	Can identify illustrations or tactile graphics/objects that reflect aspects of a familiar text, such as setting, characters, or action if it is a story or a person, place, thing, or idea if it is an informational text	Can identify the main idea for a paragraph in an informational text that lacks an explicit statement of the topic	Can summarize the information in a familiar informational text
Target	Can identify the concrete details mentioned in beginner level informational texts	Can determine more than one main idea in an informational text	Can pick out the details that are relevant and contribute to the understanding of the central idea of an informational text
Successor	Can identify explicit details in an informational text	Can summarize the information in a familiar informational text	Can support the identification of the implicit and explicit meaning of an informational text using specific details and citations

Development of the EEs began in 2011. Stakeholder meetings were held via webinar in March 2011 to prepare materials for development meetings. State partners recruited content experts and educators of students with significant cognitive disabilities to serve as panelists on the committees that drafted the EEs. A series of content-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. Face-to-face meetings were attended by DLM project staff, Edvantia, Inc. staff, and SEA and LEA representatives, in addition to the content and special education experts who served on the panels.

Led by Edvantia, Inc., representatives from each of the then thirteen DLM partner state education agencies 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 member states. A similar meeting was held to draft the mathematics EEs in May 2011, with more than 70 participants representing 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.

Concurrent with the development of the DLM EEs, the DLM Consortium was actively engaged in building learning map models in mathematics and ELA, as described in Chapter II. The DLM maps are highly connected representations of how academic skills are acquired, built through a research synthesis process. In the case of the DLM project, the Common Core State Standards helped to specify academic targets, while the surrounding map content clarified how students could reach the specified standard. Learning map models of this size had not been previously developed, and as a result, alignment between the DLM EEs and the maps was not possible until the fall of 2012, when an initial draft of the maps was available for review.

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 Common Core State Standards 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 growth 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 partner states in early 2013, with an approved final version of the EEs published in May 2013.[9] Final documents for ELA and mathematics are available publically at http://dynamiclearningmaps.org/content/essential-elements.

III.1.B. TEST BLUEPRINTS

The DLM test blueprints specify the pool of available EEs and requirements for coverage within each conceptual area. The precise test experience could vary across students within the boundaries of required coverage.

Blueprint development began with a proposed plan in October 2013 and was discussed extensively through September 2014, after which state partners finalized those blueprints for the

[9] Each state chose whether to formally adopt the EEs as alternate or extended content standards for students with the most significant cognitive disabilities.

2014–15 assessment year. Content teams in each content domain developed blueprint options following several guiding principles. Member state representatives and content experts then reviewed multiple iterations of blueprints, as did the senior DLM staff and psychometricians.

III.1.B.i. Guiding Principles

DLM partner states identified three overarching needs for blueprints. First, the blueprint in each content area should have broad coverage of academic content as described by the EEs. This emphasis maintains the connection to grade-level content standards for SWSCDs and ensures that there is appropriate breadth of content coverage within the domain. Second, the blueprints in both content areas should emphasize connections in skills and understanding from grade to grade. The third need was to limit the administration burden of assessing SWSCDs. The learning map models developed by DLM project staff were used to prioritize EEs for inclusion in the blueprint in each content area. 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 growth 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 content area,
- 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 SWSCDs.

In both content areas, some EEs were not included on the blueprint. Some reasons for excluding EEs from the blueprint were:

- 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 RL.3.7, "Use information gained from visual elements and words in the text to answer explicit *who* and *what* questions."[10]) and,

[10] In this case, a different EE in the same grade, describing a similar construct, 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.

- 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. It is important to recognize that these principles were not implemented as rules, but as guidelines for prioritization of the content of the EEs within and across the grades.

III.1.B.ii. Blueprint Development Process

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

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 DLM partner states and Technical Advisory Committee (TAC) members in early 2014. In order to ensure coverage of content across conceptual areas, there is a required minimum number of EEs to be assessed in certain conceptual areas at each grade level. States have the flexibility to require or recommend higher numbers of EEs covered during the school year.

III.1.B.ii.a English Language Arts

After seeking input and consent from state partners, content in the areas of Claim 1 (reading) and Claim 2 (writing) was prioritized for inclusion in the ELA blueprint. In addition to a variety of reading testlets at each grade level, all students complete structured writing assessments in

[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.

which a test administrator engages the student 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.

Table 9 shows the number of EEs included in the ELA blueprint by grade level compared to the total number of EEs in each conceptual area. As grade level increases, more EEs are located in more cognitively complex conceptual areas.

Table 9. Number of EEs in the ELA YE Blueprint/Total Number of EEs per Conceptual Area

ELA Conceptual Areas (CA)						
Grade	C1.1 Deter-mine critical elements of text	C1.2 Construct understandings of text	C1.3 Integrate ideas and information from text	C2.1 Use writing to communicate	C2.2 Integrate ideas and Information in writing	Total
3	7/12	5/9	2/2	2/12		16/35
4	7/10	6/9	1/5	3/9	0/1	17/34
5	3/6	8/10	4/8	2/7	0/1	17/32
6	1/3	10/13	3/9	2/8	0/3	16/36
7	1/3	8/12	4/10	5/9	0/4	18/38
8	0/3	9/12	3/10	5/11	0/4	17/40
9	0/2	9/11	3/11	3/9	2/6	17/39
10	0/2	9/11	3/11	3/9	2/6	17/39
11	0/2	8/11	4/11	4/9	2/7	17/40

*Note: * "7/12" indicates the blueprint contains 7 of 12 EEs in a grade and CA combination. Empty cells represent grades with no EEs assigned to the CA.*

The *DLM English Language Arts Year-End Assessment Model Blueprint* (2014) is available on the DLM website.

III.1.B.ii.b Mathematics

Like ELA, the breadth of mathematics EEs available for assessment was deliberately broad. In each grade, the approved blueprint addresses all four claims and each conceptual area relevant to the grade. All but a few EEs are 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 power for future mathematical learning during and beyond school, and
- acknowledge mathematical learning trajectories that connect the EEs over the course of several grades.

Table 10 shows the number of EEs by grade and conceptual areas included in the blueprint for grades 3-8. Note that not all grades have EEs in all nine conceptual areas.

Table 10. Number of EEs in the YE Blueprint for Grades 3–11 and Total Number of EEs per Conceptual Area

Grade	**Mathematics Conceptual Areas**									**Total**
	C1.1	**C1.2**	**C1.3**	**C2.1**	**C2.2**	**C3.1**	**C3.2**	**C4.1**	**C4.2**	
3	3/4*		1/1	0/1	1/1	2/3	1/1	2/2	1/1	11/14
4	2/2	2/2	1/1	¾	1/2	3/5	1/2	2/3	1/1	16/22
5	2/2	3/4	2/2	2/2	1/1	3/3	1/1		1/1	15/16
6	1/1	2/2	2/2		2/2		1/2	3/3		11/12
7	2/2	1/1	3/3	3/4	1/2		2/3	1/2	1/1	14/18
8	1/1	1/2	2/2	4/4	1/1		1/1	1/1	3/5	14/17
9			3/6	2/4	1/1	0/1	0/3	2/4	0/7	8/26
10			1/6	1/4	0/1	1/1	2/3	2/4	2/7	9/26
11			2/6	1/4	0/1	0/1	1/3	0/4	5/7	9/26

*Note: * "3/4" indicates the blueprint contains 3 of 4 EEs in a grade and CA combination. Empty cells represent grades with no EEs assigned to the CA. All 26 EEs in grades 9, 10 and 11 are considered together in a grade band.*

The high school EEs are defined for the high school grade band (grades 9–12) as a whole. In the blueprint, mathematics high school EEs are organized by grade level: Math 9, Math 10, and Math 11. All of the EEs except two are each assigned to one of the three grade-level blueprints.

The *DLM Mathematics Year-End Assessment Model Blueprint* (2014) is available on the DLM website. Item maps showing the number of items for each EE in ELA and mathematics are provided in Appendices C.10 and C.11.

III.1.C. ITEMS AND TESTLETS

Testlets are the basic units of the DLM Alternate Assessment System. These testlets are short, instructionally relevant measures of student skills and understandings and contain an engagement activity that includes a stimulus related to the assessment designed to help the student focus on the task at hand followed by three to nine items. ELA reading testlets also contain a story or informational text. Each testlet includes items from one or more EEs in the blueprint. By completing all the testlets assigned in the spring window, students cover all the EEs in the blueprint.

III.1.C.i. Overview of the Testlet Development Process

Every testlet went through multiple rounds of review by DLM staff, internal content and accessibility specialists, editors, and educators in DLM states who served as external reviewers. The full set of test development steps are outlined below.

1. Item writer is trained.
2. Item writer is assigned testlet specification and Essential Element Content Map (EECM) with other supporting materials.
3. Item writer develops a draft testlet and associated metadata.
4. Content team completes first internal quality control review.
5. Testlet receives first editorial review. Where applicable, graphics needed for engagement activities and items are inserted.
6. Content and accessibility specialists complete internal quality control review.
7. Content team completes second internal quality control review.
8. Testlet is entered into the content management system in KITE.
9. Testlet receives second editorial review.
10. Content team completes third internal quality control review.
11. External reviewers review testlet for content, accessibility, and bias and sensitivity.
12. Synthetic read-aloud tagging is applied to the testlet.
13. Test production team completes first quality control review.
14. Testlet is prepared for delivery in KITE.
15. Testlet receives testing window delivery quality control checks by test production, content, and psychometric teams for accessibility, display, content, and associated test delivery resources.
16. The testlet is delivered for field testing.
17. Field test data is reviewed by psychometric and content teams.
18. Testlets and items that do not require revision are made operational.

Each review group was carefully trained to look for potential problems with the academic content, accessibility issues, and concerns about bias or sensitive topics. After testlets were externally reviewed, they were scheduled for field-testing. DLM staff reviewed results from field tests to determine which testlets met quality standards and were ready for operational assessment. Security of materials was maintained through the test development process. Paper materials were kept in locked facilities. Electronic transfers were made on a secure network drive or within the secure content management system in KITE.

III.1.C.ii. General Testlet Structure and Item Types

In reading and mathematics, testlets are based on nodes for one linkage level of one or more of EEs. Writing testlets cover multiple EEs and linkage levels. Each testlet contains an engagement activity and three to nine items. All testlets begin with a non-scored engagement activity.

Several item types are used in DLM testlets. Most types are used in both ELA and math testlets. Some types are used only in testlets for one content area. The following item types are used in DLM testlets:

- Multiple choice single select (MCSS)
- Multiple choice multiple select (MCMS)
- Select text (ELA only)
- Matching lines (mathematics only)
- Drag-and-drop (mathematics only)

Most items within the testlets have three answer options presented in a multiple-choice format using either text or images. Technology-enhanced items 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 in order to avoid having to use many multiple choice items to assessment same construct. Evidence for the accessibility and utility of technology-enhanced items was collected from item tryouts and cognitive labs. See Chapter IX for a description of item tryouts and cognitive labs.

There are two general modes for DLM testlet delivery: computer-delivered and teacher-administered (see Chapter IV). 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. 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 IV.

Teacher-administered testlets are designed for educator to administer outside the system, with the test administrator recording responses in the system rather than the student recording his or her own responses. These teacher-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 directions screen(s), teacher-administered testlets include instructions for the engagement activity. After the engagement activity, items are presented. All teacher-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 teacher-administered) and content being assessed (reading, writing, or mathematics). Descriptions of engagement activities and items are found in this section for ELA reading, writing, and mathematics testlets. 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.

III.1.C.iii. English Language Arts Reading Testlets

ELA reading testlets were built around texts adapted from or related to grade-level appropriate general education texts. Short narrative passages were constructed from books commonly taught in general education, and short informational texts were written to relate to thematic elements from narratives. All passages were deliberately written to provide an opportunity to assess specific nodes in the maps associated with different EEs and linkage levels. Text complexity for passages was reduced from the grade level texts for students without significant cognitive disabilities, focusing on core vocabulary, simple sentence structure, and readability.

Above all, texts were written with an emphasis on readability. ELA Claim 1 states, "Students can comprehend text in increasingly complex ways." To provide access to a wide range of student needs, the surface complexity of the text was held relatively constant, but the complexity of cognitive tasks needed to answer items was increased. Texts are generally very brief and allow for paired readings, that is, two readings by the student, without posing an undue burden on test administration. Texts are presented with 1–3 sentences on a screen with an accompanying photograph. One screen is presented at a time. Students and educators can navigate forward and backward between screens. ELA passages contain between 6 and 25 screens. Texts are between 50 and 200 words in length.

ELA reading testlets follow a basic structure, with variations for some teacher-administered testlets or testlets assessing nodes that require students to compare more than one text. Figure 12 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 teacher-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 passage is complete and that next, students will re-read the passage and respond to some questions. After these directions, the student begins the second reading. The second reading is presented in exactly the same format as the first reading, with items embedded as appropriate. Embedded items are placed between the screens from the text.

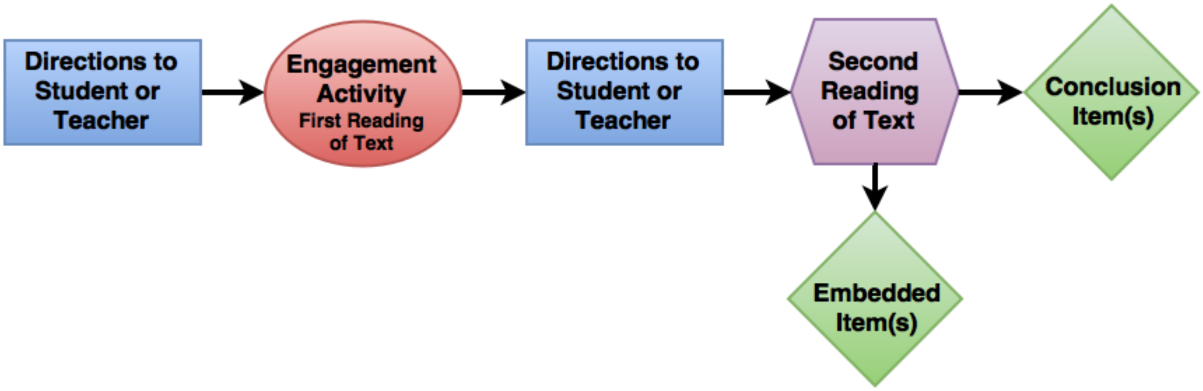


Figure 12. Elements of an ELA reading testlet.

The decision to use paired readings of the same passage 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 of the curriculum 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, which 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 as appropriate into the second reading between screens that present the text. Items that are associated with a node that describes a cognitive *process* related to the conceptual area and EE are generally embedded in the text during the second reading. That way, the item will be able to measure information in the current working memory of the reader. 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.

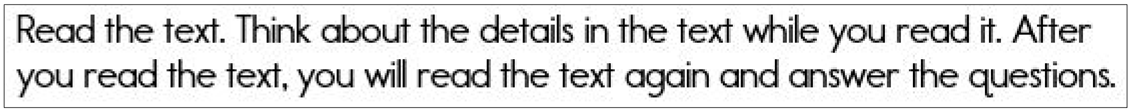
The use of embedded items means that rather than having students read a story once and then recall how a character felt at some prior point in the story, the embedded question is presented when the character's feeling state is active in working memory.

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.

III.1.C.iii.a Engagement Activities

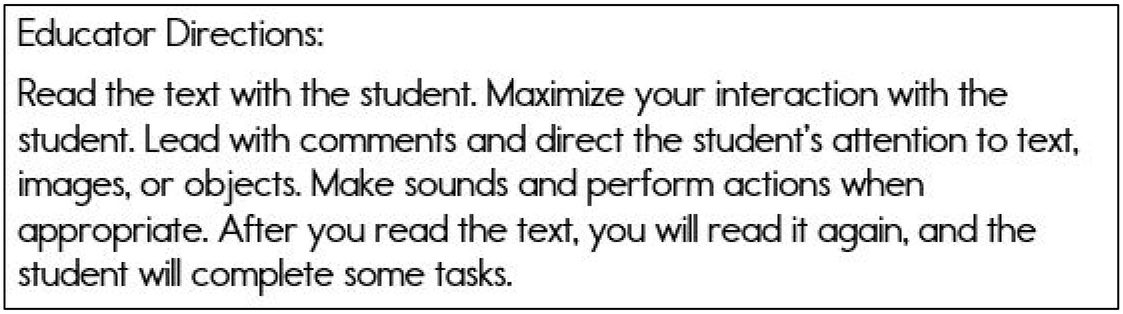
ELA reading testlets include an engagement activity, which outlines the structure of the testlet and instructs the student and/or test administrator how to proceed through the testlet. In reading testlets, the first reading of the text is considered a part of the engagement activity. In computer-delivered testlets, the engagement activity instructs students to read the text on their own or with read-aloud support as a selected accessibility support (see Chapter IV). In teacher-administered testlets, the engagement activity introduces the testlet to the test administrator, who will read the story or text with the student. An example of a computer-delivered engagement activity screen is shown in Figure 13.



Read the text. Think about the details in the text while you read it. After you read the text, you will read the text again and answer the questions.

Figure 13. Example ELA Computer-delivered Reading Engagement Activity.

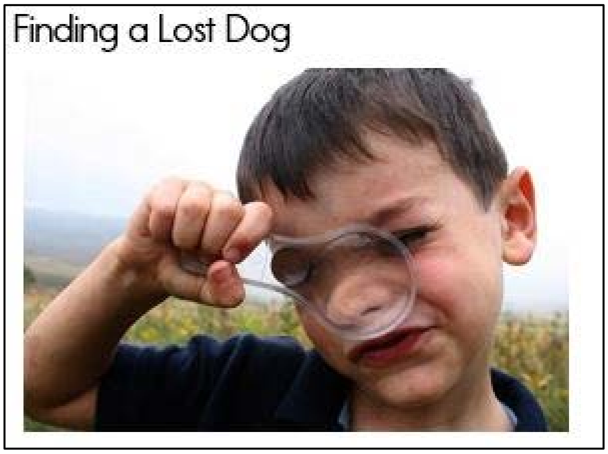
Teacher-administered testlets require the test administrator to assess the student outside the KITE system and enter responses. For ELA reading teacher-administered testlets, the engagement activity is also the first reading of the text. In this case, the directions for the engagement activity are presented to the test administrator. An example of a screen included in a teacher-administered engagement activity screen is shown in Figure 14. The first screen contains directions written for the test administrator. The second screen is the first page of the text that is used in the testlet (Figure 15).



Educator Directions:

Read the text with the student. Maximize your interaction with the student. Lead with comments and direct the student's attention to text, images, or objects. Make sounds and perform actions when appropriate. After you read the text, you will read it again, and the student will complete some tasks.

Figure 14. *Example ELA Teacher-administered Reading Engagement Activity.*



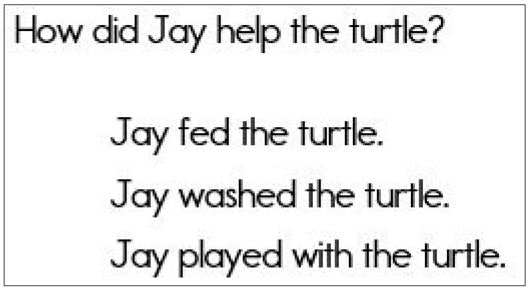
Finding a Lost Dog

Figure 15. *Example ELA first text page.*

III.1.C.iii.b Items

Computer-delivered ELA reading testlets contain three item types: multiple choice, multiple-select multiple choice, and select text. Items of all three types can be embedded items, which occur throughout the second reading of the text used in the testlet, or conclusion items, which occur at the end of the second reading. Teacher-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 16.



How did Jay help the turtle?

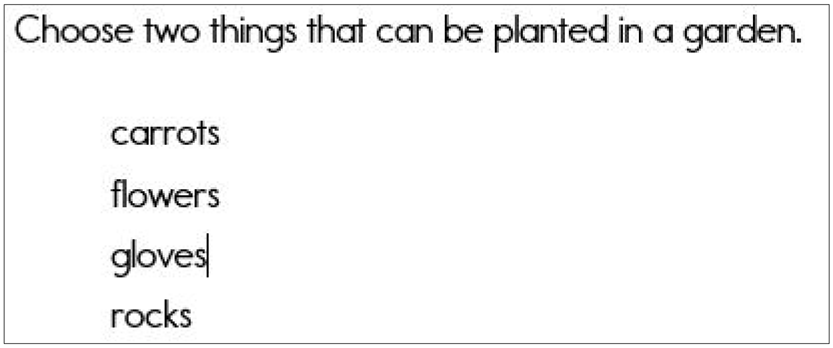
Jay fed the turtle.

Jay washed the turtle.

Jay played with the turtle.

Figure 16. Example ELA Computer-delivered Multiple-Choice Item.

For multiple-select multiple-choice 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-select multiple-choice items allow students to choose up to four answer options. An example of an ELA multiple-select multiple-choice item is shown in Figure 17.



Choose two things that can be planted in a garden.

carrots

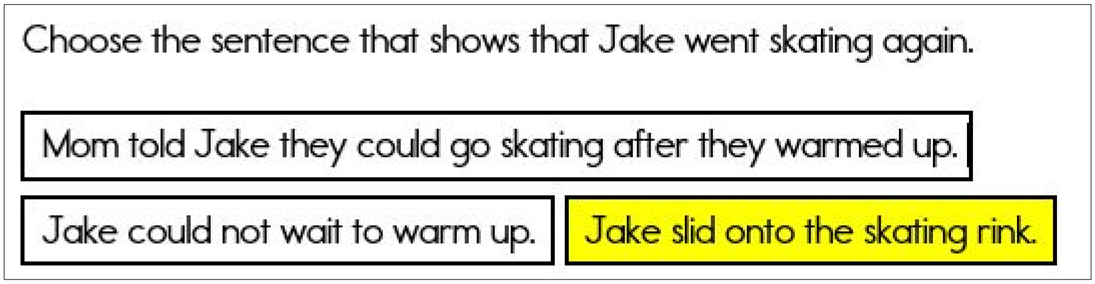
flowers

gloves

rocks

Figure 17. Example ELA Computer-delivered Multiple-Select Multiple-Choice Item.

Select-text items direct students to select an answer from a passage taken from the story or text. In Figure 18, 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.



Choose the sentence that shows that Jake went skating again.

Mom told Jake they could go skating after they warmed up.

Jake could not wait to warm up.

Jake slid onto the skating rink.

Figure 18. Example ELA Select-Text Item.

III.1.C.iv. English Language Arts Writing Testlets

Writing testlets cover multiple EEs. All ELA writing testlets are teacher-administered. For writing testlets, the test administrator engages in a scripted activity with a student outside the KITE system and then enters observations and ratings of the student's writing process and product into KITE. Figure 19 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.

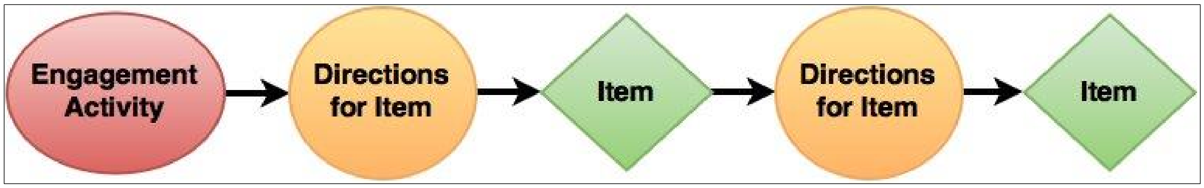
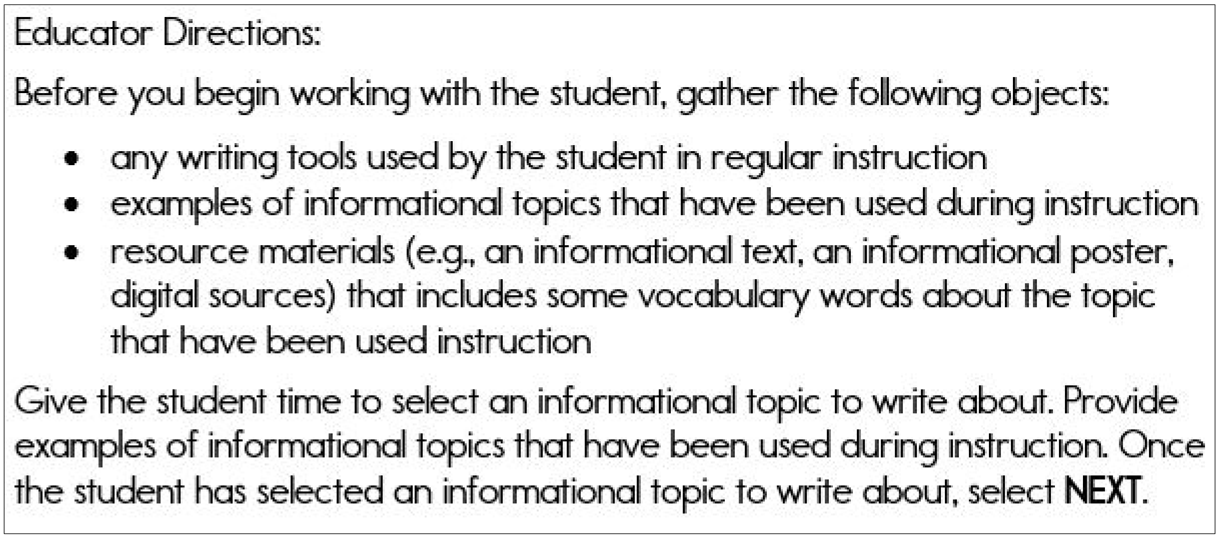


Figure 19. Elements of an ELA writing testlet.

Every grade level has an Emergent and Conventional writing testlet, each comprised of several EEs. Emergent writing describes the marks, scribbles, and random selection of letters seen in beginning writers (Erickson, Hatch & Clendon, 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 are comprised of letters, that words have meanings, and that written words can be put together in order to communicate to others. Key behaviors associated with conventional writing include writing letters and words through the use of a traditional writing tool or alternate pencil.

III.1.C.iv.a 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 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 20 shows an example.



Educator Directions:

Before you begin working with the student, gather the following objects:

- any writing tools used by the student in regular instruction
- examples of informational topics that have been used during instruction
- resource materials (e.g., an informational text, an informational poster, digital sources) that includes some vocabulary words about the topic that have been used instruction

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**.

Figure 20. Example ELA Teacher-administered Writing Engagement Activity.

III.1.C.iv.b Items

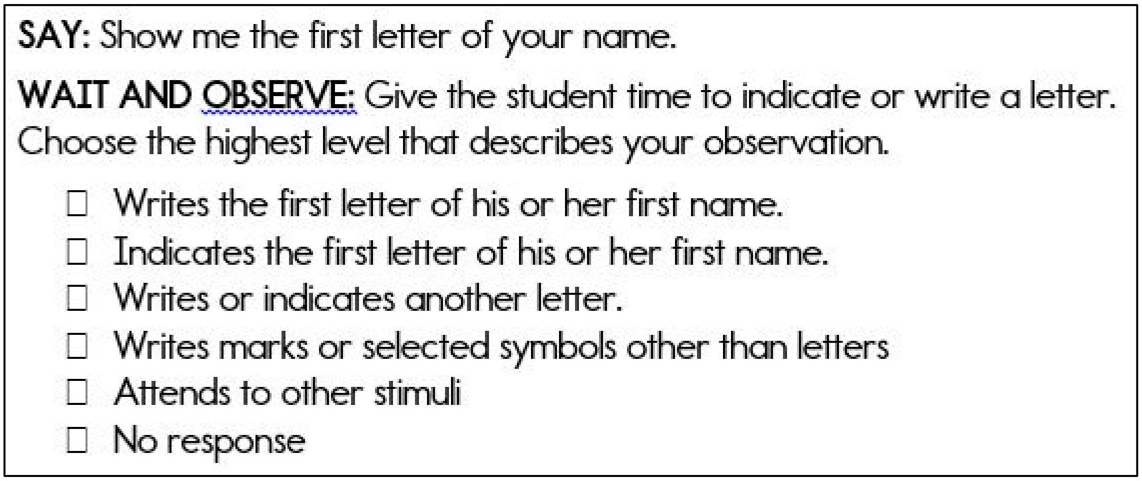
In writing testlets, the engagement activity is followed by items that require the test administrator to evaluate the student's writing process. Some writing testlets also evaluate the student's writing product, and these product items occur at the end of the testlet. Process and product items are MCSS or MCMS items with answer choices that are judgments made by the test administrator. 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 ratings of the test administrator's observations of the student as he or she completes items in the testlet. Figure 21 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 21). The inclusion of multiple correct answer options was designed to ensure that this testlet was accessible to emergent writers who were beginning to write letters and emergent writers who had not yet developed writing production skills but were still able to identify the first letter of their name.

Items that assess writing products are the test administrator's ratings of the product created by the student as a result of the writing processes completed in the administration of the testlet.

Figure 22 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.

Writing testlets are constructed to provide test administrators with a coherent structure for delivering an instructionally-relevant writing tasks 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 includes 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. Since 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 linkage levels and different EEs. For example, in Figure 21, 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.

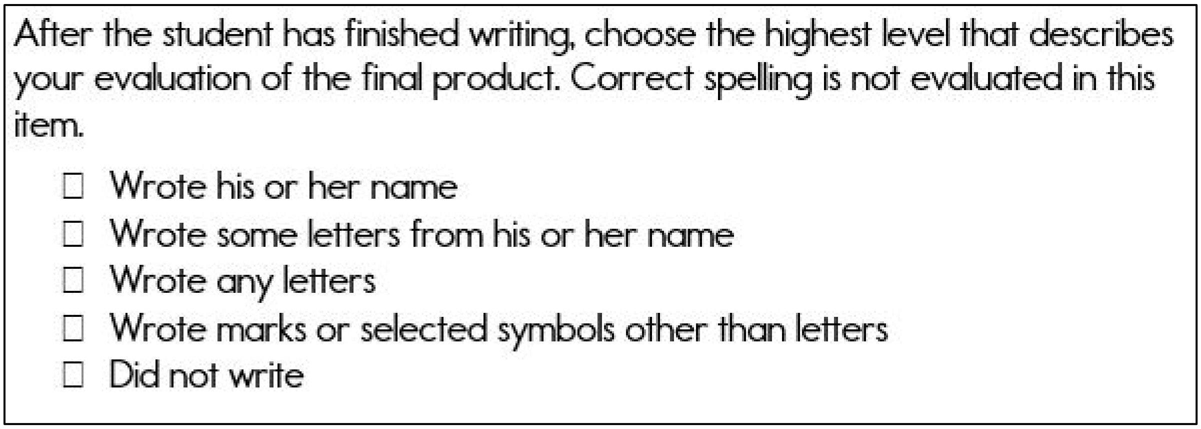


SAY: Show me the first letter of your name.

WAIT AND OBSERVE: Give the student time to indicate or write a letter. 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 selected symbols other than letters
- □ Attends to other stimuli
- □ No response

Figure 21. Example ELA Emergent Writing Item Focused on Process.



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 his or her name
- □ Wrote some letters from his or her name
- □ Wrote any letters
- □ Wrote marks or selected symbols other than letters
- □ Did not write

Figure 22. *Example ELA Conventional Writing Item Focused on Product.*

III.1.C.v. 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 23 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 EEs and linkage levels.

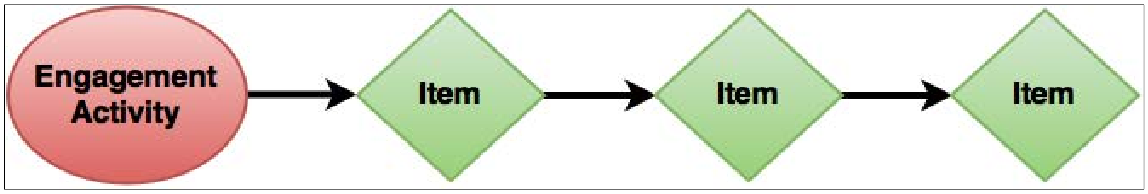
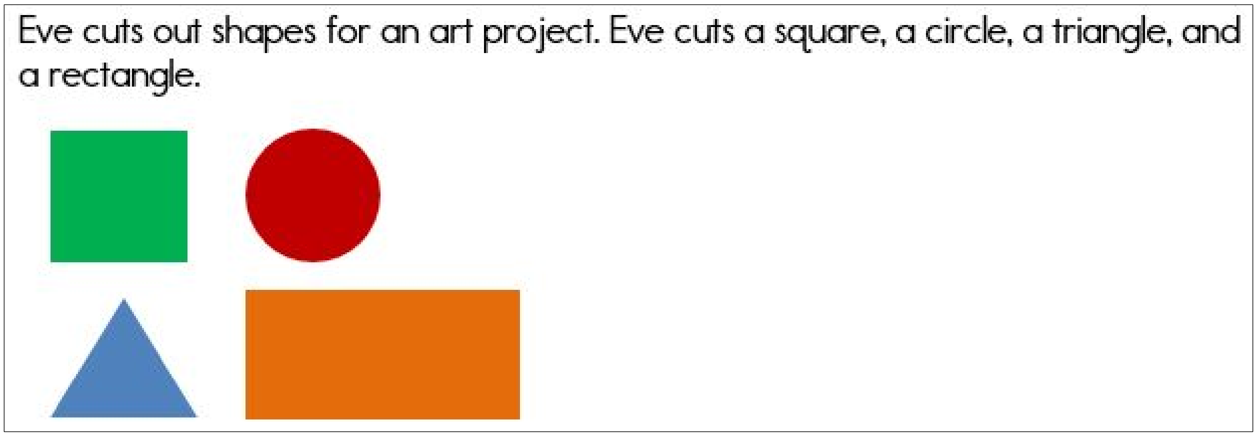


Figure 23. Elements of a mathematics testlet.

Following the engagement activity, three to eight items are presented to the student. The number of items varies based on blueprint and test specifications. Teacher-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.

III.1.C.v.a Engagement Activities

Mathematics testlets start with an engagement activity that provides a context for the questions. Mathematics testlets are built around a common scenario activity to investigate related facets of student understanding of the targeted content. The mathematics engagement activity in Figure 24 provides a context related to shapes and activates a cognitive process about putting things together. This example was 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.



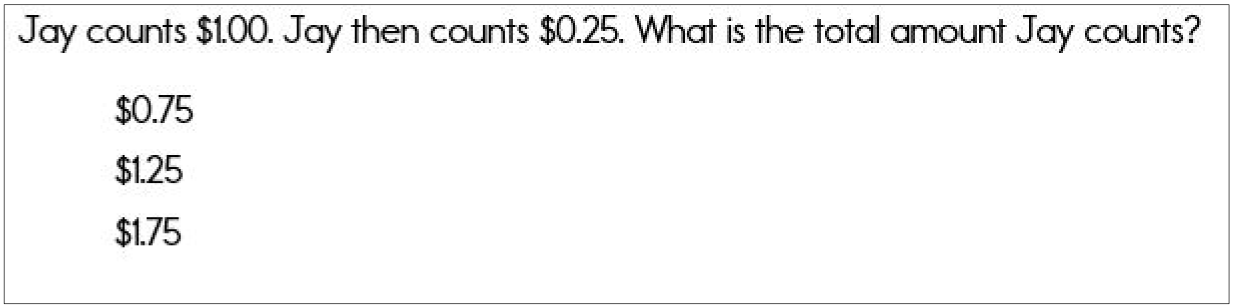
Eve cuts out shapes for an art project. Eve cuts a square, a circle, a triangle, and a rectangle.

Figure 24. Example Mathematics Engagement Activity.

III.1.C.v.b Items

Computer-delivered mathematics testlets contain four item types: multiple choice, multiple-select multiple-choice, matching, and drag-and-drop. Technology-enhanced items such as multiple select, matching, and drag-and-drop were 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. Items that require students to sort multiple objects were better assessed by using a drag-and-drop item where the structure task onscreen was representative of the cognitive process being assessed. Teacher-administered mathematics testlets used 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 25. An example multiple-choice mathematics item using pictures as response options is shown in Figure 26.



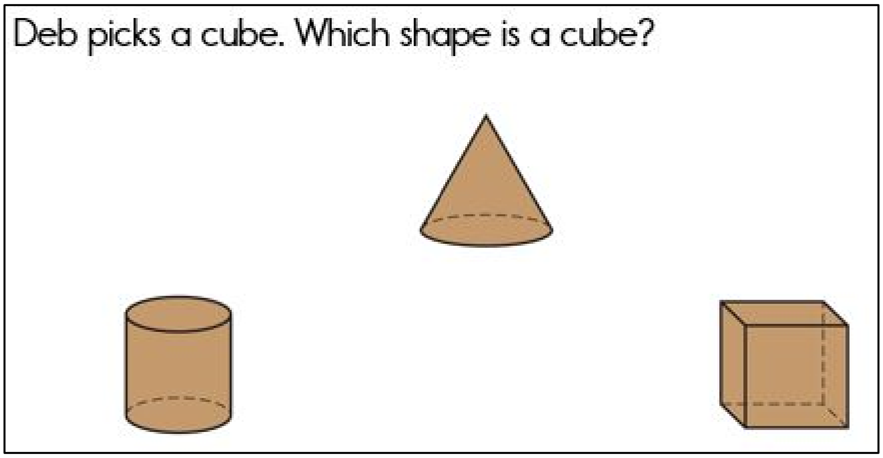
Jay counts $1.00. Jay then counts $0.25. What is the total amount Jay counts?

$0.75

$1.25

$1.75

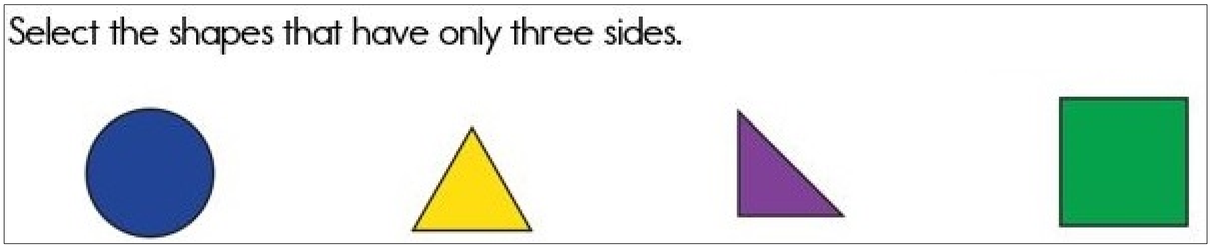
Figure 25. Example Mathematics Multiple-Choice Item with Text.



Deb picks a cube. Which shape is a cube?

Figure 26. Example Mathematics Multiple-Choice Item with Pictures.

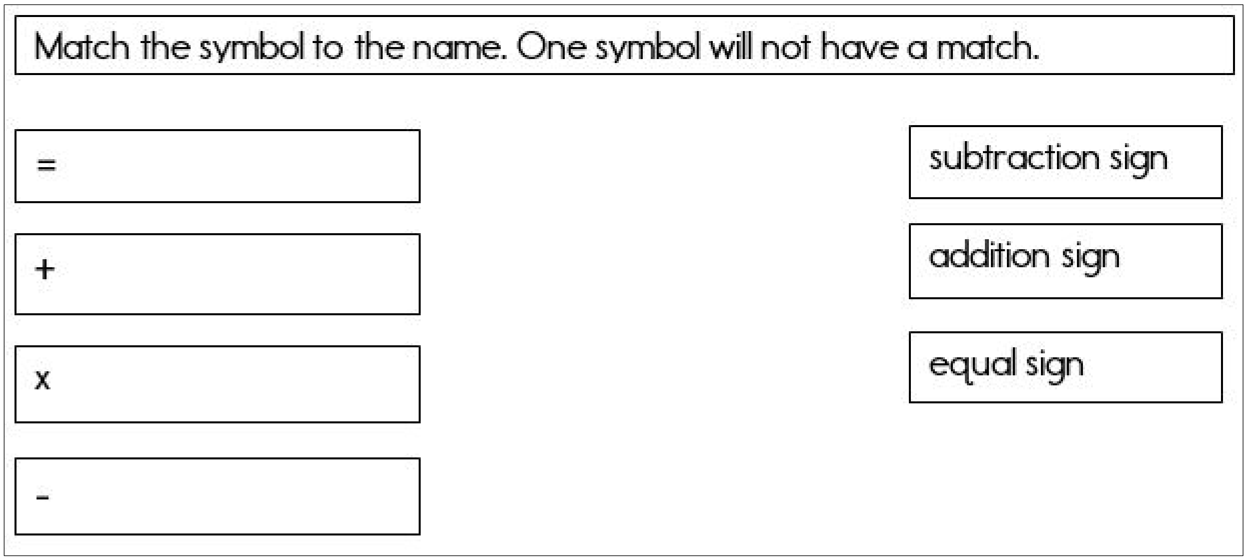
Multiple-select multiple-choice items provide the student with the opportunity to make more than one answer choice. An example of a multiple-select multiple-choice item is shown in Figure 27.



Select the shapes that have only three sides.

Figure 27. Example Mathematics Multiple-Select Multiple-Choice 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 28. 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.



Match the symbol to the name. One symbol will not have a match.

=

+

x

-

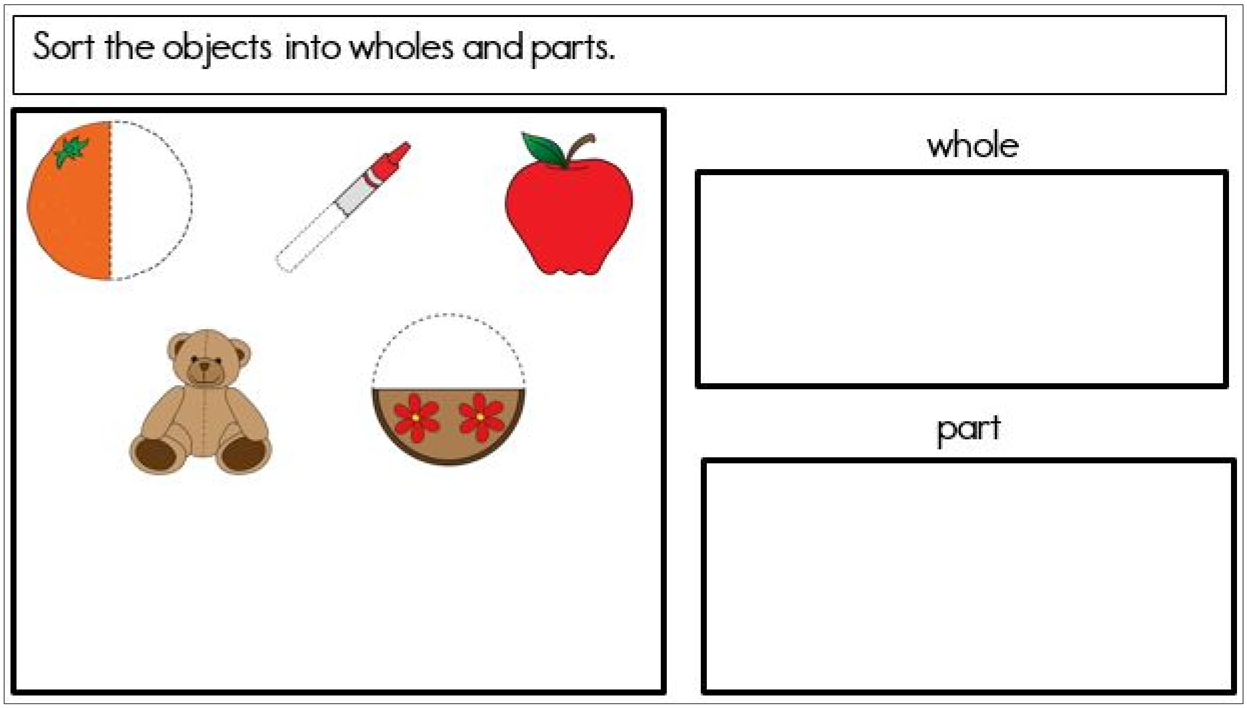
subtraction sign

addition sign

equal sign

Figure 28. *Example Mathematics Matching Item.*

Students also encounter questions asking them to sort objects into categories. An example of a drag-and-drop item, as it would look before the student responded, is shown in Figure 29. In this item, the student clicks and holds on an object from the left side box and drags the object to one of the two boxes on the right side. When the student releases the object, it stays in the right side box where it was placed.



Sort the objects into wholes and parts.

Figure 29. Example Mathematics Drag-and-Drop Item.

III.1.C.vi. Alternate Testlets for Students who are Blind or Have Visual Impairments

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

1. Alternate testlets, called BVI forms, were created when nodes were difficult to assess online for students who had visual impairments, even with features such as read aloud or magnification. 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. In teacher-administered BVI testlets, test administrator 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 teacher-delivered) are provided on the Testlet Information Page (see Chapter IV).
2. Braille forms were available for grades 3-5 at the Target and Successor levels and in grades 6-HS at the Proximal Precursor, Target, and Successor levels. Braille was intentionally limited to these grades and linkage levels as alternate forms. Braille forms were provided when sighted students were 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 were at levels where students were 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. Since general versions of testlets at those EEs and levels did not require reading, braille was not provided at those levels.

III.1.D. Essential Element Concept Maps for Testlet Development

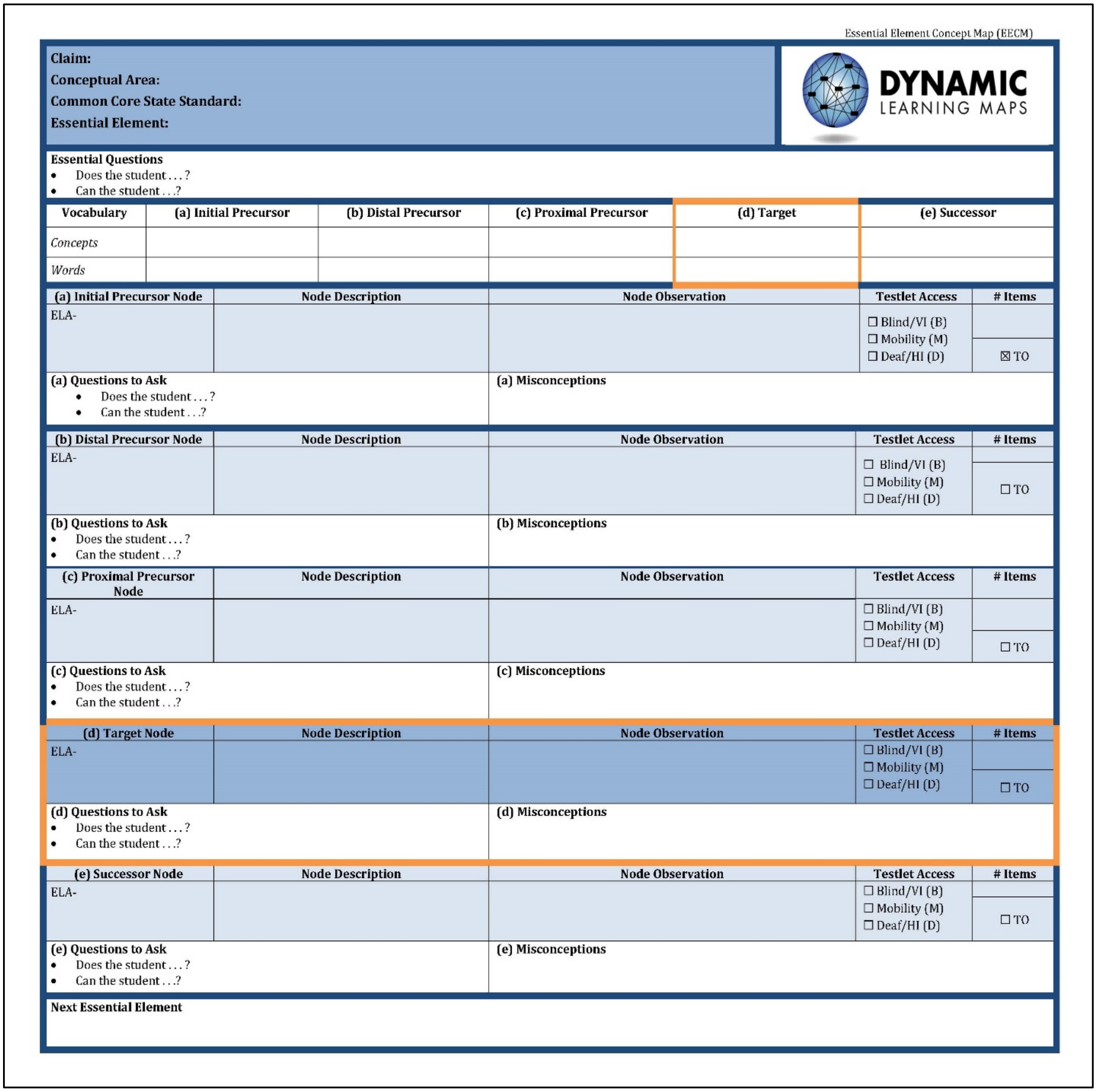
Evidence-centered design (ECD) describes a conceptual framework for designing, developing, and administering educational assessments (Miselvy, Steinberg & Almond, 1999). The use of an ECD framework in developing large-scale assessments supports arguments for validity of the interpretations and uses of the assessment results. ECD requires test designers to make 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 explicit. 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, Seeratan, Cameto, Haertel, Knokey, & Morrison, K., 2011; Flowers, Turner, Herrera, Towles-Reeves, Thurlow, Davidson, & Hagge, 2015).

Item and testlet writing was based on Essential Element Concept Maps (EECMs). These graphic organizers used principles of ECD to define ELA and mathematics content specifications for assessment. ELA and mathematics content teams developed the EECMs. Developers selected nodes from the learning map models to be assessed at different linkage levels based on an analysis of the map structure. Staff with student population expertise also reviewed EECMs. Item writers use the EECM, which is 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 and identifies key concepts and vocabulary at each level. It 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 templates were developed and adopted by states in the DLM Alternate Assessment Consortium and utilized in the development of assessments for ELA and mathematics (Bechard and Sheinker, 2012). The templates were specifically designed for clarity and ease of use, as the project engaged non-professional item writers from participating consortium states who needed to create a large number of items in a constricted timeframe. An example of a blank EECM is shown in Figure 19. An example EECM that was used for item development is included in Appendix B.1.

The EECM has seven functions:

- Identify the targeted standard by claim, conceptual area, CCSS, 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 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.



Essential Element Concept Map (EECM)

Claim:
Conceptual Area:
Common Core State Standard:
Essential Element:

Essential Questions

- Does the student . . . ?
- Can the student . . .?

Vocabulary	(a) Initial Precursor	(b) Distal Precursor	(c) Proximal Precursor	(d) Target	(e) Successor
Concepts					
Words					

(a) Initial Precursor Node	Node Description	Node Observation	Testlet Access	# Items
ELA-			☐ Blind/VI (B) ☐ Mobility (M) ☐ Deaf/HI (D)	☒ TO

(a) Questions to Ask

- Does the student . . . ?
- Can the student . . .?

(a) Misconceptions

(b) Distal Precursor Node	Node Description	Node Observation	Testlet Access	# Items
ELA-			☐ Blind/VI (B) ☐ Mobility (M) ☐ Deaf/HI (D)	☐ TO

(b) Questions to Ask

- Does the student . . . ?
- Can the student . . .?

(b) Misconceptions

(c) Proximal Precursor Node	Node Description	Node Observation	Testlet Access	# Items
ELA-			☐ Blind/VI (B) ☐ Mobility (M) ☐ Deaf/HI (D)	☐ TO

(c) Questions to Ask

- Does the student . . . ?
- Can the student . . .?

(c) Misconceptions

(d) Target Node	Node Description	Node Observation	Testlet Access	# Items
ELA-			☐ Blind/VI (B) ☐ Mobility (M) ☐ Deaf/HI (D)	☐ TO

(d) Questions to Ask

- Does the student . . . ?
- Can the student . . .?

(d) Misconceptions

(e) Successor Node	Node Description	Node Observation	Testlet Access	# Items
ELA-			☐ Blind/VI (B) ☐ Mobility (M) ☐ Deaf/HI (D)	☐ TO

(e) Questions to Ask

- Does the student . . . ?
- Can the student . . .?

(e) Misconceptions

Next Essential Element

Figure 30. Example Essential Element Concept Map (EECM) Graphic Organizer.

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 31 shows an example of a mathematics mini-map.

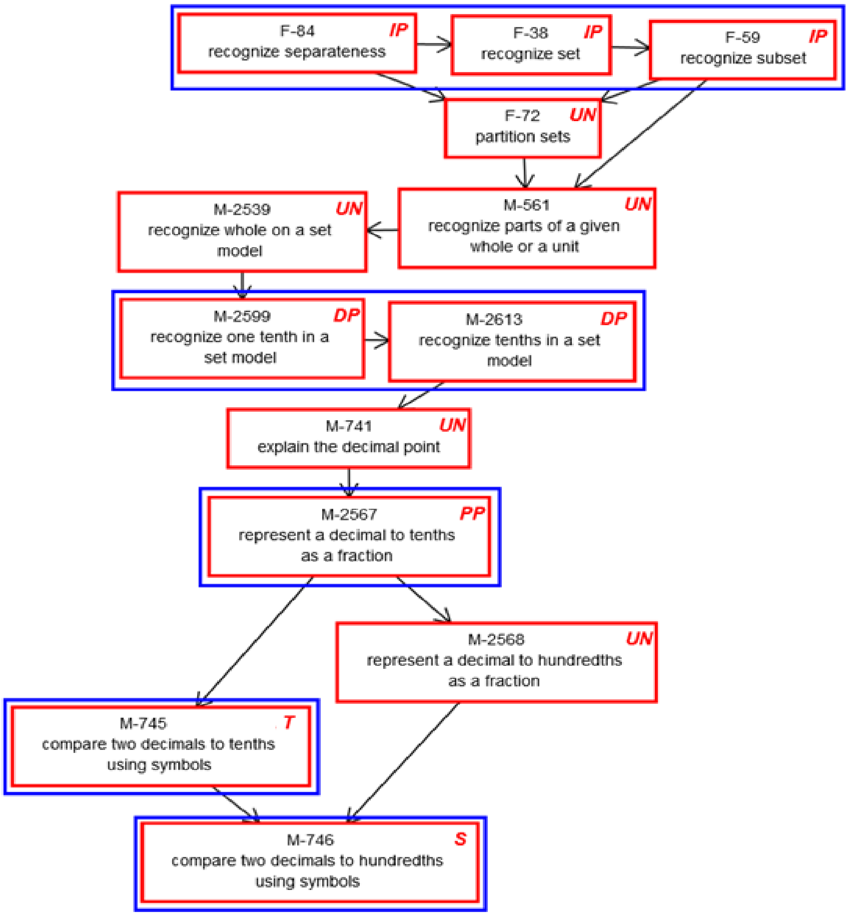

F-84 IP
recognize separateness
F-38 IP
recognize set
F-59 IP
recognize subset
F-72 UN
partition sets
M-2539 UN
recognize whole on a set model
M-561 UN
recognize parts of a given whole or a unit
M-2599 DP
recognize one tenth in a set model
M-2613 DP
recognize tenths in a set model
M-741 UN
explain the decimal point
M-2567 PP
represent a decimal to tenths as a fraction
M-2568 UN
represent a decimal to hundredths as a fraction
M-745 T
compare two decimals to tenths using symbols
M-746 S
compare two decimals to hundredths using symbols

Figure 31. Example Mathematics Mini-map for 7.NS.3—"Compare quantities represented as decimals in real world problems to tenths."

III.1.E. ITEM WRITING

DLM items and testlets were developed beginning in the summer of 2013. Additional items and testlets were developed during 2014. Most item writing occurred during summer events in which content and special education specialists worked on-site in Lawrence, Kansas, to develop DLM assessments. In addition to item writers, DLM staff and graduate research assistants supported item-writing efforts by developing supporting resources and EECMs, serving as internal reviewers, and in some cases, writing testlets.

III.1.E.i. Recruitment and Selection

The item writer recruitment and selection process secured qualified and experienced individuals to write high-quality testlets. ELA and mathematics content teams used several recruitment strategies to solicit applicants. An electronic recruitment survey was sent to state partners to be distributed to mathematics, ELA, and special education educators in DLM member states. This recruitment survey included a brief description of the job and inquired about skills and availability. Additionally, the job description was posted in several area newspapers and online sources, sent to school districts within a 50-mile radius of Lawrence, Kansas, and was sent to DLM state partners for distribution. Content teams screened applicant materials, conducted interviews, and made hiring offers to selected candidates. Applicants were evaluated on the following required qualifications: experience with ELA or mathematics academic content, knowledge of how individuals develop and learn, attention to detail in written work, time management skills, self-direction with assigned work tasks, flexibility and willingness to adapt to redirection, excellent oral and written communication in English, proficiency with basic technology skills, and ability to commute daily to a work site in the Lawrence area. The preferred qualifications included teaching experience in ELA or mathematics, experience working with or instructing students with significant cognitive disabilities, and experience with or knowledge of large-scale assessments, item development, state testing, and/or state standards. The hired applicants comprised a balance of expertise in mathematics, ELA, and special education. All item writers signed security agreements and were trained on item security procedures.

III.1.E.ii. Item Writer Characteristics

At the 2013 item-writing event, there were 53 mathematics item writers and 55 ELA item writers. At the 2014 item-writing event, there were 15 mathematics item writers and 17 ELA item writers. In 2014, of the 32 item writers, 23 had participated previously in 2013. In addition to the item writers from both the 2013 and 2014 events, 14 graduate research assistants that work on the DLM project have written testlets.

An item writer survey was used to collect demographic information about the educators and other professionals who were hired to write DLM assessments during the 2013 and 2014 summer item-writing events. This survey was used before item-writing events held in Lawrence, Kansas, during the summers of 2013 and 2014. In total, 117 item writers responded to the item writer surveys across both years. There were 58 respondents from Math and 59 respondents from ELA, which represents 97.5% of all items writers across both years. Data gathered through this survey included years of teaching experience, grades taught, degree type, experience with the population, experience with alternate assessment based on alternate achievement standards (AA-AAS), and whether the item writer currently taught students eligible for AA-AAS. Additionally, item writers were asked whether they were National Board Certified educators. Each survey category is described below, with an accompanying table when applicable. Data were aggregated across both years. The median and range of number of years of teaching experience for ELA and mathematics item writers is shown in Table 11.

Table 11. Item Writers' Years of Teaching Experience

	ELA		**Mathematics**	
	Median	**Range**	***Median***	**Range**
PreK–12	11	0–32	9.5	0–37
ELA	9	0–31	6	0–34
Mathematics	7	0–31	9	0–35
Special Education	5.5	0–32	6	0–37

The distribution of grade levels taught by item writers was similar in both content areas. Nineteen item writers with high-school teaching experience participated on each content team. There were 36 ELA item writers with experience at the elementary level, grades 3–5, and 30 with experience in middle school, grades 6–8. There were 31 mathematics item writers with experience at the elementary level, grades 3–5, and 37 with experience with middle school, grades 6–8. See Table 12 for a summary.

Table 12. Item Writers' Grade Level Teaching Experience

	ELA		**Mathematics**	
	n	**%**	***n***	**%**
Grade 3	14	23.73	12	20.69
Grade 4	11	18.64	11	18.97
Grade 5	11	18.64	8	13.79
Grade 6	6	10.17	17	29.31
Grade 7	12	20.34	10	17.24
Grade 8	12	20.34	10	17.24
High School	19	32.20	19	32.76

Note: Multiple grades could be selected on the survey of item writers. Percentages do not equal 100%.

The 117 item writers represented a highly qualified group of professionals in the education and assessment field. Over 90% of the item writers on both content teams—92% of ELA item writers and 91% of mathematics item writers—held a bachelor's degree. Master's level degrees were held by 59% of the ELA item writers and 71% of the mathematics item writers. Among all item writers, 14% held doctoral degrees. The number and types of degrees held by item writers are shown in Table 13.

Table 13. Item Writers' Level of Degree

	ELA Item Writers		**Mathematics Item Writers**	
	n	%	*n*	%
Bachelor's	54	91.53	54	93.10
Master's	35	59.32	41	70.69
Other	10	16.95	7	12.07

Most item writers had experience working with students with disabilities. The highest levels of experience occurred in the Emotional Disability, Mild Cognitive Disability, and Specific Learning Disability categories. The disability categories of Blind/Low Vision and Traumatic Brain Injury had the fewest number of responses. All disability categories reported on the survey are listed in Table 14.

Table 14. Item Writers' Population Experience

Disability Category	**ELA Item Writers**		**Mathematics Item Writers**	
	n	%	*n*	%
Blind/Low Vision	16	27.12	16	27.59
Deaf/Hard of Hearing	18	30.51	16	27.59
Emotional Disability	41	69.49	36	62.07
Mild Cognitive Disability	38	64.41	37	63.79
Multiple Disabilities	30	50.85	31	53.45
Orthopedic Impairment	18	30.51	18	31.03
Other Health Impairment	36	61.02	33	56.9

Disability Category	ELA Item Writers		Mathematics Item Writers	
	n	%	*n*	%
Severe Cognitive Disability	18	30.51	25	43.10
Specific Learning Disability	42	71.19	38	65.52
Speech Impairment	30	50.85	25	43.10
Traumatic Brain Injury	13	22.03	15	25.86
None of the above	6	10.17	8	13.79

Note: Multiple categories could be selected on the survey of item writers. Percentages do not equal 100%.

Of the item writers, 42% had experience administering an Alternate Assessment of Alternate Achievement Standards (AA-AAS) prior to their work on the DLM project, with 48%, or 57 out of 117, reporting that at the time of the survey, they worked with students eligible for AA-AAS. Twenty-nine item writers held a National Board certification. Fifteen of these National Board Certified educators were ELA item writers, 14 were mathematics item writers.

III.1.E.iii. Item Writer Training

Training for item writers consisted of multi-day sessions at the beginning of the 2013 and 2014 summer item-writing events. There were five full days of training preceding the 2013 item-writing event, and three days of training preceding the 2014 item-writing event. The redesigned, condensed training in 2014 (see Appendix B.2 for example slides) was the result of the test development team's evaluation of lessons learned in 2013. Additional changes to the 2014 training included increasing feedback opportunities during training, additional hands-on training activities, and the use of revised formatting and style guides as part of training activities. Additionally, many processes for test development were streamlined between the item-writing events, requiring less training for item writers in 2014.

After training on confidentiality and signing security agreements (see Appendix B.3), item writers were introduced to the DLM system and given time to complete DLM professional development modules before beginning specific training on the writing process. The purpose of using the modules for training was to ensure a common level of knowledge about DLM and the student population before beginning to write items. Modules focused on assessment system design, population of students, accessibility, and specific information related to either ELA or mathematics. There was a brief quiz at the end of each module that item writers were required to pass with 80% accuracy.

Training was divided into sections that focused on accessibility, content development, use of images and graphics, bias and sensitivity, use of a cognitive process dimension taxonomy, and appropriate assignment of item metadata for the content management system in KITE.

The ELA and mathematics content teams, DLM test development staff internal reviewers, and editors were all involved in monitoring and retraining item writers to ensure the quality of the testlets produced. Retraining opportunities were held for item writers during item-writing events when content teams and editors identified patterns of errors or problems with content, accessibility, or bias and sensitivity. Editors held periodic retraining sessions with item writers to review the most common errors and solutions for resolving them. Editors rated the first three testlets each item writer wrote and provided specific, individualized feedback during individual and group retraining sessions. The content teams led retraining sessions with item writers as needed, providing examples, visuals, and additional documentation. Specifically, the ELA content team generated a list of common vocabulary for ELA item writers to use when developing testlets at linkage levels that contain foundational nodes. The mathematics team developed a document that listed common questions and answers for item writers to refer to when writing items and testlets. Internal reviewers also provided feedback (e.g., vocabulary too complex) for targeted retraining.

III.1.E.iv. Item Writing Resource Materials

Item writers used the EECMs to develop testlets at different linkage levels for each EE. In addition to the EECMs, item writers used material developed by content teams to support the development of testlets. All item writers used the DLM Core Vocabulary list. A 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, spanning grades K–12, that reflects the research in core vocabulary in Augmentative and Alternative Communication (AAC) and words needed to successfully communicate in academic settings where the EEs are being taught (Dennis, Erickson & Hatch, 2013).

Additionally, all item writers used a guide to good practices in item writing, which included a checklist of common item writing challenges and errors. Both content teams prepared additional materials to support item writing, including materials specially prepared to support writing items for testlets designed for students who were blind or had visual impairments. The ELA content team used guides to passages and question writing to assist item writers in designing testlets. The mathematics content team prepared a list of mathematics vocabulary and definitions to support item writers. In both subjects, prototypes of testlets were used during training and available for item writer review. These prototypes went through multiple rounds of input from state partners and other stakeholders, internal content reviews, and editorial reviews. Prototypes were written at all five linkage levels in both subjects and included examples of teacher-administered and computer-administered testlets.

III.1.E.v. Item Writing Process

Item writers were given writing assignments for EEs, including all linkage levels outlined on the EECM. DLM assessments are built as testlets. Each testlet is associated with a linkage level. Because testlets were conceived as a short, coherent, instructionally relevant assessments, item writers produced entire testlets rather than stand-alone items. Item writers frequently wrote testlets for the same EE at different linkage levels to encourage item writers to use the DLM map relationships in the EECM to think about the content of testlets at different linkage levels.

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 complexity, from less complex to more complex linkage levels. Using the EECMs, item writers selected specific vocabulary for each testlet that matched the cognitive complexity of the node being assessed.

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 and answer options for items in testlets.

Item writers focused on all of the students who might receive each testlet and considered any accessibility issues. The goal for the item writer was to create testlets that were accessible to the greatest number of students possible, and to be specific about the conditions necessary to achieve that. Writers were prompted to ask questions such as: "Are there accessibility tools (online or offline) that may be necessary for some students?" They 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. Then, item writers focused on access to the testlet, asking, "Is this testlet designed for a particular group of students who will need a specific approach due to their disability?"

During item development, item writers and DLM staff maintained the security of materials. Item writers all signed security agreements. Training about best practices to maintain test security was provided to item writers and staff. Materials were stored in locked facilities. Electronic transfers were made on secured network drives and within the secure content management system in KITE.

III.1.E.vi. Item Writer Evaluations

An evaluation survey of the item-writing experience was sent to all participating item writers after the summer 2013 item-writing event. Item writers were asked to provide feedback on the perceived effectiveness of training and the overall experience in the summer item-writing event, as well as narrative comments on their experience and suggestions for future DLM item-writing events. Ninety-seven of the 108 item writers that participated in the summer 2013 item-writing event responded.

Of the 97 respondents across both ELA and mathematics, 25 felt training activities were very effective, 63 felt the first week of training was somewhat effective, and nine felt the training activities were not at all effective. Fifty-eight of the 97 stated the second week of extensive guided practice and peer review activities were very effective, and 36 felt the activities were effective, with only 3 responding that the activities were not at all effective. Brainstorming with colleagues was perceived as very effective by 90 of the 97 item writers that responded. Table 15 and Table 16 detail responses to the perceived effectiveness questions from the survey from ELA and mathematics item writers, respectively.

Table 15. Perceived Effectiveness of Training, English Language Arts Item Writers (n = 47)

	Very Effective		Somewhat Effective		Not At All Effective	
	n	%	*n*	%	*n*	%
Initial training week	13	27.6	31	65.9	3	6.3
Second week—extensive guided practice and peer review	29	61.7	18	38.2	0	0.0
Contents of the peer review checklist	20	42.5	25	53.1	2	4.2
Contents of the content/special education review checklist	25	53.1	20	42.5	2	4.2
Brainstorming with colleagues	42	89.3	5	10.6	0	0.0
Feedback from DLM staff	33	70.2	11	23.4	3	6.3
Resource materials (printed and on flash drive)	32	68.0	14	29.7	1	2.1

Table 16. Perceived Effectiveness of Training, Mathematics Item Writers (n = 50)

	Very Effective		Somewhat Effective		Not At All Effective	
	n	%	*n*	%	*n*	%
Initial training week	12	24.0	32	64.0	6	12.0
Second week—extensive guided practice and peer review	29	48.0	18	36.0	3	6.0
Contents of the peer review checklist	23	46.0	26	52.0	1	2.0

	Very Effective		Somewhat Effective		Not At All Effective	
	n	%	*n*	%	*n*	%
Contents of the content/special education review checklist	17	34.0	28	56.0	5	10.0
Brainstorming with colleagues	48	96.0	1	2.0	1	2.0
Feedback from DLM staff *	23	46.0	20	40.0	6	12.0
Resource materials (printed and on flash drive)	31	62.0	18	36.0	1	2.0

*Note: * Only 49 respondents*

Overwhelmingly, responding item writers agreed or strongly agreed that the DLM item-writing process was a valuable professional development experience (95 out of 96, or 99%). Almost all respondents (97%) agreed or strongly agreed that the project goals were clear and that they were developing good assessments for students with the most significant cognitive disabilities (98%). They were equally as confident that the testlets they wrote would be perceived as instructionally relevant (98%).

Table 17 and Table 18 detail responses to the overall experience questions from the survey from ELA and mathematics item writers, respectively.

Table 17. Overall Experience, English Language Arts Item Writers (n = 47)

	Strongly Agree		Agree		Disagree		Strongly Disagree	
	n	%	*n*	%	*n*	%	*n*	%
The overall goals of the summer item-writing project were clear.	25	53.1	20	42.5	2	4.2	0	0.0
I had enough time to complete my testlets each week.	38	82.6	8	17.3	0	0.0	0	0.0
My content leaders were knowledgeable about academic content.	25	53.1	20	42.5	2	4.2	0	0.0

	Strongly Agree		**Agree**		**Disagree**		**Strongly Disagree**	
	n	%	*n*	%	*n*	%	*n*	%
My room leaders were knowledgeable about testlet development procedures.	27	57.4	17	36.1	3	6.3	0	0.0
The content of the EECMs (questions, misconceptions, observations) guided my decisions regarding testlet creation.	20	42.5	23	48.9	3	6.3	1	2.1
I am confident that the testlets I produced will be good assessments for students with significant cognitive disabilities.	20	42.5	27	57.4	0	0.0	0	0.0
Other educators would find the testlets I wrote to be instructionally relevant.	20	42.5	26	55.3	1	2.1	0	0.0

Table 18. Overall Experience, Mathematics Item Writers (n = 50)

	Strongly Agree		**Agree**		**Disagree**		**Strongly Disagree**	
	n	%	*n*	%	*n*	%	*n*	%
The overall goals of the summer item-writing project were clear.	25	50.0	24	48.0	1	2.0	0	0.0
I had enough time to complete my testlets each week.	38	76.0	12	24.0	0	0.0	0	0.0

	Strongly Agree		Agree		Disagree		Strongly Disagree	
	n	%	*n*	%	*n*	%	*n*	%
My content leaders were knowledgeable about academic content.	34	68.0	15	30.6	1	2.0	0	0.0
My room leaders were knowledgeable about testlet development procedures.	34	68.0	16	32.0	0	0.0	0	0.0
The content of the EECMs (questions, misconceptions, observations) guided my decisions regarding testlet creation.	21	42.0	27	54.0	2	4.0	0	0.0
I am confident that the testlets I produced will be good assessments for students with significant cognitive disabilities. *	22	44.0	25	50.0	2	4.0	0	0.0
Other educators would find the testlets I wrote to be instructionally relevant.	28	56.0	21	42.0	1	2.0	0	0.0

*Note: * Only 49 respondents*

Item writers were asked three open-ended response questions.

- What else helped you write high-quality assessments this summer?
- What recommendations do you have for training future DLM item writers?
- What other comments would you like to share about the experience as a DLM item writer?

In general, comments from item writers were positive and constructive. Item writers felt that the small-group work and interaction with peers was helpful during the writing and review process. They believed that previous experience with alternate assessment and/or the population, and assistance from knowledgeable DLM staff was helpful during the item-writing process.

Some recommendations for future training included increased feedback opportunities, hands-on activities throughout the week of training, and a formatting guide to be provided prior to item development.

Overall, item writers were pleased with the training received, the writing process, and their accomplishments throughout the summer session. They expressed a genuine appreciation of the knowledge gained through the item-writing event and the opportunities to collaborate with peers.

III.1.F. External Reviews

The purpose of external review is to evaluate items and testlets developed for the DLM Alternate Assessment System. Using specific criteria established for DLM assessments, reviewers decided whether to recommend that the content be accepted, revised, or rejected. Feedback from external reviewers was used to make final decisions about assessment items before they were 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 KITE. Educators nominated by consortium governance partners, and several governance partners 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. Minor modifications were made to each of these as a result of the pilot event. Once the online external review capability was available in KITE, six educators tried out the online system. They used the online training and external review manual to guide their work as they evaluated testlets in the KITE system. DLM staff observed and provided assistance if the educator had difficulty with the platform or the rating process. The external review manual was revised to address those difficulties prior to finalizing the materials for the 2013-14 external review.

III.1.F.i. Overview of Review Process

External reviews occurred after the initial internal reviews, which involved a comprehensive editorial review and an internal content review by individuals with content expertise and/or experience with students with the most significant cognitive disabilities. Figure 32 shows the order and relationship of reviews in the DLM test development process. Based on these initial reviews, DLM staff made final decisions, revised as needed, and performed a final editing review. Each testlet was then sent for external review. External reviews were conducted online, independently, and asynchronously through an application in the secure content management system in KITE. The descriptions in this chapter include external review of DLM testlets in 2013–2014 and in 2014–2015 as content development for operational delivery in the spring of 2015 was ongoing.

Resulting ratings were compiled with ratings from other reviewers and submitted to DLM staff, and DLM staff made final decisions regarding whether the testlet should be rejected, accepted as is, or revised before pilot/field-testing.

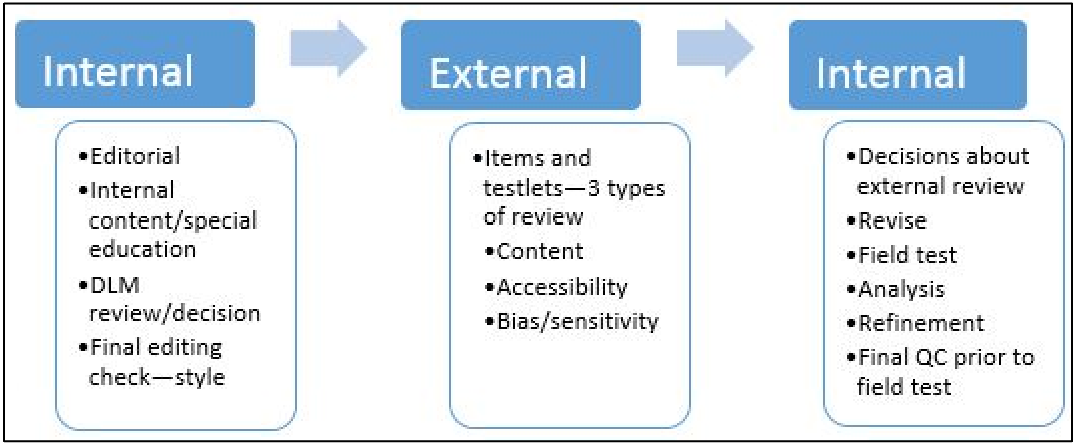

Internal
•Editorial
•Internal content/special education
•DLM review/decision
•Final editing check—style
External
•Items and testlets—3 types of review
•Content
•Accessibility
•Bias/sensitivity
Internal
•Decisions about external review
•Revise
•Field test
•Analysis
•Refinement
•Final QC prior to field test

Figure 32. Overview of the Item Review Processes prior to field testing for the Dynamic Learning Maps Alternate Assessment System.

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 and used the criteria for that panel to conduct reviews. Reviewers evaluated items grouped together in testlets. For each item and each testlet, reviewers made one of three decisions: accept, requires critical revision, or reject. Reviewers made decisions independently and without discussion with other reviewers.

Reviews of testlets for students who are blind or have visual impairments were also conducted during the 2014–2015 academic year. These testlets were assigned to volunteers who had experience working with students with significant cognitive disabilities or experience working with students who are blind or have low vision. The results of these reviews are included with the results of the other external reviews in the following sections.

III.1.F.ii. Review Assignments and Training

For external reviews in 2014-2105, 391 people responded to a volunteer survey used to recruit panelists. Volunteers for the External Review process completed a Qualtrics survey to capture demographic information as well as information about their education and experience. This data is then used to identify panel types for which the volunteer would be eligible. Of the 391 respondents, 226 people completed the required training and 181 of those were placed onto external review panels. Each reviewer was assigned to one of the three panel types. There were 91 ELA reviewers, 30 on accessibility panels, 30 on content panels, and 31 on bias and sensitivity panels. There were 90 math reviewers, with 30 people assigned to each panel type.

The current professional roles reported by reviewers is shown in Table 19. Reviewers who reported "other" included SEA and LEA staff, university professors, independent special education consultants, and reading specialists.

Table 19. Professional Roles of External Reviewers

	Math		**ELA**	
	n	%	*n*	%
Classroom Teacher	49	54.44	48	52.74
District Staff	7	7.77	10	10.98
Homebound Teacher	0	0.00	1	1.09
Instructional Coach	5	5.55	6	6.59
Other	29	32.22	26	28.57

Reviewers were had experience teaching SWSCDs. ELA reviewers had a median of 10 year years of experience and mathematics reviewers had a median of 12 years of experience.

Review assignments were made throughout the year. Reviewers were notified by email each time they were given an assignment of collections of testlets. Each review assignment took 1.5 to 2 hours. In most cases, reviewers had two weeks to complete an assignment.

Before reviewing testlets, participating reviewers were required to complete several online training modules. These modules included detailed instructions on the review process, a quiz, and a practice activity. This training had to be completed successfully before reviewers began reviewing for the year. Training was completed in segments, taking 60 to 75 minutes total. Training information was made available online.

III.1.F.iii. Reviewer Responsibilities

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

III.1.F.iv. Decisions and Criteria

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

III.1.F.iv.a 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 20.

Table 20. General Review Decisions for External Reviews

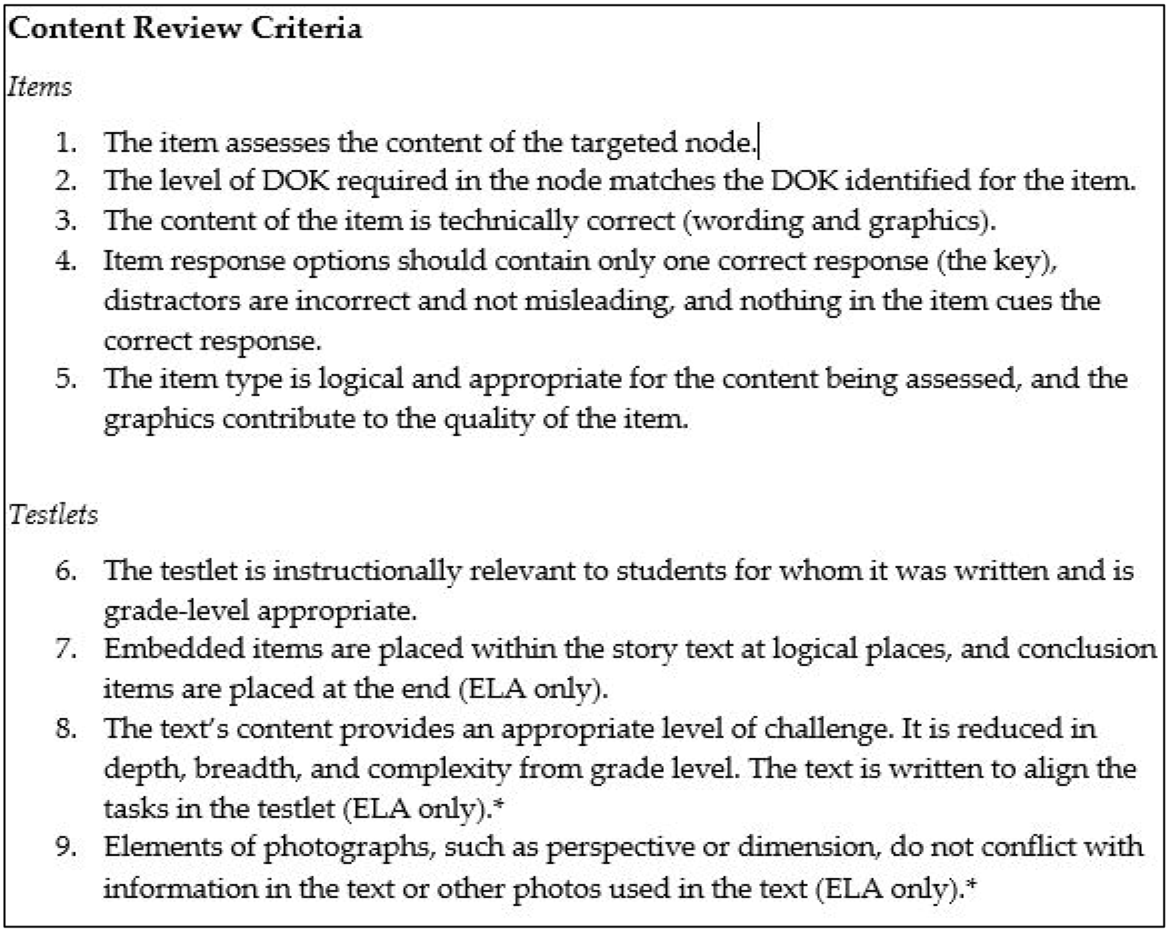
Decision	Definition
Accept	Item/testlet is within acceptable limits. It may not be perfect, but it is worth putting through field tests and seeing how it goes.
Critical Revision Required (Revise)	Item/testlet violates one or more criteria. It has some potential merits and can be acceptable for field-testing after revisions to address the criteria.
Reject	Item/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 were used for items and testlets. Therefore, it was possible to recommend revisions or rejections to items without automatically having to recommend revision or rejection to the testlet as a whole. If a reviewer recommended revision or rejection, he or she was required to provide an explanation that included identification of the problem and, in the case of revision, a proposed solution.

III.1.F.iv.b Review Criteria

While most of the external review process was the same in 2013-2014 and 2014-2015, the 2014–2015 academic year reviews included some changes based on outcomes from the initial year. First, recruitment timelines were made more specific. Rather than having a single ongoing volunteer window, three phases of volunteering were implemented. Each phase had a deadline for reviewers to submit the volunteer survey, complete the required training, and receive the first assignment. In addition, the openings of the three phases was staggered across the months of August and September to account for varying school-year start times across the states in the consortium. Some external review criteria were also changed. These criteria are noted with footnotes in the following section.

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 practice activity. There were specific criteria for external reviewers of content, accessibility, and bias and sensitivity. Figure 33, Figure 34, and Figure 35 show the review criteria.



Content Review Criteria

Items

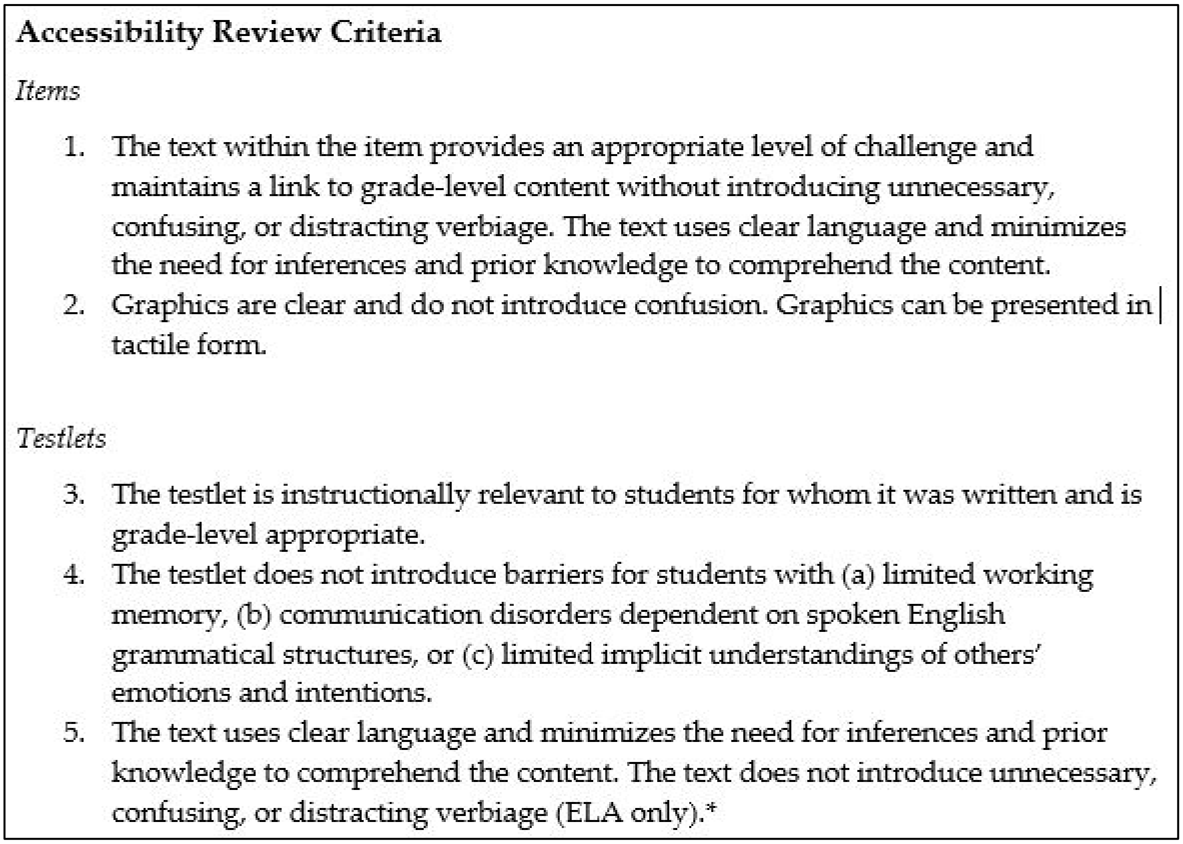
1. The item assesses the content of the targeted node.
2. The level of DOK required in the node matches the DOK identified for the item.
3. The content of the item is technically correct (wording and graphics).
4. Item response options should contain only one correct response (the key), distractors are incorrect and not misleading, and nothing in the item cues the correct response.
5. The item type is logical and appropriate for the content being assessed, and the graphics contribute to the quality of the item.

Testlets

6. The testlet is instructionally relevant to students for whom it was written and is grade-level appropriate.
7. Embedded items are placed within the story text at logical places, and conclusion items are placed at the end (ELA only).
8. The text's content provides an appropriate level of challenge. It is reduced in depth, breadth, and complexity from grade level. The text is written to align the tasks in the testlet (ELA only).*
9. Elements of photographs, such as perspective or dimension, do not conflict with information in the text or other photos used in the text (ELA only).*

Figure 33. Content Review Criteria.

Note: Asterisks (*) indicate criteria that were added for 2014–15.



Accessibility Review Criteria

Items

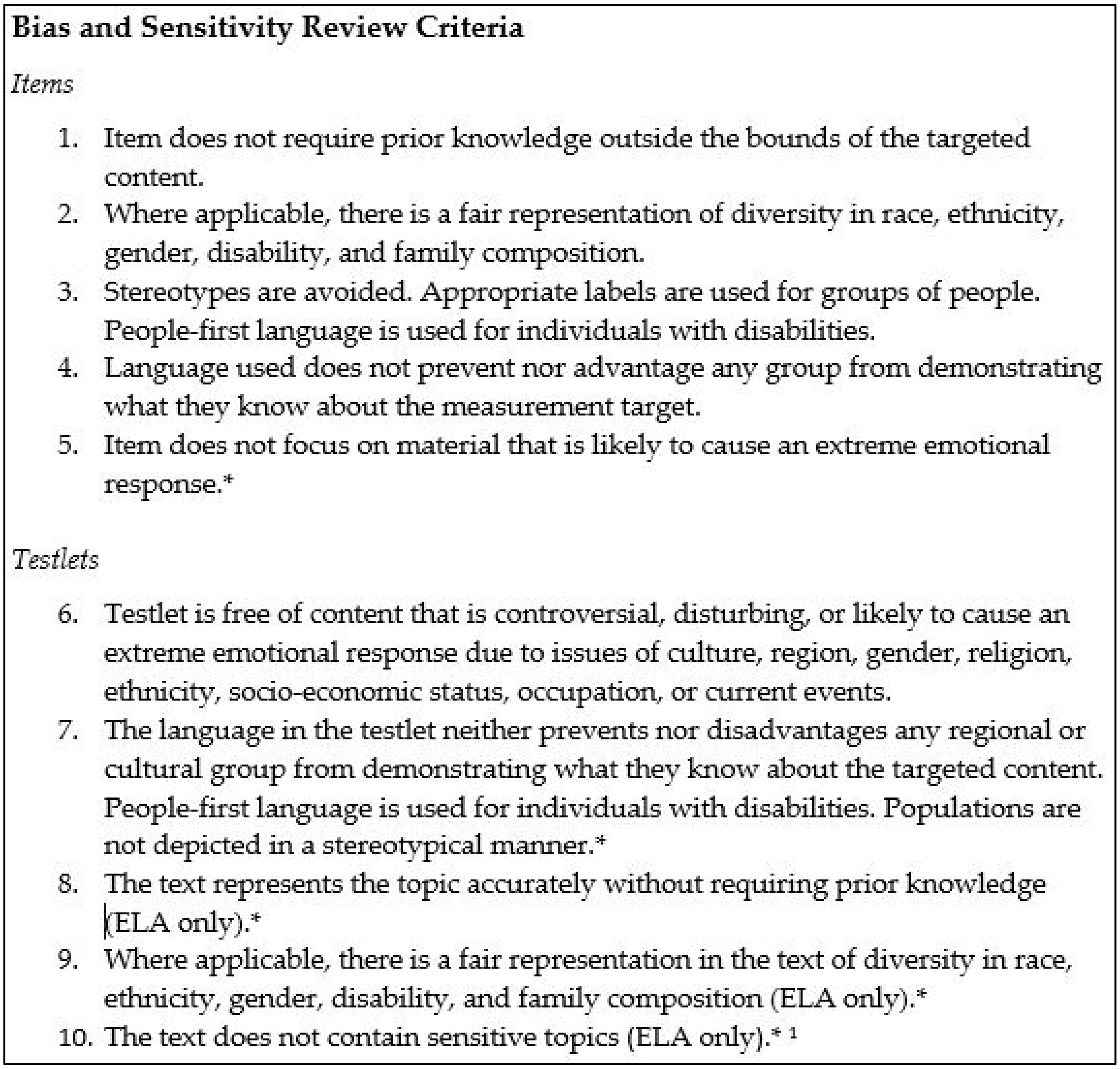
1. The text within the item provides an appropriate level of challenge and maintains a link to grade-level content without introducing unnecessary, confusing, or distracting verbiage. The text uses clear language and minimizes the need for inferences and prior knowledge to comprehend the content.
2. Graphics are clear and do not introduce confusion. Graphics can be presented in tactile form.

Testlets

3. The testlet is instructionally relevant to students for whom it was written and is grade-level appropriate.
4. 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.
5. The text uses clear language and minimizes the need for inferences and prior knowledge to comprehend the content. The text does not introduce unnecessary, confusing, or distracting verbiage (ELA only).*

Figure 34. Accessibility Review Criteria.

Note: Asterisks (*) indicate criteria that were added for 2014–15.



Bias and Sensitivity Review Criteria

Items

1. Item does not require prior knowledge outside the bounds of the targeted content.
2. Where applicable, there is a fair representation of diversity in race, ethnicity, gender, disability, and family composition.
3. Stereotypes are avoided. Appropriate labels are used for groups of people. People-first language is used for individuals with disabilities.
4. Language used does not prevent nor advantage 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.*

Testlets

6. 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.
7. The language in the testlet neither prevents nor disadvantages any regional or cultural group from demonstrating what they know about the targeted content. People-first language is used for individuals with disabilities. Populations are not depicted in a stereotypical manner.*
8. The text represents the topic accurately without requiring prior knowledge (ELA only).*
9. Where applicable, there is a fair representation in the text of diversity in race, ethnicity, gender, disability, and family composition (ELA only).*
10. The text does not contain sensitive topics (ELA only).* [1]

Figure 35. Bias and Sensitivity Review Criteria.

Note: Asterisks (*) indicate criteria that were added for 2014–15.

All three types of reviews focused on both items and testlets. Content reviews of items included alignment to the targeted node in the DLM maps, congruence of the item and node, level of cognitive process dimension taxonomy, quality and appropriateness of the content, accuracy of response options, and appropriateness of distractors. Testlet content reviews focused on the instructional relevance to students and grade-appropriateness, as well as the logic of item placement within narrative text. Accessibility item reviews focused on appropriate challenge levels and the maintenance of links to grade-level content. For accessibility reviews, testlets were checked for instructional relevance at grade level and minimizing of barriers to students with specific needs. Finally, item-level bias and sensitivity reviews included identifying items that require prior knowledge outside the bounds of the targeted content, ensuring fair representation of diversity, avoiding stereotypes and negative naming, removing language that affects a student's demonstration of their knowledge on the measurement target, and removing any language that was likely to cause strong emotional response. For testlet reviews, criteria were applied similar to item-level reviews, with emphasis on reducing the chance of construct-irrelevant variance due to inadvertent use of controversial, disturbing, stereotypic, or negative language or graphics. The texts used in ELA reading testlets were reviewed using criteria related to content, accessibility, and bias and sensitivity as a part of the external review of testlets.

III.1.G. Results of Reviews

The majority of the content reviewed during the 2013–2014 academic year was included in the fall pilot and spring field-testing events. On a limited basis, content for the upcoming 2014–2015 school year was also reviewed. For ELA, the percentage of items or testlets rated as "accept" ranged across grades, pools, and rounds of review from 72% to 91%. The rate at which content was recommended for rejection ranged from 1% to 5% across grades, pools, and rounds of review. For mathematics, the percentage of items or testlets rated as "accept" ranged from 76% to 88%. The rate at which content was recommended for rejection ranged from 2% to 3%. A summary of the content team decisions and outcomes is provided here. A more detailed report and outcomes from external reviews is included in the external review technical report for 2013-2014 (Clark, Karvonen, & Swinburne Romine, 2014) and the external review technical report for 2014-2015 (Clark, Swinburne Romine, Bell, & Karvonen, 2015).

III.1.G.i.a Content Team Decisions

Because multiple reviewers examined each item and testlet, external review ratings were compiled across panel types. DLM staff reviewed and summarized the recommendations provided by the external reviewers for each item and testlet. Based on that combined information, staff had five decision options: (a) no pattern of similar concerns, accept as is; (b) pattern of minor concerns, will be addressed; (c) major revision needed; (d) reject; and (e) more information needed.

Content teams documented the decision category applied by external reviewers to each item and testlet. Following this process, content teams made a final decision to accept, revise, or reject each of the items and testlets. The ELA content team retained 98% of items and testlets sent out for external review. Of the items and testlets that were revised, most required only minor changes (e.g., minor rewording but concept remained unchanged), as opposed to major changes (e.g., stem or option replaced). The ELA team made a total of 124 minor revisions to items and 84 minor revisions to testlets. The mathematics content team retained 99% of items and testlets sent out for external review. As with ELA, most revisions made to items and testlets were minor. The mathematics team made a total of 387 minor revisions to items and 186 minor revisions to testlets. Additional detail on review outcomes is included in the external review technical reports (Clark, Karvonen, & Swinburne Romine, 2014; Clark, Swinburne Romine, Bell, & Karvonen, 2015).

III.1.H. The First Contact Survey

The linkage level for the student's first testlet is determined based on responses to the 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 assessments. Supporting procedures and a complete list of First Contact questions is included in the *Test Administration Manual 2014-15* (Dynamic Learning Maps, 2014). Test administrators are trained on the role of First Contact in testlet assignment as part of required test administrator training (see Chapter X).

Three sections of the First Contact survey are used to provide an optimal match between student and testlet during the initial DLM testing experience: (1) Expressive Communication, (2) Reading Skills, and (3) Math Skills. From these responses, the student's assigned complexity band is calculated automatically and stored in the system. For English language arts reading and writing testlets, the responses to the Expressive Communication and Reading Skills questions are used. For assignment to mathematics testlets, the responses to the Expressive Communication and Math Skills questions are used. If a different complexity band is indicated between the two sets of questions, the lower band is selected. The goal is to present a testlet that is approximately matched to a student's knowledge, skills, and abilities. That is, within reason, the system should present a testlet that is neither too easy nor too difficult and should provide a positive experience for the student entering the assessment.

Based on the complexity band assigned by the First Contact survey, the student's first testlet could be at delivered at one of four levels. The Foundational band will deliver a testlet written at the Initial Precursor level, to be appropriate for students who do not use speech, sign or AAC, do not read any words when presented in print (ELA) or do not sort objects (math). Band 1 will deliver a testlet at the Distal Precursor level for students who use one word, sign, or symbol to communicate, recognizes symbols (ELA), or sorts symbols (math). Band 2 will deliver a testlet at the Proximal Precursor level for students who use 2 words, signs, or symbols to communicate, reads at the primer to 2nd grade level (ELA), or adds/subtracts up to 80% of the time (math). Band 3 will deliver a testlet at the Target level for students who regularly combine three or more spoken words to communicate for a variety of purposes and are able to read print at the 3rd grade level or above (ELA) or regularly adds/subtracts and forms groups of objects (math). Because there are only four complexity bands, no testlets written at the Successor level are delivered as the first testlet. However, a student is able to route to the Successor level by providing correct responses to items on a Target level testlet.

III.1.I. PILOT ADMINISTRATION

A pilot administration of the DLM Alternate Assessment System was conducted in the fall of 2013. The purpose of the pilot assessment was to evaluate the following research questions:

1. Will complexity bands support the online KITE administration system to present a testlet that is relatively well matched to students' knowledge, skills, and abilities, as evidenced by educator responses to the First Contact survey?
2. What feedback did educators have about student and educator experience with testlet contents and the testing platform?

Additionally, data from the pilot was used to conduct the first exploratory modeling work when fitting cognitive diagnostic models and Bayesian networks. A complete description of the findings from the pilot can be found in Clark, Kingston, Templin, & Pardos (2014).

III.1.I.i. Overview

The pilot assessment was available to educators and students in states belonging to the Dynamic Learning Maps Consortium from October 21 to November 22, 2013. A total of 1,409 students completed assessments, and 597 educators responded to educator surveys that were administered within the KITE platform.

Content was available for both ELA and mathematics. Each DLM content team selected a single EE to be assessed in each of three grade bands: third-fourth, seventh-eighth, and high school. A fixed-form assessment was built for each grade band and assessed the chosen EE at three different linkage levels. All forms consisted of three testlets at three different linkage levels: Initial Precursor, either Distal or Proximal Precursor, and Target. All students started at the lowest linage level and progressed to the highest linkage level. Educators were asked to administer as much of the form as possible but were given the option to exit at any time. By administering the same set of testlets for a single EE to all students in the grade band, the DLM test development team was able to gauge how a range of students responded to the varied levels of complexity and used that data to inform initial assignment to a complexity band.

III.1.I.ii. Initialization

In response to the first research question, student initialization into the DLM system was examined to evaluate the match of testlet linkage level to the student. Responses to the First Contact survey were used to determine the student's complexity band. Two approaches to testlet assignment based on First Contact responses were compared for students taking the pilot assessment:

1. Assign each student to a complexity band based solely on First Contact responses that pertain to academic performance in ELA or mathematics; or

2. Assign each student to the lower of two complexity bands:
   a. First Contact responses that pertain to academic performance in ELA or mathematics or
   b. First Contact responses that pertain to expressive communication skills.

III.1.I.ii.a Comparison of Approaches for Determining Complexity Band

Data collected from the pilot assessment were used to evaluate the two different initialization methods. The percentages of students classified in each complexity band for ELA and mathematics are presented in Table 21. Similar values are evident in ELA and mathematics. These values indicate that the combined approach of using the lower band for content area or expressive communication provides a slightly more conservative classification to complexity bands; a small percentage of students are placed at a lower complexity band after taking into account their expressive communication ability. The percentage of students impacted in the pilot sample ranged from 5% to 9% based on grade and content area.

Table 21. Percentages of Students Classified into Complexity Bands

Complexity Band	ELA		Mathematics	
	Content Only (%)	Combined (%)	Content Only (%)	Combined (%)
Foundational	20	23	20	24
Complexity Band 1	31	33	32	32
Complexity Band 2	33	31	36	36
Complexity Band 3	16	13	12	10

These findings were presented to the TAC and shared with the state partners. Based on their review of the results, the combined algorithm, which takes the lower band between content and expressive communication, was selected for initialization for the first field testing events. Although the decision would result in a small portion of students being placed at an initially lower complexity level, the DLM test development team believes it is preferential to have students enter the assessment with items that are too easy than with items that are too difficult. The conservative approach would potentially provide more students with a positive initial testing experience.

III.1.I.ii.b Student Performance Within and Across Complexity Bands

To determine whether the complexity bands provided meaningful distinctions between students at varying levels of knowledge, skill, and ability, analyses were conducted to determine the extent that students categorized in the four complexity bands differed from one another in their item responses. One hypothesis was that if the complexity bands provide meaningful distinctions between students, then the percentage of students responding correctly to items should be higher on the same testlet for student classified into higher complexity bands. Also, the percentage of students within an assigned complexity band who respond correctly to items should decrease as linkage level increases.

A set of example findings are presented in Table 22. The complete results can be found in the *Summary of results from the fall 2013 pilot administration of the Dynamic Learning Maps™ Alternate Assessment System* (Clark, Kingston, Templin, & Pardos, 2014). In Table 22, the rows represent students grouped by complexity bands, increasing from foundational (F) to complexity band 3 (CB 3). The table provides the percentage of correct responses, including non-attempts as incorrect responses, for each item administered in the seventh–eighth grade band assessment for ELA.

Table 22. Seventh–Eighth Grade ELA Percentage Correct by Item

Complexity Band	Initial Precursor Testlet			Distal Precursor Testlet			Target Testlet			
	Item 1 (%)	Item 2 (%)	Item 3 (%)	Item 1 (%)	Item 2 (%)	Item 3 (%)	Item 1 (%)	Item 2 (%)	Item 3 (%)	Item 4 (%)
F (*N*=90)	39	36	43	24	28	27	27	24	26	22
CB 1 (*N*=92)	75	46	62	32	39	42	40	34	28	41
CB 2 (*N*=114)	96	82	79	77	59	72	50	53	75	67
CB 3 (*N*=54)	100	94	94	93	67	93	67	78	81	83

The percentage of correct responses at the item level was lowest for students at the foundational level, as expected, and increased as the complexity band increased from complexity band 1 to complexity band 3. Similarly, because the testlets were ordered from lowest linkage level to highest, the percentage of correct responses generally decreased from testlet 1 to testlet 3. Similar results were found across grade bands and content areas. These findings are one source of evidence indicating that the complexity bands create a meaningful distinction among students in order to provide them with the best match of item complexity during the initial testing experience.

Because students were able to exit the assessment at any time, the DLM test development team was interested in determining how many students within each complexity band attempted all three testlets. This information served as another source of evidence regarding appropriateness of the linkage level assignment based on complexity band. Table 23 provides the percentages of students who attempted all three testlets, by grade band and content area. Students at the foundational level attempted all three testlets less frequently than students at higher complexity bands. This is an expected finding, as students at the foundational level would typically only be administered testlets at the Initial Precursor level, and only the first testlet in the pilot was at this linkage level. Students at complexity band 3 would typically be assigned items at the Target level or beyond and thus would be expected to be able to respond to all content presented in the pilot. Future analyses will examine completion rates within each testlet.

Table 23. Percentage of Students Who Attempted All Testlets by Grade and Content Area

Complexity Band	**ELA**			**Mathematics**		
	3rd–4th (%)	**7th–8th (%)**	**HS (%)**	**3rd–4th (%)**	**7th–8th (%)**	**HS (%)**
F	53	77	81	63	69	68
CB 1	85	79	79	86	76	72
CB 2	84	92	90	89	88	98
CB 3	100	94	100	75	98	95

III.1.I.ii.c Regression Analyses

A series of regression analyses were conducted to further evaluate the extent that the proposed initialization algorithm was supported by the pilot data. A hierarchical ordinary least squares regression model was used to predict the total number of correct items for each content area assessment using the previously specified First Contact survey variables. Many of the survey variables had to be dummy-coded because they are categorical variables. This created a large number of predictors, so a reduced set of variables was used to remove redundancy in the number and overlap of variables for each skill. Variables included addition/subtraction and sorting for mathematics; two reading levels (up to primer level and beyond primer level) and symbol recognition for ELA; and a single expressive communication variable reflecting the student's highest level of expressive communication using spoken word, sign, or AAC. The sequential nature of the model was such that the mathematics and ELA First Contact predictors were added to the model first, followed by the expressive communication variable, to determine the extent to which additional variance was explained by its inclusion.

The hierarchical ordinary least squares regression models were statistically significant for both ELA and mathematics across all three grade-band assessments (see Table 24). The amount of variance explained by the mathematics First Contact predictors was between 14% and 38%, with an additional 3% to 5% of variance explained by including the expressive communication variable. For ELA, the amount of variance explained by the First Contact predictors was between 17% and 53%, with an additional 6% to 9% of variance explained by the inclusion of the expressive communication variable. For all grade bands and content areas, the addition of expressive communication resulted in a significant change to model-data fit.

Table 24. Ordinary Least Squares Regression Results by Grade and Content Area

Grade and Content Area	*F*	*df*	*p*	*R^2*
3rd–4th grade mathematics	16.9	2, 258	< .001	.14
7th–8th grade mathematics	59.4	2, 247	< .001	.35
High school mathematics	21.3	2, 107	< .001	.38
3rd–4th grade ELA	14.3	2, 258	< .001	.17
7th–8th grade ELA	96.8	2, 247	< .001	.52
High school ELA	14.0	2, 107	< .001	.26

Next, a hierarchical ordinal logistic regression model was used to predict the probability of success for students at each linkage level testlet. Success at the testlet level was determined by obtaining a threshold of 67% correct. The same First Contact variables were used as predictors. Again, the mathematics variables were significant predictors of mathematics linkage level, χ^2 (7) = 165.24, $p < .001$, with a Nagelkerke pseudo R^2 value of .26. The expressive communication variable raised the value by .02. Similar findings were evident for ELA, χ^2 (4) =117.21, $p < .001$, with a Nagelkerke value of .18. The inclusion of the expressive communication variable increased the value by .04. For both content areas, the addition of the expressive communication variable resulted in a significant change to Nagelkerke pseudo R^2 values. Similar findings were obtained using binary logistic regression models to predict success at each linkage level testlet independently.

Predicted and observed values were also compared, and the root mean squared error (RMSE) was calculated to quantitatively capture how accurate each model was in predicting actual student values for the three linkage level categories. The complete set of findings can be found in the pilot technical report (Clark, Kingston, Templin, & Pardos, 2014). Overall, the addition of expressive communication variables to the models resulted in a slightly smaller RMSE value for both content areas, and more conservative classification to linkage levels.

III.1.I.iii. Educator Survey

As part of the pilot testing event, educators were asked to complete a survey about each participating student's experience with the assessment. This survey was designed to provide feedback on several aspects of the educator and student experience, including testlet contents and delivery via the user interface. Survey items pertaining to item and test development are presented here. The complete survey results can be found in the pilot technical report (Clark, Kingston, Templin, & Pardos, 2014). All participating educators were presented with seven survey items on a common form, followed by one of five randomly administered forms containing between two and twelve additional items. The survey contained a mix of selected response and open-ended response items. Educators were not required to respond to all items and could exit the survey at any time.

A total of 1,209 educator responses to the survey were recorded, indicating a response rate of around 86%. The distribution of responses by grade band is presented in Table 25.

Table 25. Educator Responses to Survey by Grade Band

Grade band	Students assessed	Educator responses	%
3rd–4th	477	400	84
7th–8th	546	464	85
High school	393	324	82

One way of evaluating student engagement with testlet content was to investigate when and why educators chose to exit a testlet. Respondents indicated that a total of 436 students, or 36%, exited a testlet prior to its completion. Reasons for exiting a testlet prior to completion were examined across complexity bands to determine where similarities or differences were evident. Because a separate complexity band was calculated for each content area, results were prepared by content area even though the survey question was not content specific. Findings for each complexity band are presented in Table 26 and Table 27 for ELA and mathematics respectively. Note that percentages add up within a column rather than across a row. The percentage of students who did not exit a testlet prior to completion increased across complexity bands. Of those students in the foundational band who did exit a testlet, the most frequent reason was the student did not know the content, while for students in complexity band 3, the most common reason was frustration or disengagement.

Table 26. Reasons for Exiting Testlets by ELA Complexity Band

Reason for exiting	F		CB 1		CB 2		CB 3	
	n	%	*n*	%	*n*	%	*n*	%
Did not exit	162	69	162	70	169	80	71	89
Extreme frustration or disengagement	15	6	21	9	4	2	4	5
Student's behavior or health interfered	7	3	10	4	1	1	0	0
Accidental exit	6	3	7	3	16	7	1	1
Student did not know anything about the content	36	15	22	9	9	4	0	0
Accessibility features were not working	1	1	4	2	1	1	1	1
Other reason	7	3	6	3	11	5	3	4

Table 27. Reasons for Exiting Testlets by Mathematics Complexity Band

Reason for exiting	F		CB 1		CB 2		CB 3	
	n	%	*n*	%	*n*	%	*n*	%
Did not exit	151	65	169	71	184	83	60	94
Extreme frustration or disengagement	15	6	21	9	8	4	0	0
Student's behavior or health interfered	8	3	9	4	1	0	0	0
Accidental exit	7	3	9	4	13	6	1	2
Student did not know anything about the content	43	19	20	8	4	2	0	0
Accessibility features were not working	1	0	2	1	4	2	0	0
Other reason	7	3	9	4	8	4	3	5

Because the pilot was the first opportunity for students to interact with the system, DLM test development staff were interested in evaluating how independently students interacted with the system. Approximately 45% of students required prompting, support, or redirection from their educator during the assessment *and* could not enter their own responses on the computer.

These survey responses were further examined by complexity band to determine whether level of independence varied by complexity band. Table 28 and Table 29 present the findings for the mathematics and ELA complexity bands, respectively. Students classified in lower complexity bands had less independence when interacting with the system, while students classified in higher complexity bands had greater levels of independence. Few students at any complexity band interacted with the assessment system without any prompting, redirection, or support from an educator.

Table 28. Student Interaction with the System by Mathematics Complexity Band

Type of Interaction	**F**		**CB1**		**CB2**		**CB3**	
	n	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Did not require supports and entered responses independently	1	2	3	6	8	17	1	8
Required supports and entered responses independently	2	5	6	13	27	58	9	76
Did not require supports and did not enter responses independently	2	5	8	17	10	21	1	8
Required supports and did not enter responses independently	39	88	31	64	2	4	1	8

Table 29. Student Interaction with the System by ELA Complexity Band

Type of Interaction	F		CB1		CB2		CB3	
	n	%	*n*	%	*n*	%	*n*	%
Did not require supports and entered responses independently	1	2	1	2	7	14	4	36
Required supports and entered responses independently	2	5	9	18	28	57	5	45
Did not require supports and did not enter responses independently	2	5	8	16	9	18	2	18
Required supports and did not enter responses independently	39	89	31	63	5	10	0	0

These additional findings were also related to testlet design:

- Educators indicated that the multiple-choice item type met student's needs for 68% of students.
- The drag-and-drop item type, used only in high school ELA testlets, met student's needs for 65% of students.
- The amount of text presented on a single screen met student's needs for 69% of students.
- The engagement activities at the beginning of each testlet met student's needs for 56% of students.

In addition to responding to selected response items in the pilot, educators were also asked to provide open-ended feedback.

Many educators commented on the level of difficulty of the items included in the pilot, stating that items were either too challenging or too easy for their students. These comments were expected due to the structure of the fixed-form pilot assessment that included testlets at multiple linkage levels in order to obtain information about the ideal point of entry for students with varying levels of knowledge, skill, and ability. Because of this, students were presented with a wider range of testlet complexity than they ordinarily would receive during a DLM testing session. While upcoming field tests continued to evaluate initial linkage level placement, data obtained from the pilot helped the DLM test development team determine how to administer content more closely aligned with each student's knowledge and skills.

Another frequently received comment pertained to the desire for a greater number of images in the items. For ELA testlets, the DLM test development team had previously determined through cognitive labs that a single picture would be presented with each screen, which typically contains a sentence or two of text. The DLM test development team also had designed texts to de-emphasize image use as answer options in items, based on research on the development of early reading skills and the bias that would be introduced for students who are blind or have visual impairments. As such, the images included do not convey information a student would need to know beyond the information presented in the text. These comments suggested a need for more educator education about the intentional design of ELA texts. For mathematics, the pilot testlets contained fewer items with images than is typical in the pool of mathematics items overall. Many mathematics items included in future testing events include images in the stem and/or answer options.

III.1.J. FIELD TESTING

The 2014 and 2015 DLM field tests were administered to evaluate item quality for EEs assessed at each grade level for mathematics and ELA. In addition to evaluating item quality, the field tests also continued to evaluate student initialization into the assessment system and gain educator feedback on aspects of the assessment system. A complete summary of the field test events can be found in *Summary of results from the 2014 and 2015 field test administrations of the Dynamic Learning Maps™ Alternate Assessment System* (Clark, Karvonen, & Wells Moreaux, 2016).

For each field test window, the mathematics and ELA content teams selected testlets to be assessed for grades three through twelve. Testlets were made available at five linkage levels for each EE. A description of the specific characteristics of each field test is provided.

III.1.J.i. Description of Field Tests

A total of six field test events were conducted in 2014 and 2015. Table 30 summarizes the dates of each field test window. The lengths of each field test window ranged from 10 business days to 9 weeks.

Table 30. Date Ranges for Each Field Test Window

Field Test	Open Date	Close Date
Field Test 1	February 10, 2014	February 21, 2014
Field Test 2	March 17, 2014	April 11, 2014
Field Test 3	May 1, 2014	June 13, 2014
Phase A	October 13, 2014	October 31, 2014
Phase B	November 10, 2014	December 19, 2014
Phase C	January 5, 2015	March 6, 2015

Field Tests 1 and 2 occurred prior to the adoption of operational test blueprints, and as a result, only included optionally available instructionally-embedded testlets. Two EEs were assessed at each grade and content area. The initialization algorithm developed and tested during the pilot study was used to select the linkage level testlets the student received during Field Test 1 and Field Test 2. A total of 199 testlets were administered during Field Test 1. A total of 296 testlets were administered during Field Test 2. Of those testlets, 44 were administered in Field Test 1, as well.

Field Test 3 was designed to more closely reflect the operational assessments that would be available in the 2014–15 year. Similarly to Field Tests 1 and 2, only optionally available instructionally-embedded testlets were included. During Field Test 3, students received three testlets, all at the same linkage level, based on initialization from responses to the First Contact survey. These three testlets each assessed a different EE out of the five available for each grade and content area. A total of 738 testlets were administered across all grades and content areas. No testlets were re-administered from Field Tests 1 or 2 during Field Test 3.

The Phase A field test was structured similar to Field Test 3 in preparation for the opening of operational testing. Students were assigned between three and four testlets per content area at a single linkage level based on their First Contact survey results.

Because blueprints were developed and approved by states in spring 2015, the Phase A window was the first to include testlets designed to cover the year-end blueprint. A total of 331 testlets were available across grades and content areas.

Phase B was the first field test window to include complete coverage of all EEs required by the blueprints and all linkage levels for both content areas. Testlets available during Phase B were delivered using the same method in Phase A: an enrollment process automatically assigned up to four testlets, all at a single linkage level. A total of 808 testlets were available during Phase B.

During Phase C, testlets were delivered following the sequencing and adaptive algorithm rules planned for the spring operational testing window (See Chapter IV for a description of algorithm). Testlets were available to cover the complete blueprint, with students receiving between 4 and 7 testlets. Phase C included 810 testlets.

States provided their users with guidance on the number of field test testlets to complete during Phase C. In most states, participation was voluntary.

A summary of educator, district, and state participation during each of the field test windows is presented in Table 31. The counts are based on students with at least one testlet complete or in progress during the window dates, and include all consortium states.

Table 31. Participation Summary during Field Test Windows

	Field Test 1	**Field Test 2**	**Field Test 3**	**Phase A**	**Phase B**	**Phase C**
Educators	3,288	3,673	3,375	3,490	4,895	5,870
Districts	608	648	654	936	1,087	1,470
States	14	16	17	8	12	17

Students and educators were recruited for participation in each of the six field test events by state and district education agencies within the DLM Consortium. In most states, participation was voluntary. Students and educators participated in anywhere from one to all of the field test events during the 2014 and 2015 years. A summary of the number and demographic percentages for students participating in each field test window are presented in Table 32. Included in Table 32 are reported percentages of gender, disability, race, ethnicity, and complexity band (Chapter IV) for both ELA and mathematics.

Table 32. Demographic Summary of Students Participating in Field Test Windows (Year-End)

	FT1	FT2	FT3	A	B	C
Demographic Group	n=4538	n=5103	n=4701	n=2049	n=2080	n=5362
	%	%	%	%	%	%
Gender						
Female	18.07	16.99	17.36	32.45	34.52	33.18
Male	34.38	31.53	32.82	65.50	63.94	64.99
Missing	47.55	51.48	49.82	2.05	1.54	1.83
Primary Disability						
Autism	4.74	4.62	4.57	12.98	8.37	9.46
Deaf/blindness	0.04	0.04	0.04	NA	0.19	0.06
Developmentally delayed	1.90	1.70	1.83	2.34	3.56	2.63
Emotional disturbance	0.29	0.24	0.28	0.20	0.29	0.35
Hearing impairment	0.11	0.10	0.11	0.10	0.34	0.24
Intellectual disability	9.74	8.60	10.08	11.62	14.33	12.53
Specific learning disability	0.79	0.78	0.91	1.07	0.87	1.23

	FT1	FT2	FT3	A	B	C
Demographic Group	n=4538	n=5103	n=4701	n=2049	n=2080	n=5362
	%	%	%	%	%	%
Multiple disabilities	2.97	3.43	3.49	7.08	5.48	5.59
Mental retardation	1.50	1.53	1.19	1.02	0.67	0.88
Other health impairment	2.58	2.27	2.30	2.39	1.63	2.13
Orthopedic impairment	0.46	0.47	0.51	0.39	0.24	0.22
Speech/language disability	0.11	0.20	0.11	1.56	0.53	0.50
Traumatic brain injury	0.33	0.27	0.32	0.15	0.29	0.24
Visual impairment	0.04	0.06	0.06	0.05	0.19	0.13
Missing	74.39	75.68	74.20	59.05	63.03	63.80
Comprehensive Race						
White	0.11	0.16	0.43	55.39	61.20	57.50
Black or African American	9.23	6.90	15.12	10.10	11.25	14.23
Asian	0.22	0.41	0.09	2.98	2.12	2.28
American Indian or Alaska Native	18.60	14.11	14.27	NA	NA	NA
American Indian	NA	NA	NA	12.10	15.10	14.45
Alaska Native	NA	NA	NA	0.10	0.38	0.80
Two or More Races	NA	NA	NA	8.69	4.76	4.40
Native Hawaiian/Pacific Islander	0.29	0.39	0.55	0.54	1.06	0.76
Missing	71.55	78.03	69.54	10.10	4.13	5.58
Hispanic Ethnicity						
No	NA	NA	NA	35.09	33.99	46.57
Yes	NA	NA	NA	4.00	2.60	3.80
Missing	1.00	1.00	1.00	60.91	63.41	49.63
ESOL Participation						
Not ESOL eligible/monitored student	NA	NA	NA	97.32	97.55	96.42
ESOL eligible/monitored student	NA	NA	NA	2.68	2.45	3.58

	FT1	FT2	FT3	A	B	C
Demographic Group	n=4538	n=5103	n=4701	n=2049	n=2080	n=5362
	%	%	%	%	%	%
ELA Band						
Foundational	18.47	18.85	17.95	13.62	14.18	16.13
Band1	27.63	28.28	28.19	26.40	28.27	28.33
Band2	36.32	36.02	36.55	38.21	39.33	38.18
Band3	16.88	16.64	16.57	21.72	18.22	17.36
Missing	0.71	0.22	0.74	0.05	NA	NA
Math Band						
Foundational	20.27	20.58	19.93	16.84	16.30	18.00
Band1	29.88	30.20	29.76	29.72	29.81	30.77
Band2	38.25	37.76	38.72	38.65	42.79	40.00
Band3	10.84	11.25	10.85	14.74	11.11	11.23
Missing	0.75	0.22	0.74	0.05	NA	NA

III.1.J.ii. Field Test Results

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

III.1.J.ii.a Item Flagging Criteria

In order to focus the content teams' review of field test items, flagging criteria were developed to identify items in need of review by the teams. Items were flagged for review by content teams if they met any of the following statistical criteria:

- The item was too challenging, as indicated by a percent correct (*p*-value) less than 35%. This value was selected as the threshold for flagging due to most DLM items consisting of three response options, so a value less than 35% may indicate chance selection of the option.

- The item was significantly easier or harder than other items assessing the same node within the grade level, as indicated by a weighted standardized difference greater than two standard deviations from the mean *p*-value for that node.

Items that had a sample size of at least 20 cases were reviewed, and those items with a sample size of less than 20 were retested to collect additional data prior to making item-quality decisions.

III.1.J.ii.b Item Data Review Decisions

Content teams made four types of item-level decisions as they reviewed field test results:

1. No changes made to item. Content team decided item can go forward to operational assessment.
2. Content team identified concerns that required modifications. Modifications were clearly identifiable and were likely to improve item performance.
3. Content 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. Reject item. Content team determined the item was not worth revising.

For an item to be accepted as-is, the content teams had to have determined that the item was consistent with DLM item-writing guidelines and the item was aligned to the node. An item/testlet was rejected completely if it was inconsistent with DLM item-writing guidelines, the EE and linkage level were covered by other testlets that had better performing items, or there was not a 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 an item was flagged for modification included items that were incorrectly keyed (i.e., no correct answer or incorrect answer option was labeled as the correct option), items that were misaligned to the node, distractors that could be argued as partially correct options, 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 the testlet would be retained or rejected. Here, the content team could elect to keep the testlet (with or without revision) or reject it. If an edit was to be made, it was assumed the testlet needed retesting. If the testlet included a majority of flagged items without obvious edits that the content team believed could address the problem with the items, the entire testlet was rejected. As a general rule of thumb, DLM field-tests all content prior to making it operational; revised items and testlets were treated as new content.

III.1.J.ii.c Results of Item Analysis and Content Team Review

Table 33 summarizes the number and percentage of items flagged for each field test window. Across both content areas, a total of 515 items (12.2% of total) were flagged in Field Test 1 through Field Test 3, and 999 items (17.0%) were flagged during Phases A through C as needing review by content teams. Items were included in the count of flagged items if they were flagged for one or more criteria. The complete breakdown of items flagged by grade are included in the *Summary of results from the 2014 and 2015 field test administrations of the Dynamic Learning Maps™ Alternate Assessment System* (Clark, Karvonen, & Wells Moreaux, 2015).

Table 33. Item Flags for Content Administered During Field Test 1 through Field Test 3

	ELA			**Mathematics**		
	Count of Flagged Items	**Total Items**	**% Flagged**	**Count of Flagged Items**	**Total Items**	**% Flagged**
Optionally available instructionally-embedded FT 1-3	177	1,925	9.2	338	2,311	14.6
Phase A-C	311	2,531	12.3	688	3,330	20.7

Content teams reviewed all flagged items to determine possible reasons for the flag and whether an edit was likely to resolve the issue.

Table 34 provides the content team accept, revise, and reject counts by content area for all the field test events. In ELA, a total of 79 items were rejected. The ELA content team elected to reject some items outright when the testlet already had four or five items, rather than make edits to one poorly performing item. In mathematics, a total of 74 items were rejected. The complete content team response to item flags by grade and content area can be found in the field test technical report (Clark, Karvonen, & Wells Moreaux, 2015).

Table 34. Content Team Response to Item Flags for Each Field Test Window

Grade	**Flagged Item Count**	**Accept**	**% Accept**	**Revise**	**% Revise**	**Reject**
ELA FT 1-3	177	83	46.9	18	10.2	76
ELA A-C	311	263	84.6	45	14.5	3
Math FT1-3	338	220	65.1	92	27.2	26
Math A-C	688	276	40.1	364	52.9	48

III.1.J.ii.d Educator Survey Results

As part of each field test event, educators were asked to complete a survey about each participating student's experience with the assessment. The surveys were designed to provide the DLM test development team feedback on various aspects of the assessment experience. The complete summary of field test survey results can be found in Clark, Brussow, & Karvonen (2016).

The surveys for Field Tests 1 and 3 were randomly assigned to a subset of educators administering the DLM assessment to their students. The Field Test 2 survey was available to all educators with students participating in the DLM field test. During Field Test 1, a total of 1,402 educators completed surveys for 4,077 students. During Field Test 2, a total of 2,582 educators completed surveys for 7,471 students. During Field Test 3, a total of 1,580 educators completed surveys for 4,166 students.[12]

Survey topics included accessibility, tested content, and training resources. Topics related to tested content are reported in this chapter. For topics related to accessibility, see Chapter IV, and for topics related to training resources, see Chapter X.

Findings Related to Assessment Content

During Field Test 2, which used matrix sampling across multiple linkage levels, educators were asked to evaluate whether the content of the testlet had been taught prior to administering the assessment. Educators reported that most students (58%) had been instructed on the content assessed in the first (lowest linkage level) ELA testlet. Similarly, 60% of students had been instructed on the content of the first mathematics testlet. Slightly fewer students had been instructed on the content assessed on the last testlet (highest linkage level), with a total of 53% for ELA and 50% for mathematics.

During Field Test 3, which assigned three testlets from among five available EEs, a majority of educators (86%) agreed or strongly agreed that the field tested content measured important academic skills for the student being assessed. Similarly, 85% of educators agreed or strongly agreed the field tested content reflected high expectations for the student being assessed.

Educators were also asked to report their views on testlet difficulty during Field Test 3. Testlets administered during Field Test 3 presented items at a single linkage level that was determined from responses to the First Contact survey. Across all four complexity bands, educators indicated most testlets were *about right* for the student. Educators also indicated testlets administered at the foundational (lowest) band were too hard for many students. Table 35 summarizes the reported difficulty levels by student complexity band.

[12] Response rate information is not available due to the method of survey delivery.

Table 35. Educator Reported Testlet Difficulty

Band	Too Easy		About Right (%)		Too Hard (%)	
	n	%	*n*	%	*n*	%
ELA Foundational	11	4	119	46	128	50
ELA Band 1	56	12	298	62	122	26
ELA Band 2	92	18	379	73	50	10
ELA Band 3	64	23	191	67	27	10
Math Foundational	17	5	182	54	137	41
Math Band 1	86	17	340	67	85	17
Math Band 2	142	24	383	66	58	10
Math Band 3	53	30	113	63	7	7

During Field Test 1, educators were asked to rate the *text* complexity for the ELA testlets administered to each student. Table 36 shows educator perceptions of text complexity for content assessed in Field Test 1. Most educators reported that the text was of appropriate complexity for the student taking the DLM assessment. Approximately 35% of educators reported that the text was too complex for the student taking the test. Because of the matrix-sampling approach used during Field Test 1, a range of text complexity was presented to each student, so it is likely that some were at a higher level than the student would ordinarily be administered.

Table 36. ELA Field Test 1 Text Complexity

Resource	*n*	%
Not complex enough	384	9.4
Appropriate complexity	2,303	56.2
Too complex	1,412	34.5

III.1.K. Operational Assessment Items for 2014-15

Operational assessments were administered during the spring window. Table 37 gives the participation numbers. One test session is one testlet taken by one student. Only test sessions that were complete or in progress at the close of the window counted towards the total test sessions by model.

Table 37. Operational Window Participation

Participation	*N*
Test Sessions	591,814
Students	50,080
Educators	13,187
Schools	7,467
Districts	2,410

Participation by grade level ranged from 606 students in twelfth grade to 6,874 students in sixth grade.

Testlets were made available for operational testing following promotion from field test item review. Table 38 and Table 39 give the total number of operational testlets by content area for 2015. There were a total of 669 operational testlets available across grades and content areas. This also included 701 EE/linkage level combinations that had more than one testlet available during an operational window due to having both a braille and general version of the testlet available.

Table 38. 2014–15 ELA Operational Testlets

Grade	*n*
3rd	38
4th	41
5th	36
6th	34
7th	29
8th	24
9th	29
10th	26
11th	29
9th–10th	n/a

Grade	*n*
11th–12th	n/a
Grand Total	286

Table 39. 2014–15 Mathematics Operational Testlets

Grade	*n*
3rd	41
4th	50
5th	42
6th	33
7th	38
8th	37
9th	37
10th	34
11th	71
9th–12th	n/a
Grand Total	383

Similar to the field test item review, *p*-values were calculated for all operational items to provide information about item difficulty. Figure 36, Figure 37, and Figure 38 include the *p*-values for each operational item for ELA and math. The sample size cutoff for inclusion in the *p*-values plots that follow was 20, to prevent items with small sample size from potentially skewing the results. In general, ELA items were easier than the math items, as evidenced by more items falling in the higher bin ranges. Writing items are omitted from this plot due to scoring occurring at the option level rather than item level.

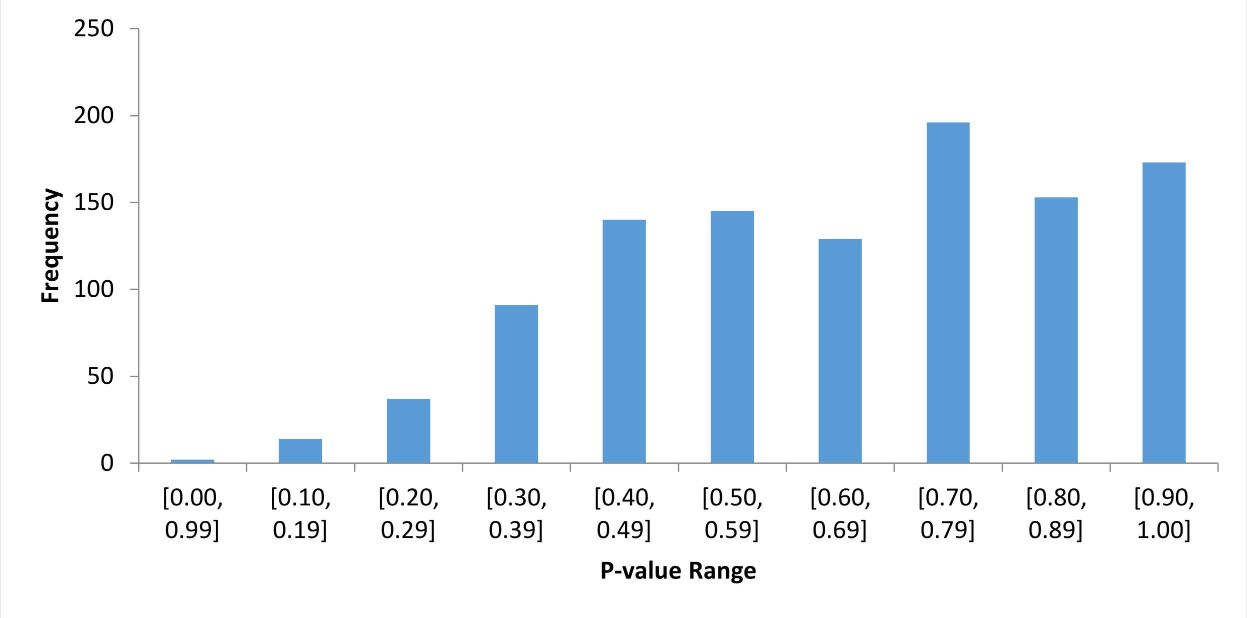


Figure 36. *P-value for ELA operational items. Writing items and items with a sample size less than 20 were omitted.*

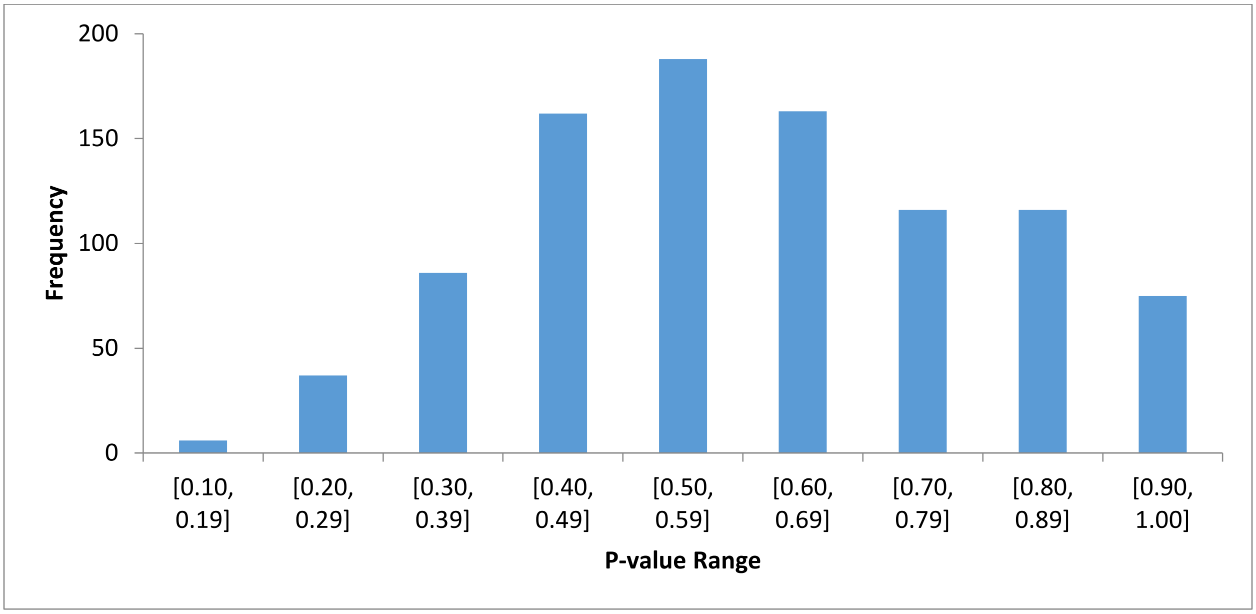


Figure 37. *P-value for mathematics operational items. Items with a sample size less than 20 were omitted.*

Standardized difference values were also calculated for all operational items with a sample size of at least 20. However, due to the modeling approach used for generating operational scores, the standardized difference values were calculated to compare the *p*-value for the item to all other items measuring the EE and linkage level, rather than by node, as they were for field test item review. Figure 36, Figure 37, Figure 38, and Figure 39 summarize the standardized difference values for operational items. Most items fell within two standard deviations from the mean for the EE and linkage level.

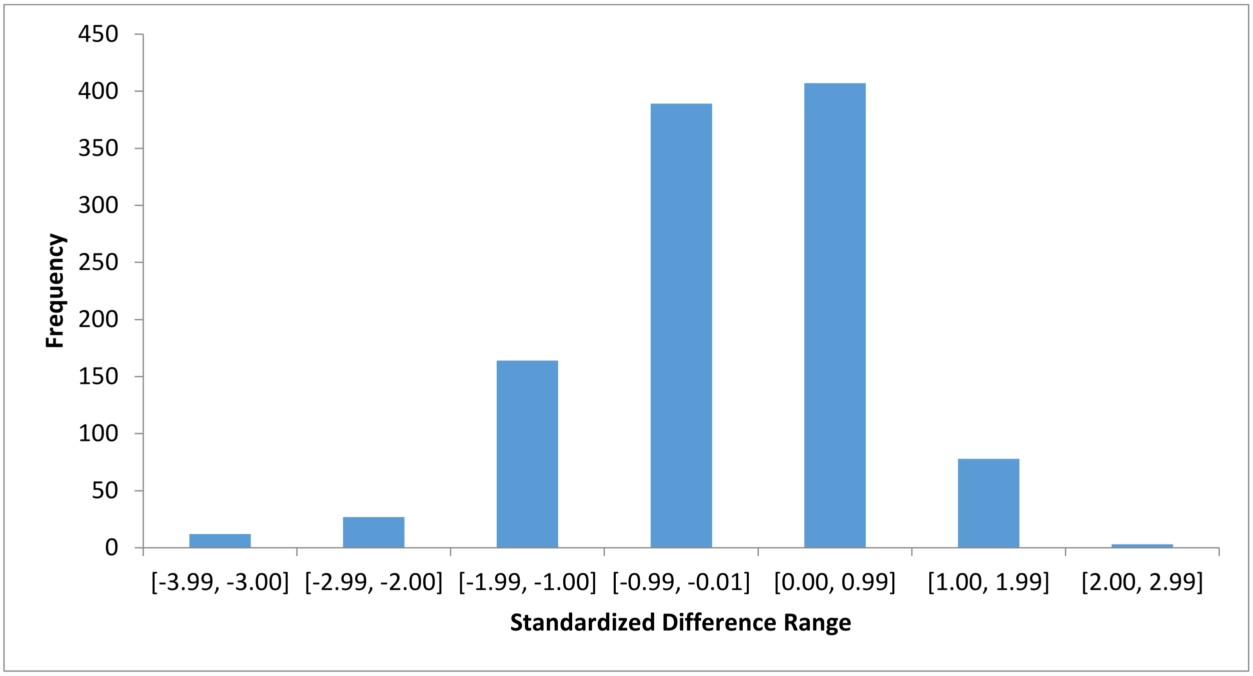


Figure 38. *Standardized difference z scores for ELA operational items. Items with a sample size less than 20 were omitted.*

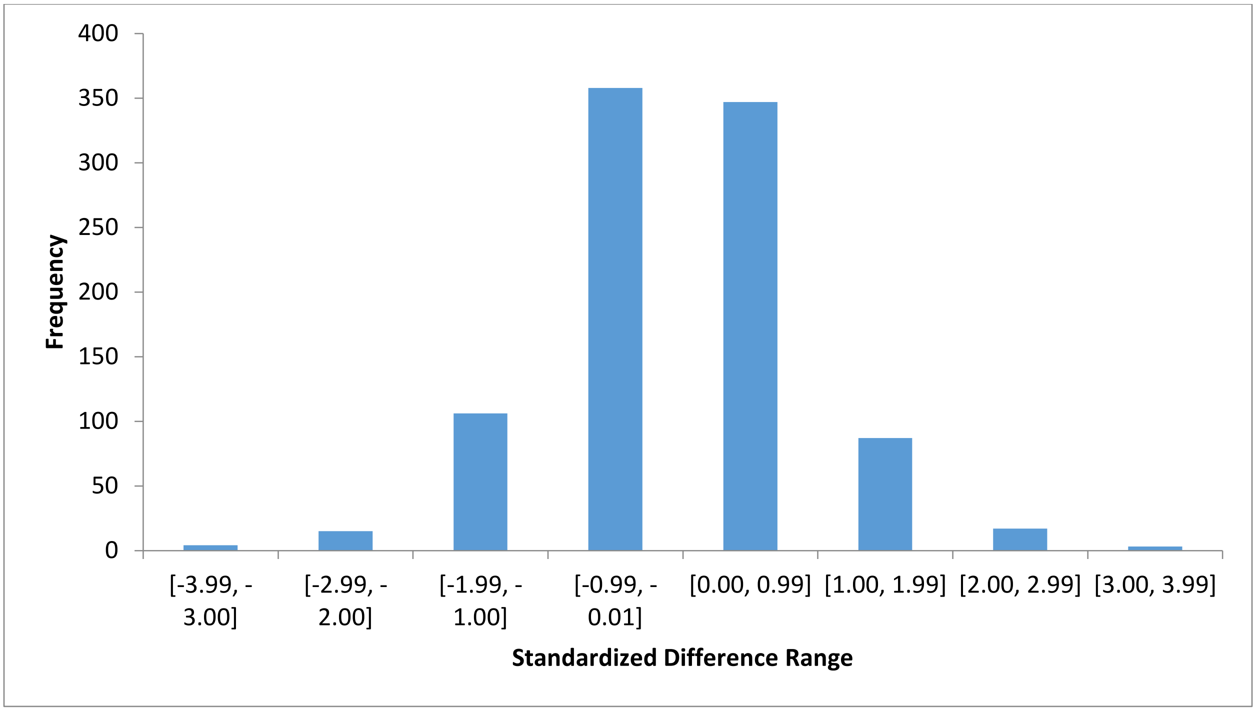


Figure 39. *Standardized difference z scores for mathematics operational items. Items with a sample size less than 20 were omitted.*

For information on a summary of the total linkage levels mastered during operational testing and the distribution of students by performance level, see Chapter VI.

IV. TEST ADMINISTRATION

Chapter IV presents the processes and procedures used to administer the Dynamic Learning Maps® (DLM®) alternate assessments in 2014–2015. As described in earlier chapters, the DLM Consortium developed adaptive computer-delivered alternate assessments that provide the opportunity for students with the most significant cognitive disabilities to show what they know and are able to do in mathematics and English language arts (reading and writing) in grades 3-12.[13] The DLM assessments are based on DLM maps, highly connected representations of how academic skills are acquired, as demonstrated in research literature. Assessment blueprints are composed of Essential Elements (EEs), which are alternate content standards that describe what students with the most significant cognitive disabilities should know and be able to do at each grade level. The 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.

First, Chapter IV provides an overview of the key features of test administration. In this overview, we explain how students are assigned their first testlet using the First Contact survey results. The chapter also describes testlet formats (computer-delivered and teacher-delivered) and the assessment window. Sections that follow define test administration protocols, accessibility tools and features, test security, and evidence of educator and student experiences with test administration in 2014–2015.

IV.1. Overview of Key Administration Features

Consistent with the DLM Theory of Action described in Chapter I, the DLM test administration features reflect the 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 and prior performance. Based on students' support needs, some DLM assessments are designed to be administered in a one-to-one, student/test administrator format. Most test administrators are the special education teachers of the students, as they are best equipped to provide the most conducive conditions to elicit valid and reliable results. Test 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 test content and materials and to limit construct-irrelevant variance.

This section describes the key, overarching features of the DLM test administration. First, we explain the year-end assessment model, which yields summative results based on spring assessments which cover the test blueprints. Next, we describe the two assessment delivery modes and the online testing platform, the KITE™ system. Finally, we describe the system-

[13] Specific high school grades required are determined by each state.

driven adaptive delivery that determines the linkage levels of testlets assigned during the spring assessment window.

IV.1.A. The Year-end Assessment Model

As briefly described in Chapter I, there are two variations on the DLM assessment system. This manual supports the year-end assessment model, which is described here.

In the year-end assessment model, the DLM system is designed to assess a student's learning consistent with the theory of action (see Chapter I). The year-end model uses testlets that assess one or more EEs delivered in the spring of each year. Additional optionally available, instructionally-embedded assessments are available throughout the year, but since results from these assessments are not used for accountability purposes and programs, they are not addressed in this manual. In the year-end model all students are assessed during the spring window on the entire breadth of the blueprints in each content area.

IV.1.A.i. Assessments

The DLM alternate assessments are delivered in testlets. In reading and math, testlets are based on nodes for one or more EEs. Each testlet contained an engagement activity and three to nine items. Writing testlets covered multiple EEs. In the spring testing window, students received as few as five and as many as seven testlets, depending on the grade and subject. The system delivered only one testlet at a time in each subject. The system used the First Contact survey information to initiate the first testlet assigned in both ELA and mathematics. After the student took the first testlet, the system delivered the next testlet. The student's performance on the previous testlet determined how the system selected and delivered the second testlet. An explanation of the selection procedures that assigned the first and subsequent testlets is described in the Adaptive Delivery section in this chapter.

IV.1.A.ii. Calculation of Summative Results

Summative results are based on student responses on testlets and information about the structure of the DLM map. Together, this information is used to determine which linkage levels the student has likely mastered. Results for each linkage level are determined based on the probability that the student has mastered the skills at that linkage level (see Chapter V for a full discussion of modeling).

Summative results are determined from the linkage level mastery data. The information about each linkage level leads to a summary of the student's mastery of skills in each conceptual area and for the subject overall. See Chapter VII for a full description of how summative results are calculated.

IV.1.B. Assessment Delivery Modes

The DLM system includes testlets designed to be delivered via computer directly to the student and testlets designed for the teacher to administer outside the system and record responses in the system. The majority of testlets were developed for the computer-administered mode because evidence suggests that 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 test on the computer with navigation assistance from a test administrator (Nash, Clark, & Karvonen, 2015). Teacher-administered testlets, designed for teacher delivery, included all testlets at the Initial Precursor linkage level, all writing testlets, some higher linkage level mathematics testlets requiring manipulatives, and some alternate forms for students who are blind or who have visual impairments. The 2014–2015 operational testlet pool was comprised of 66.4% computer-delivered testlets and 33.6% teacher-administered testlets.

IV.1.B.i. Computer-Delivered Assessments

Most DLM alternate assessments could be delivered directly to students by computer through the KITE system. 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.[14]

The computer-delivered testlets included various item types, including single-select multiple choice with three response options and text or images as answer choices, multi-select multiple choice with text or images as answer choices, matching items from two lists, sorting objects into categories, and highlighting selected text. See Chapter III for more information about item types.

IV.1.B.ii. Teacher-Delivered Assessments

Some testlets were designed to be administered directly by the test administrator outside of the KITE system. The KITE system delivered the test, but the test administrator played a more direct role than in computer-delivered testlets. In teacher-administered testlets, the test administrator was responsible for setting up the assessment, delivering it to the student, and recording responses in the DLM system.

There were three general categories of teacher-administered testlets.

1. Testlets with content designed for students who are developing symbolic understanding or who may not yet demonstrate symbolic understanding (Initial Precursor).[15]

[14] For students who cannot interact independently with the computer, test administration procedures allow for the student to indicate a response through any mode of expressive communication and for the test administrator to enter the response on the student's behalf. See the Accessibility section in this chapter for details.

[15] These testlets tend to occur at lower linkage levels, and the test administrator must be very familiar with the student’s typical modes of expressive communication.

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 were often used in this case, especially for students with blindness or visual impairment.
3. All writing assessments.

All three types of teacher-administered testlets had some common features. See Chapter III for a description of the structure of teacher-delivered testlets.

IV.1.C. THE KITE SYSTEM

The DLM alternate assessments are managed and delivered using the Kansas Interactive Testing Engine (KITE) platform, which was designed and developed to meet the needs of the next generation of large-scale assessments. The KITE system consists of four applications. Teachers and students see two of these applications: Educator Portal and KITE Client (Test Delivery Engine). The KITE system has been 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) Profile and World Wide Web Consortium Web Content Accessibility Guidelines in the KITE Client. Minimum hardware and operating system requirements for KITE and supported browsers for Educator Portal are published on the DLM website and in the *DLM Technical Liaison Manual* linked on each state's DLM webpage.

IV.1.C.i. Educator Portal

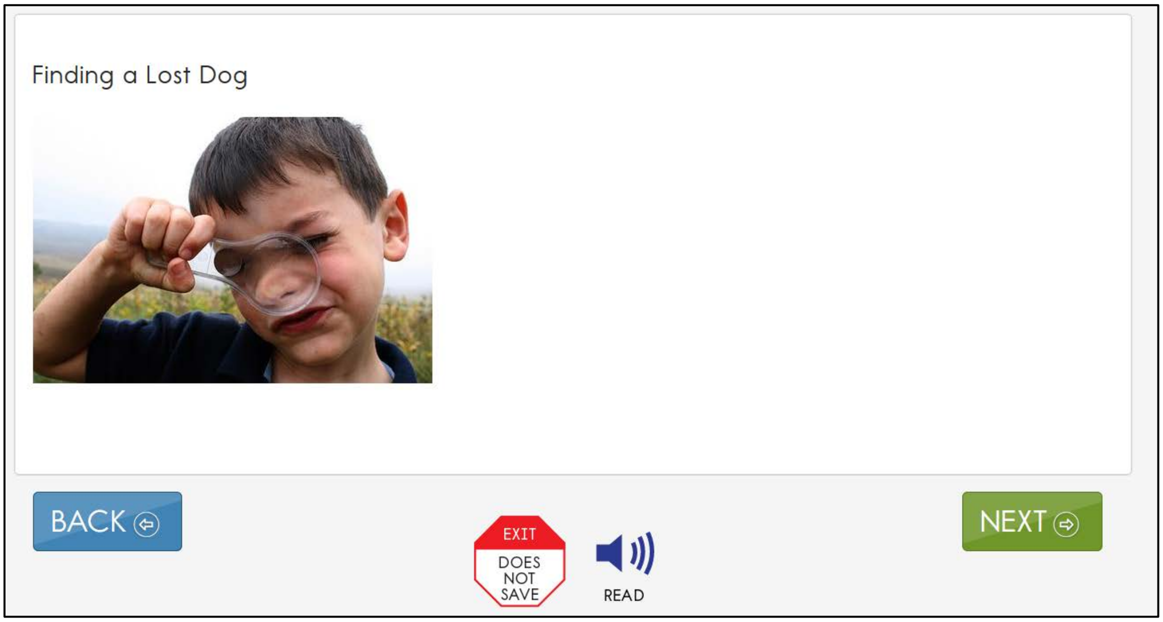
Educator Portal is the administrative application where staff and educators manage student data, complete required test administration training, assign optional instructionally embedded assessments, retrieve resources needed for each assigned testlet, and retrieve reports.

- Test administrators, usually teachers, use Educator Portal to manage all student data. They are responsible for checking class rosters of the students who are assigned to take DLM alternate assessment testlets and for completing the PNP and First Contact survey for each student.
- Educator Portal hosts the required test administrator training modules (since 2014–2015). Teachers complete facilitated or self-directed training and take post-tests to demonstrate their understanding of the material. (See Chapter X for more information.)
- After each testlet is assigned to a student, the system delivers a Testlet Information Page (TIP) through 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 the Resources and Materials section of this chapter for more information.)

IV.1.C.ii. KITE Client (Test Delivery Engine)

The KITE Test Delivery Engine (TDE) is the portal that allows students to log in and complete assigned testlets. Practice activities and released testlets are also available to students through TDE. Students access TDE via KITE Client, a customized version of Firefox, which launches in kiosk mode and prevents students from accessing unauthorized content or software while taking secure, high-stakes assessments. The TDE interface is supported on desktops and laptops running Windows or OS X, on Chromebooks, and on iPad tablets.

The KITE system 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 within a testing session, and minimizes the cognitive load related to test navigation and response entry. An example of a screen used in an English language arts testlet is shown in Figure 40. 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 is provided in the Accessibility section in this chapter.



Finding a Lost Dog

Figure 40. An example screen from the student interface in KITE.

IV.1.C.iii. 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 Client in an internal database; therefore, if the upstream network connection becomes unreliable or variable during testing, students can still continue testing, and their responses will be 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 in order to ensure the next testlet is delivered correctly.

IV.1.D. Adaptive Delivery

As discussed in Chapter III, the DLM assessments are delivered in testlets. In reading and mathematics, items in a testlet are aligned to nodes at one of five linkage levels for one or more EEs. Writing testlets cover multiple EEs and are delivered at one of two levels: Emergent (which corresponds with Initial Precursor and Distal Precursor) or Conventional (which corresponds with Proximal Precursor, Target, and Successor linkage levels). While blueprints determine the EEs that are selected for assessment, the adaptive delivery mechanism determines the linkage level for each testlet assigned to students. The linkage level of the first assigned testlet in both ELA and mathematics was determined based on teacher responses to the First Contact survey, which is an inventory of learner characteristics in a variety of areas, including communication and academic skills. Three sections of the First Contact survey were used to provide an optimal match between student and testlet during the initial DLM testing experience: Expressive Communication, Reading Skills, and Math Skills. First Contact survey items used for initialization purposes are included in Appendix C1. Based on the teacher's responses, the student's assigned complexity band was automatically calculated and stored in the system.

- For the English language arts (reading and writing) testlets, the KITE system used the responses from the Expressive Communication and Reading Skills questions to assign a student's complexity band.
- The KITE system used the responses from the Expressive Communication and Math Skills questions when calculating a complexity band for the mathematics assessment.
- For either subject, if a different complexity band was indicated between the two sets of questions (Expressive Communication and the subject area questions), the system selected the lower band. The goal was to present a testlet that is approximately matched to a student's knowledge, skills, and abilities. That is, within reason, the system should have presented a testlet that was neither too easy nor too difficult and that provided a positive experience for the student entering the assessment.

Research supporting the use of this algorithm for classifying students to complexity bands is summarized in the Pilot – Initialization section of Chapter III.

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 40.

Table 40. 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 AAC; does not read any words when presented in print (ELA); or does not sort objects (math)	Foundational	Initial Precursor
Uses one word, sign, or symbol to communicate; recognizes symbols (ELA) or sorts symbols (math)	Band 1	Distal Precursor
Uses 2 words, signs, or symbols to communicate; reads at the primer to second grade level (ELA); 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 (ELA) or regularly add/subtract and form groups of objects (math)	Band 3	Target

Note: AAC = augmentative or alternative communication device; ELA = English language arts.

The educator must complete the student's First Contact survey before assessments are delivered. Supporting procedures and a complete list of First Contact survey questions are included in the *Test Administration Manual 2014–2015* (Dynamic Learning Maps, 2014). Test administrators are trained on the role of the First Contact survey in testlet assignment as part of required test administrator training (see Chapter X). Each student was assigned as few as five to as many as seven testlets per subject during the spring window. The system determined the linkage level for each testlet. The assignment was adaptive between testlets. Each spring testlet was packaged and delivered separately, and the test administrator determined when to schedule each testlet within the larger window. See Spring Operational Assessments (Dynamic Learning Maps, 2014, p. 73) for more detail.

The second and subsequent testlets were assigned based on previous performance. That is, the linkage level associated with the next testlet a student received was based on the student's performance on the previously administered testlet. The goal was to maximize the match of student knowledge, skill, and ability to the appropriate linkage level content. Specifically:

- The system adapted up one linkage level if students responded correctly to 80% or more of the items measuring the previously tested EE. If testlets are already at the highest level (i.e., Successor), they remain there.

- The system adapted down one linkage level if students responded correctly to less than 35% of the items measuring the previously tested EE. If testlets are already at the lowest level (i.e., Initial Precursor), they 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, <.35), the linkage level of the testlet is likely too challenging. To provide a better match to the student's knowledge, skills, and ability, 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 ability being assessed at that linkage level. Therefore, the student is neither routed up nor down for the subsequent testlet.

Figure 41 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 of the items correctly, so the next testlet was assigned at the Proximal Precursor level. The next two testlets adapted up or down a level, whereas the fifth testlet remained at the same linkage level as the previous testlet.

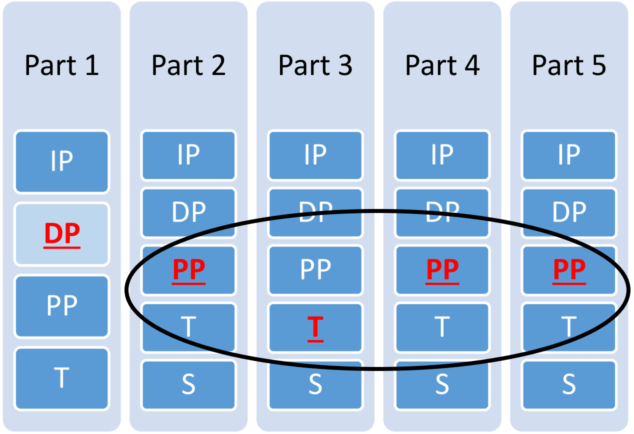


Figure 41. Linkage level adaptations for a student who completed five testlets.

Note: IP = Initial Precursor; DP = Distal Precursor; PP = Proximal Precursor; T = Target; S = Successor.

IV.2. Test Administration

This section overviews general test administration processes and procedures. For more detail, see the *Test Administration Manual 2014–2015* (Dynamic Learning Maps, 2014). Test administration guidelines provide teachers with the information necessary to administer the assessments with fidelity and for students to demonstrate their knowledge and skills at appropriate breadth, depth, and complexity of the content.

IV.2.A. Test Windows

During the consortium-wide spring testing window, which occurred between March 16 and June 12, 2015, all students were assessed on the EEs on the blueprint in both ELA and mathematics. Each state set its own testing window within the larger consortium window.

IV.2.B. Administration Time

During the spring testing window, the estimated total testing time was 60–75 minutes per student in English language arts and 35–50 minutes in mathematics.

The published estimated total testing time per testlet averaged 5–10 minutes in mathematics, 10–15 minutes in reading, and 10–20 minutes for writing. Published estimates were slightly longer than anticipated real testing times because of the assumption that teachers would need time for setup. Actual testing time per testlet varied depending on each student's unique characteristics.

The KITE system captured start and end dates and time stamps for every testlet. To calculate the actual testing time per testlet, the difference between these start and end times was calculated for the spring 2015 operational administration. As the KITE system was still in development, the 2015 time-stamp data included some impossible values (i.e., negative times, values greater than 24 hours). Implausible values comprised 5% of the data.

Table 41 the distribution of test times per testlet after removing negative values and test times greater than 8 hours (i.e., approximate maximum length of a school day). Given the wide range of testlet response times (up to 8 hours), the interquartile range values most likely describe the typical range of testing time per testlet. Most testlets took around 5 minutes or less to complete, with mathematics testlets generally taking less time than English language arts testlets.

Table 41. Distribution of Response Times in Minutes – Year-End Model

Subject/Grade	Min	Max	Mean	Median	25%Q	75%Q	IQR
ELA							
3	0.27	450.68	11.22	5.42	3.67	8.68	5.02
4	0.35	452.48	10.29	5.42	3.73	8.40	4.67
5	0.23	477.17	10.27	5.65	4.03	8.53	4.50

Subject/Grade	Min	Max	Mean	Median	25%Q	75%Q	IQR
6	0.27	452.03	10.07	5.63	4.05	8.52	4.47
7	0.25	434.42	9.63	4.97	3.52	7.85	4.33
8	0.28	400.03	9.87	5.78	4.23	8.62	4.38
9	0.47	419.88	9.79	5.53	3.95	8.52	4.57
10	0.40	418.13	10.07	5.70	4.08	8.80	4.72
11	0.42	460.08	12.34	6.42	4.50	10.08	5.58
Math							
3	0.23	448.38	9.14	3.00	1.65	6.52	4.87
4	0.13	412.67	7.60	2.38	1.48	4.45	2.97
5	0.25	464.20	8.60	2.90	1.75	5.40	3.65
6	0.22	448.97	6.63	2.83	1.82	4.80	2.98
7	0.13	459.92	6.34	2.37	1.55	4.37	2.82
8	0.15	402.80	6.35	2.53	1.68	4.35	2.67
9	0.18	468.65	6.46	2.57	1.40	4.87	3.47
10	0.15	413.98	6.53	3.02	1.67	5.38	3.72
11	0.13	433.95	8.28	2.93	1.65	5.65	4.00

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

Time per testlet may have been impacted by student breaks during the assessment. In the DLM Test Administration Manual, test administrators were encouraged to allow students to take breaks in the case of fatigue, disengagement, or behavioral problems that are likely to interfere with a valid assessment of what the student knows and can do. The KITE system allowed for up to 90 minutes of inactivity without timing out to allow teachers and students to pause for breaks during administration of a testlet. In cases in which administration had begun but the student was unable to engage and respond for any reason and a short break was not sufficient, the **EXIT DOES NOT SAVE** button could be used to exit the testlet, allowing the teacher and student to return to it at another time.

Test administrators were asked to estimate the amount of time it took to administer a testlet both in English language arts reading and writing and in mathematics. This information included teacher judgment of the total amount of time spent on administration of each type of testlet, including setup, student engagement, navigation, and where applicable, preparation of materials needed for administration. Table 42 shows the number and percentage of test administrators reporting administration times of less than 10 minutes, 10–20 minutes, and more than 20 minutes for each type of testlet. English language arts testlets, in general, took more time to administer than mathematics testlets. Results were generally consistent with the published estimates provided in the Test Administration Manual.

Table 42. Teacher-Reported Length of Time to Administer a Testlet by Subject

Subject	Time	*n*	%
Reading	Less than 10 minutes	1010	35.6
	10–20 minutes	1558	54.9
	More than 20 minutes	272	9.6
Writing	Less than 10 minutes	947	36.2
	10–20 minutes	1235	47.2
	More than 20 minutes	433	16.6
Mathematics	Less than 10 minutes	1252	44.6
	10–20 minutes	1344	47.8
	More than 20 minutes	214	7.6

IV.2.C. Resources and Materials

Test administrators, school staff, and Individualized Education Program (IEP) teams had multiple resources throughout the test administration process, some of which were required consortium wide. Some states provided additional materials on their websites. The DLM website provided resources that covered DLM background and assessment administration training information; student and roster data management; test delivery protocols and setup; accessibility features, protocols, and documentation; and practice activities. This section provides an overview of all resources and materials for test administrators as well as more detail regarding the critical resources of TIPs and Practice Activities and Released Testlets.

IV.2.C.i. The DLM Website

The DLM website served as a way to communicate assessment information to educators. Pages such as Essential Elements, Accessibility, and Test Development covered topics related to the DLM system as a whole and may be of interest to a variety of audiences. To support assessment administration, each state also had its own customized landing page with an easy-to-remember URL (i.e., dynamiclearningmaps.org/statename). For an example see Appendix C.9. Through training, manuals, webinars, and replies from Service Desk inquiries, educators were made aware of their state-specific webpage to locate consortium-level resources and state-customized resources.

To provide consortium-wide updates and reminders, the DLM website also featured a Test Updates webpage. This was a newsfeed-style page that addresses timely topics such as assessment deadlines, resource updates, or system status. Additionally, the Test Updates page offered educators the option to subscribe to an electronic mailing list to automatically receive the same message via email without visiting the website.

IV.2.C.ii. Test Administration Resources

The DLM website provided specific resources designed for test administrators. These resources were available to all states (Table 43) to ensure consistent test administration practices.

Table 43. DLM Resources for Test Administrators and States

Test Administration Manual (PDF)	**Supports the test administrator in preparing themselves and students for testing.**
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.
Educator Resource Page (Webpage)	Includes additional resources for test administrators, such as familiar texts, materials collection, and testlet overview videos, tested Essential Elements and their associated mini-maps.
Guide to DLM Required Training and Professional Development 2014-15 (PDF)	Helps users access DLM Required Test Administration Training and instructional professional development in Educator Portal.
Guide to Practice Activities & Released Testlets	Supports the test administrator in accessing practice activities in KITE Client.
Test Updates Page (Webpage)	Breaking news on test 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 Required Test Administration Training modules.

IV.2.C.iii. District-Level Staff Resources

Resources were available for three district-level supporting roles: Assessment Coordinator, Data Steward, and Technical Liaison. The Assessment Coordinator oversaw 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 Steward managed educator, student, and roster data. The Technical Liaison verified that the network and testing devices were prepared for test administration.

Resources for each of these roles were made available on the state's customized DLM webpage. Each role had its own manual, a webinar, and a FAQ compiled from webinar questions. Each role was also guided to supporting resources for other roles where responsibilities overlapped. For example, Data Stewards were also guided to the TEST ADMINISTRATION MANUAL to support data-related activities assigned to the test administrator and connected to troubleshooting data issues experienced by the test administrator. Technical Liaisons were also guided to the KITE and Educator Portal webpage for information and documents connected to KITE Client, Local Caching Server use, supported browsers, and bandwidth requirements. Assessment Coordinators were also guided to resources developed for the Data Steward, Technical Liaison, and test administrator for specific information and supplemental knowledge of the responsibilities of each of those roles. Some of those resources include:

- GUIDE TO DLM REQUIRED TRAINING & PROFESSIONAL DEVELOPMENT 2014-15
- TEST ADMINISTRATION MANUAL 2014–2015
- Field Test webpage
- Test Updates webpage and electronic mailing list

Descriptions of the district-level role webinars are provided in Chapter X.

IV.2.C.iv. Testlet Information Pages

TIPs provided test administrators with information specific to each testlet. Test administrators received a TIP page after each testlet was assigned to a student, and they were instructed to review the TIP before beginning the student's assessment. (See the sample TIP in Appendix C.2.)

Each TIP stated whether a testlet was computer-delivered or teacher-administered and indicated the number of items on the testlet. The TIP also provided information for each testlet regarding the materials needed or any substitute materials allowed.

The TIP provided information on the exceptions to allowable supports. While a test administrator typically used all appropriate PNP features and other flexibility tools described in the Allowable Practices section of the Test Administration Manual, the TIP indicated when it was not appropriate to use a support on a specific testlet. This may have included limits on the use of definitions, translation, read aloud, or other supports.

If there were further unique instructions for a given testlet, they were provided in the TIP. For test administrators who delivered human read aloud that includes description of graphics, alternate text descriptions of images were provided as additional pages after the main TIP.

TIPs for English language arts testlets also provided the name of a given text, identify the text as informational or literature, and label the text as familiar or unfamiliar. TIPs included the name of the grade-level text that the DLM text was associated with and noted if test administration time was expected to be longer than usual because there were two texts (when the linkage level required a comparison between two texts). TIPs for mathematics testlets had information on specific math terminology used in the testlet and whether calculator use was appropriate.

Testlets that required special set up before test administration begins, such as some math testlets designed for students with blindness or visual impairments, had additional pages of instructions.

IV.2.C.v. Practice Activities and Released Testlets

Two practice activities and many released testlets were made available to support educators and students preparing for testing.

- The practice activities were designed to familiarize users with the way testlets and item features look in the KITE system. One activity was for teachers and the other was for students.
- The released testlets were similar to real DLM testlets in content and format.

Practice activities and released testlets were accessed through KITE in the practice section. Using login information provided by the system, both types of activities could be completed as many times as desired.

The teacher practice activity was a tutorial about testlets administered directly by the teacher. In this tutorial, teachers were instructed on how to read the instructions on the screens and follow them and how to enter the student's responses to activities or exchanges that occur outside the system. Most of these testlets required teachers to gather materials to be used in the assessment. Directions for how to prepare for the testlet were provided as Educator Directions on the first screen.

The student practice activity was a tutorial about testlets that are administered directly to the student. The student practice activity provided an opportunity for students to become familiar with navigation in the KITE system, the types of items used in DLM assessments, and the method for indicating responses to different item types.

Released testlets are similar to operational testlets. They are selected from a variety of EEs and linkage levels across grades 3-12. New released testlets are added periodically and include teacher-administered testlets and computer-delivered testlets.

IV.2.D. Test Administrator Responsibilities and Procedures

Procedures for test administrators were organized into four sets of tasks for different parts of the school year: (1) before beginning assessments, (2) spring window assessment, and (3) preparing for next year. The *Test Administration Manual* (Dynamic Learning Maps, 2014) provided detailed description of each set of tasks with specific resources to support the work.

IV.2.D.i. Before Beginning Assessments

Test administrators performed multiple steps to prepare for student testing. They confirmed student eligibility to participate in the DLM alternate assessments and shared information about the assessments with parents to prepare them for their child's testing experience. Test administrators reviewed the entire Test Administration Manual and became familiar with all available resources, including state webpages, practice testlets and available content to be assessed, and procedures for preparing to give the assessment.

Preparation included preparing for the computer-delivered aspects of the assessment system. Test administrators had to gain access to Educator Portal, activate their KITE account, complete the security agreement in their Educator Portal profile, and complete their required test administration training (see Chapter X). Test administrators also reviewed their state's guidance on required and recommended professional development modules.

Preparation also involved reviewing the *Accessibility Manual* (Dynamic Learning Maps, 2014) and working with the IEP team to determine what accessibility supports should be provided for each student taking the DLM assessments. Test administrators recorded the chosen supports in the PNP in Educator Portal and review their state's requirement for documentation of the DLM accessibility supports as testing accommodations, adjusting the testing accommodations in the IEP as necessary.

Additional preparations involved preparing student data, including a review of student demographic information and roster data in Educator Portal for accuracy. Test administrators ensured that the PNP and the First Contact survey was updated and complete in Educator Portal. School staff installed KITE Client on testing devices and familiarized both teachers and students with DLM testlets through practice activities and released testlets. Finally, student devices were checked for compatibility with KITE Client.

IV.2.D.ii. During Spring Window Assessment

The spring assessment procedures also included checking student demographic information, PNP settings, and First Contact survey responses. School staff members considered the district and school assessment schedules to ensure students could complete all DLM testlets during the spring window, and then they scheduled assessment session locations and times.

Test administrators retrieved TIPs for the assigned first testlet and gathered materials needed before beginning testing. After retrieving student usernames and passwords from Educator Portal, test administrators assessed each student with the first testlet. As each remaining testlet became available, they retrieved TIPs, gathered materials as needed, and assessed the student.

IV.2.D.iii. Preparing for Next Year

With IEP teams, teachers evaluated students' accessibility supports (PNP settings) and made decisions about supports and tools for next year. With IEP teams, they reviewed the blueprint for the next grade as one source of information to plan academic IEP goals.

IV.2.E. MONITORING ASSESSMENT ADMINISTRATION

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

IV.2.E.i. Consortium Test Administration Observation Protocol

The DLM Consortium developed a test administration observation protocol (see Appendix C.12) to standardize data collection across observers and locations. 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 (see Chapter IX for 2014–2015 results).

One observation form is completed per testlet administered. Some items are differentiated for computer-administered and teacher-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 provides 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 system, 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 used during the testing session. Finally, Observer Evaluation requires that the observer rate student overall engagement during the session and provide any additional relevant comments.

The protocol was available as a PDF to be printed for handwritten observations and as an online survey (optimized for mobile devices and with branching logic) to support electronic data collection.

Training resources were provided to state education agency staff to support fidelity of use of the test administration protocol and to increase the reliability of data collected (Table 44). State education agency staff had access to the *Test Administration Observation Training* video (see Appendix C.13) on the use of the *Test Administration Observation Protocol*. The links to this video, the *Guidance for Local Observers (*see Appendix C.13), 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 were able to use these training resources to encourage use of the protocol among local education agency staff. States were also cautioned that the protocol was only to be used to document observations for the purpose of describing the administration process. It was not to be used for evaluating or coaching teachers 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 44. DLM Resources for Test Administration Monitoring Efforts

DLM Test Administration Observation Research Protocol (PDF)	**Provides observers with a standardized way to describe the test 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.

IV.2.E.ii. Formative Monitoring Techniques

The consortium used several techniques for formative monitoring purposes in 2014–2015. First, because DLM assessments are delivered as a series of testlets, a test administration monitoring extract was available on demand in 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 lists the number of testlets completed and expected for each subject. To support local capacity for monitoring, webinars were delivered in February and March 2015 before the spring testing window opened. These webinars targeted district and school personnel who monitor assessments and had not yet been involved in DLM assessments (see Appendix C.14).

Formative monitoring also occurred through regular consortium calls including DLM staff and state partners. Throughout most of the year, these calls were scheduled twice per month. Topics related to monitoring that regularly appeared on agendas for partner calls included assessment window preparation, anticipated high-frequency questions from the field, and opportunities 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 window, check-in calls were hosted on the weeks between the regularly scheduled partner calls. The purpose of the check-in calls was to keep state partners apprised of any issues or concerns that arise during the testing window allowing them to provide timely information to districts. States are provided with a description of the issue 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.

IV.2.E.iii. Monitoring Testlet Delivery

Prior to the opening of a testing window, Agile Technology Solutions staff initiated an automated enrollment process that assigns the first testlet. Students who had missing or incorrect information in Educator Portal, preventing testlet assignment, 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 by Agile Technology Solutions staff. Once the student completed the first testlet, the adaptive delivery component of the KITE system drove the remaining testlet assignments. This process also generated error logs that could be accessed 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, or when the adaptive delivery system did not work as intended, DLM staff worked with state partners to either communicate general reminders to the field or solve problems regarding specific students.

During each operational 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 appear as expected by grade, state, and testing model and verifying correct assignment to initial testlet-based rules that govern that process. In addition, a script was run to verify student routing through the system occurred as expected, whereby students routed to the correct linkage level for each subsequent testlet based on the algorithm described earlier in this chapter in the section called Test Administration.

IV.3. ACCESSIBILITY

The DLM System was designed to be optimally accessible to diverse learners through accessible content (see Chapter III), initialization, and routing driven by First Contact survey and prior performance (Chapters III and IV) and supported by a straightforward user interface in the KITE system (Overview of Key Administration Features section, above). Consistent with the DLM map and item and test development practices described in earlier chapters (see Chapters II and III), principles of universal design for assessment were applied to design the test 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 were made available for use in the DLM assessment system. The Accessibility Manual (Dynamic Learning Maps, 2014) outlined a six-step process for test administrators and IEP teams to use in making decisions about accessibility supports. This process began with confirming the student meets the DLM participation guidelines (see Appendix C.16) and continued with the selection, administration, and evaluation of the effectiveness of the accessibility supports. Supports were selected for each student in the PNP in KITE Educator Portal. The PNP could be completed any time before testing begins. It could also be changed during testing as a student's needs change. Once updated, the changes appeared the next time the student is logged in to the KITE system. All test administrators were trained in the use and management of these features (see Chapter X).

IV.3.A. Overview of Accessibility Supports

Accessibility supports considered appropriate for use during administration of computer-delivered or teacher-delivered testlets were listed in the *Accessibility Manual* (Dynamic Learning Maps, 2014). 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 were grouped into three categories: those provided through the PNP, those requiring additional tools or materials, and those provided outside the system. Supports are listed in each of these categories in Table 45.

Table 45. Accessibility Supports in the DLM Assessment System

Supports Provided via PNP	**Supports Requiring Additional Tools/Materials**	**Supports Provided Outside the System**
• Magnification • Invert color choice • Color contrast • Overlay color	• Uncontracted braille • Single-switch system/PNP enabled • Two-switch system • Administration via iPad • Adaptive equipment used by student • Individualized manipulatives • Alternate form – visual impairment	• Human read aloud • Sign interpretation of text • Language translation of text • Test administrator enter responses for student • Partner-assisted scanning (PAS)

Note: These supports are described for the DLM system as of spring 2015. PNP = Personal Needs and Preferences.

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.

IV.3.A.i. 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

Client.

- *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.

IV.3.A.ii. 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 PNP.

- *Uncontracted braille* – Uncontracted braille testlets are available during the spring window for grades 3-5 at the Target and Successor levels and for grades 6 through high school grades at the Proximal Precursor, Target, and Successor levels.[16] The standard delivery method[17] was to deliver braille-ready files electronically to the school or district for local embossing as each testlet was assigned. The KITE system also delivered the identical general testlet form. After the student took the testlet in its embossed form, the teacher transferred the student's answers into the online version of the testlet.
- *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. The system automatically supports two-switch step scanning.
- *Administration via iPad* – Students are able to take the assessment via iPad.
- *Adaptive equipment used by student* – Educators may use any familiar adaptive equipment needed for the student.
- *Individualized manipulatives* – Individualized manipulatives are suggested for use with students rather than requiring teachers to have a standard materials kit. Recommended materials and rules governing materials selection or substitution are described in the TIP. 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.[18] BVI testlets are teacher-administered, requiring the test administrator to engage in an activity outside the system and enter responses into KITE. The general procedures for

[16] See Chapter III for further explanation of braille form availability and design.

[17] Each state had the option for other delivery methods that involved shipping embossed forms to the school.

[18] See Chapter III for further explanation of BVI form availability and design.

administering these forms are the same as with other teacher-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, teachers have the option to use tactile graphics if their student can use them fluently.

IV.3.A.iii. 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 were trained to follow guidance to ensure fidelity in the delivery of the assessment. This guidance included the typical tone and rate of speech, avoiding emphasizing the correct response or important information that would lead the student to the correct response. Teachers were 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 English language arts testlets. Finally, guidance included ensuring that answer choices were always read in the correct 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 were allowed to describe graphics or images to students who need those described. Typically, this additional support would be provided to students who are with blindness or have visual impairments. Alternate text for graphics and images in each testlet was included in the TIP as an attachment after the main TIP information. Test administrators who needed to read alternate text had the KITE system open and the TIPs in front of them while testing so they could accurately read the alternate text provided on the TIPs with the corresponding screen while the student is testing. Human read aloud was 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 could benefit from decoding support in order to demonstrate the comprehension skills in the tested EEs.
- *Sign interpretation of text* – If the student required sign language to understand the text, items, or instructions, the test administrator was allowed to use the words and images on the screen to guide while signing for the student using American Sign Language, Signed Exact English, or any individualized signs familiar to the student. The test administrator was also allowed to spell unfamiliar words when the student did not know a sign for that word and to 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 the cognitive and communication challenges for students taking DLM alternate assessments are unique 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 (a) each student's unique combination of language-related and disability-related needs and (b) 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[19] to translate the text for the student, simplify test instructions, translate words on demand, provide synonyms or definitions, and accept responses in either English or the student's native language.
- *Test administrator enters responses for student* – During computer-administered assessments, if students are unable to physically select their answer choices themselves due to a gap between their accessibility needs/supports and the KITE system, 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 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

[19] Simplified instructions, definitions, and flexible response mode are supports also allowed for non-English learner students.

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.

IV.3.B. *Additional Allowable Practices*

The KITE Client 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 Consortium recognized that students would need various levels of support to interact with the computer. Test administrators were provided general principles for the allowable practices when the supports built into the system did not support a student's completely independent interaction with the system.

When making decisions about additional supports for computer-delivered testlets, educators received training to follow two general principles**.** First, the student should be expected to respond to the content of the assessment independently. No matter what additional supports IEP teams and test administrators selected, all should have been 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, the student should be able to interact with the assessment content and use his or her normal response mode to indicate a selection for each item. Second, test administrators were to ensure that the student was familiar with the chosen supports. Ideally, any supports used during assessment were also used consistently during routine instruction. Students who had never received a support prior to the testing day would be unlikely to know how to make the best use of the support.

In order to select the most appropriate supports during testing, test administrators were encouraged to use their best professional judgment and to be flexible while administering the assessment. Test administrators were allowed to use additional supports beyond PNP options. The supports detailed in Table 46 were allowed in all computer-delivered and teacher-administered testlets unless exceptions were noted in the TIP.

Table 46. Additional Allowable Practice

Practice	Explanation
Breaks as Needed	Students could take breaks during or between testlets. Test administrators were encouraged to use their best judgment about the use of breaks. The goal should have been to complete a testlet in a single session, but breaks were allowed when the student was fatigued, disengaged, or having behavioral problems that could interfere with the assessment.

Practice	Explanation
Individualized Student Response Mode*	The nodes assessed in the teacher-administered testlets do not limit responses to certain types of expressive communication; therefore, all response modes were allowed. Test administrators could 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 could represent the answer choices in an alternate format or layout to ensure the student could indicate a clear response.
Use of Special Equipment for Positioning	For students who needed special equipment to access the test material, for instance, a slant board for positioning, or Velcro objects on a communication board, test administrators were encouraged to use the equipment to maximize the student's ability to provide a clear response.
Navigation Across Screens	For students who had a limited experience with, motor skills for, and/or devices for interacting directly with the computer, the test administrator could assist students to navigate across screens or enter the responses.
Use of Interactive Whiteboard	If the student had a severe visual impairment and needed larger presentation of content than the ×5 magnification setting provides, the test administrator could use an interactive whiteboard or projector, or a magnification device that worked 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 was allowed, as long as the representation did not favor one answer choice over another. For instance, if presenting the answer choices to a student on a communication board or using objects to represent the answer choices, the correct answer choice could not always be closest to the student or in the same position each time.
Use of Graphic Organizers	If the student was accustomed to using specific graphic organizers, manipulatives, or other tools during instruction, the use of those tools was allowable during the DLM alternate assessment.
Use of Blank Paper	If the student required blank, lined, or unlined paper, this could be provided. Once there was any writing on the paper, it became a secure testing document and needed to be disposed of and shredded at the conclusion of the testing session.
Generic Definitions*	If the student did not understand the meaning of a word used in the assessment, the test administrator could define the term generically and allow the student to apply that definition to the problem or question in which the term was used. Exceptions to this general rule were noted in the TIP for specific testlets.

*Note: *Allowed using speech, sign, or language translation unless prohibited for a specific testlet.*

Although many supports and practices were allowable for computer-delivered and teacher-administered testlets, there were also practices that test administrators were 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.

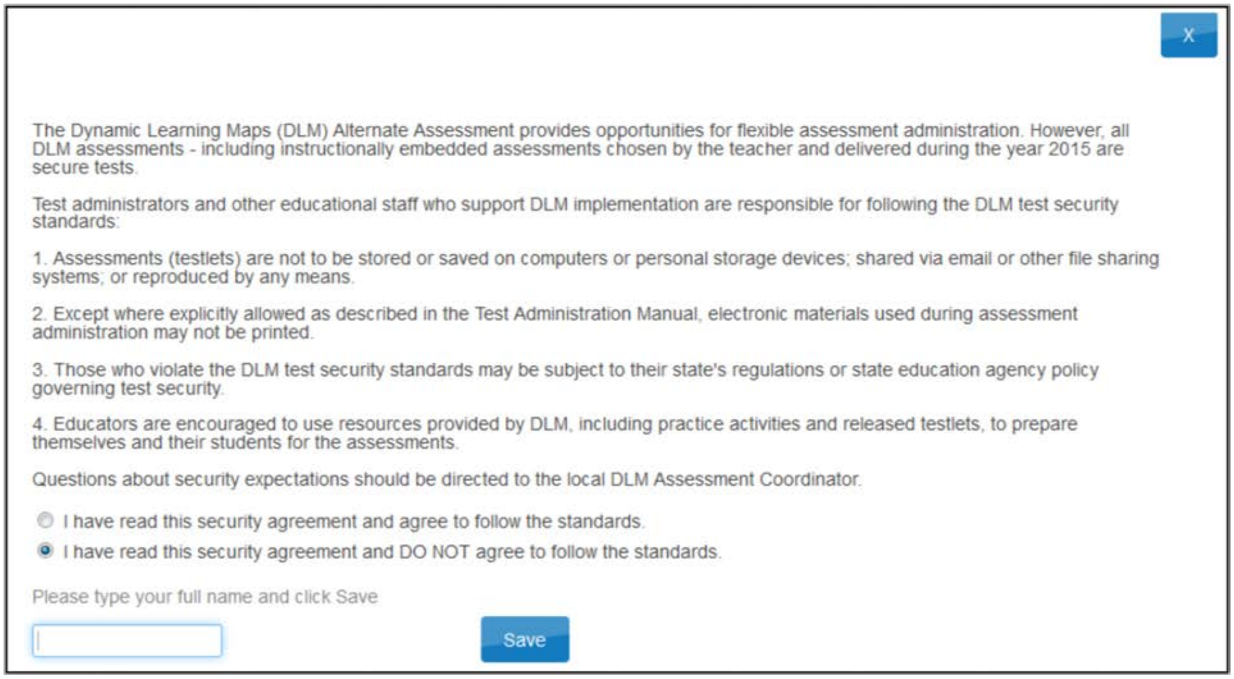
Test administrators were encouraged to ask, via the DLM Service Desk or through their state education agency, any questions regarding whether a support was allowable.

IV.4. SECURITY

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

IV.4.A. TRAINING AND CERTIFICATION

Test security is promoted through required training and certification requirements for test administrators. Test administrators are expected to deliver DLM assessments with integrity and to maintain the security of testlets. Training for test administration details test security measures (see Chapter X). Each year, test administrators must renew their DLM Security Agreement through Educator Portal. The text of the agreement is provided in Figure 42. Test administrators are not granted access to information in the Test Management portion of the Educator Portal if they have not indicated their agreement with these terms.



The Dynamic Learning Maps (DLM) Alternate Assessment provides opportunities for flexible assessment administration. However, all DLM assessments - including instructionally embedded assessments chosen by the teacher and delivered during the year 2015 are secure tests.

Test administrators and other educational staff who support DLM implementation are responsible for following the DLM 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 DLM 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 DLM, including practice activities and released testlets, to prepare themselves and their students for the assessments.

Questions about security expectations should be directed to the local DLM Assessment Coordinator.

○ I have read this security agreement and agree to follow the standards.
● I have read this security agreement and DO NOT agree to follow the standards.

Please type your full name and click Save

Figure 42. Test Security Agreement text.

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.

IV.4.B. Maintaining Security During Test Administration

There are several aspects of the DLM assessment system design that support test security and teacher integrity during use of the system. During the spring testing window, each student is tested on only one of five linkage levels for each EE and the selection of EEs is driven by the adaptive algorithm. Because of the variation in the testlets assigned to different students, test content has more limited exposure than a standardized, single-form test. Because TIPs are the only printed material, there is limited risk of exposure through printed material. Guidance is provided in the Test Administration Manual and on TIPs regarding allowable practices and limits on their use. This guidance is intended to promote implementation fidelity and reduce the risk of cheating or other types of misadministration. (See Chapter IX for test administration evidence related to implementation fidelity.)

Agile Technology Solutions, the organization that develops and maintains the KITE System and provides DLM Service Desk support to educators in the field, has procedures in place to handle alleged security breaches. 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.

IV.4.C. Security in the KITE System

As described earlier in this chapter, the KITE System consists of four applications. Teachers and students see two of these applications: Educator Portal and KITE Client (Test Delivery Engine). A third application, Content Builder, is the content authoring system where test content and associated meta-data are stored. The KITE System is developed and managed by Agile Technology Solutions at the University of Kansas. Agile Technology Solutions also administers the DLM Service Desk, which provides customer support for KITE System users in the field.

Operational access to all servers is controlled by keys that are provided only to system administrators who manage the production data center in the operations team. Access to the networking equipment and hardware consoles is limited to the data center itself; remote access to these devices is limited to the data center–specific administration host.

All KITE applications handle educator and administrative passwords using industry-standard encryption techniques. The password policy requires eight characters, including a number, uppercase letter, and a lowercase letter. Passwords expire annually. All applications generate access records that can be reviewed by system administrators to track access. Access to individual KITE applications is controlled according to the policies set forward for that application and the data the application maintains. All access policies and accounts are reviewed periodically to ensure that access to systems is limited to the appropriate populations.

In accordance with Family Educational Rights and Privacy Act (FERPA) rules, teachers', administrators', and operations' access to personal student data is limited to student records in which that person has a legitimate educational interest. All users in the system are provided the minimum amount of access required. For example, teachers can view only their students' records, and users with building-level roles can view and edit student records within a building. A user's role in an organization defines the level of access to records within that organization. Roles may only be assigned by an existing user at a higher level within the organization. For example, a district-level role may only be assigned by a user with a state-level role; district users may not assign a parallel role to other users. Security levels, groups, and the access provided are reviewed periodically to ensure continued compliance.

The KITE Client is a secure browser that prevents access to unauthorized content during a testing session. The KITE web interfaces use industry standard Secure Socket Layer and Transport Layer Security encryption to securely transfer data to and from the end user from a browser. The KITE system uses load balancing hardware and third party services to both prevent and to mitigate the effects of a distributed denial of service attack if one should occur.

IV.4.D. 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 summative 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.

IV.4.E. DATA SECURITY

Beyond uploads to KITE Educator Portal, there is occasionally a need to transfer secure data between the University of Kansas and the partner states. The consortium uses the University of Kansas' secure file transfer protocol (SFTP) system called the Hawk Drive to transfer files securely. This method is used when local educators need to share personally identifiable information with the DLM help desk agents and when DLM staff deliver score reports and data files to states. Notification of secure file transfer protocol folder links and passwords are made separately.

The consortium collects from states their personally identifiable information (PII) protocols and usage rules, as illustrated in Appendix C.3. The protocols are documented on the state summary sheet as part of the collection of policy information about the state. The consortium documents 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 the Implementation team 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 involves 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 state partners. A document describing the specific procedures is available on the state partner website (see Appendix C.5).

IV.4.F. State-Specific Policies and Practices

Some states also adopt more stringent requirements, above and beyond consortium requirements, for access to test content and for the handling of secure data. Each DLM agreement with a state partner includes a Data Use Agreement. The Data Use Agreement addresses the data security responsibilities of the consortium in regard to United States Department of Education Regulations 20 U.S.C. § 1232g; 34 CFR Part 99, also known as Family Educational Rights and Privacy Act (FERPA). The agreement details the role of the consortium as the holder of data and the rights of the state as the owner of the data. In many cases the standard Data Use Agreement is modified to include state-specific data security requirements. The consortium documents these requirements on the state summary sheet, and the Implementation and Service Desk teams implement the requirements.

The consortium's 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 onto 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 polices 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.

IV.4.G. Forensic Analysis Plans

There are a large number of possible forensic analyses available for investigating test data for possible security breaches, all of which require the collection of specific types of data. Over time, testing programs develop and refine their data collection architecture and mechanisms for the purpose of doing more sophisticated and useful data forensics. As 2015 was the first operational year for the DLM assessment system, limited forensic analyses were conducted for the following reasons:

- Limited data were available. While the goal is to collect data in the future to allow more meaningful analyses (e.g., keystroke data, item level timestamps), the data that was collected during the 2014–2015 operational year was limited to date and time stamps on testlets submitted.

- Validity of results from forensic analyses may not be as well supported as they would in subsequent operational testing administrations. Even with ample field testing and practice opportunities, the DLM assessment system is a new approach to assessing the skills of the population it serves. As such, there may be unanticipated administration situations in the system itself and in the classroom that reflect adjustments to the new assessment system rather than an intentional act or irregularity.

For 2014–2015, start and end times for all testlets were captured. Date stamps for each testlet were reported to states so they could evaluate cases where it appeared a student tested outside of the state's testing window. However, caution was warranted in drawing conclusions from these data based on the presence of implausible values found in the data. Overall, based on the limited data available for 2014–2015, forensic analyses are not planned until suitable data is available, likely in 2016 or beyond. Future analyses may include evaluation of response times to flag outliers, evaluation of answer-changing behavior, analysis of the relationship of First Contact complexity band and the linkage level of the student's last testlet, and identification of students who began the assessment at a lower linkage level and continually routed up a linkage level until reaching the successor level. Forensic analysis plans have been reviewed by the DLM Technical Advisory Committee (See Appendix C.15) and will be updated with the Technical Advisory Committee and state partners as additional data become available.

IV.5. IMPLEMENTATION EVIDENCE FROM 2014–2015 TEST ADMINISTRATION

This section describes evidence collected for 2014–2015 during the operational implementation of the DLM Alternate Assessment System. The categories of evidence include data relating to the adaptive delivery of testlets in the spring window, administration errors, user experience, and accessibility. Additional descriptions of evidence in support of the validity argument is found in Chapter IX.

IV.5.A. ADAPTIVE DELIVERY

During the spring 2015 test administration, the English language arts and mathematics assessments were adaptive between testlets. That is, the linkage level associated with the next testlet a student received was based on the student's performance on the most recently tested EE, with the specific goal of maximizing the match of student knowledge, skill, and ability to the appropriate linkage level content. Specifically:

- The system adapted up one linkage level if students responded correctly to at least 80% of the items measuring the previously tested EE. If testlets were already at the highest level (i.e., Successor), they remained there.
- The system adapted 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 remained there.
- Testlets remained at the same linkage level if students responded correctly to between 35% and 80% of the items on their previously tested EE.

- When a testlet contained items aligned to more than one EE,[20] a percentage of items answered correctly was calculated for each group of items measuring the same EE. The minimum of these values was then used to determine the next linkage level based on the above thresholds.

Threshold values for routing were selected with the number of items included in a testlet in mind. Single-EE testlets contain between three and five items. Multi-EE testlets contain between three and eight items, with between one and three items measuring each EE.

Considering a testlet that contains three items measuring the EE, if a student responds incorrectly to all items or only correctly answers one item (proportion correct, less than .35), then the linkage level of the testlet is likely too challenging, and the student's testlet would be routed to a lower linkage level to provide a better match to the student's knowledge, skills, and ability. 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, on a testlet that contains five items measuring the EE, if a student responds to at least four items correctly (proportion correct, greater than or equal to .80), then the linkage level of the testlet is likely too easy, and the student's testlet 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 ability being assessed at that linkage level. Therefore, the student's testlet is neither routed up nor down to a different linkage level for the subsequent testlet. Because most testlets were built with three to five items, and wanting to err on the side of assigning test content that students have a decent chance of succeeding on, 35% and 80% were determined to be reasonable thresholds for adapting down or up a linkage level, respectively.

The linkage level of the first testlet assigned to a student was based on prior assessment evidence or First Contact survey responses (see Adaptive Delivery earlier in this chapter). The correspondence between the First Contact complexity bands and first assigned linkage levels are shown in Table 47. Based on the concordance between complexity band and linkage level, students are not able to receive a first testlet at the Successor level.

[20] This rule only applied to testlets in the year-end and end-of-instruction models.

Table 47. 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

Depending on the testing model, grade, and subject, there were four to seven opportunities for linkage levels to be adapted between testlets. Figure 43 provides an example of a student who was administered five testlets. In the example, after the first assigned testlet (determined by the First Contact complexity band or prior performance in ITI), there were four opportunities for adaptation. The first three testlets were adapted up or down a level, whereas the fourth testlet remained at the same linkage level as the previous testlet. Overall, linkage levels adapted up or down between testlets 75% of the time.

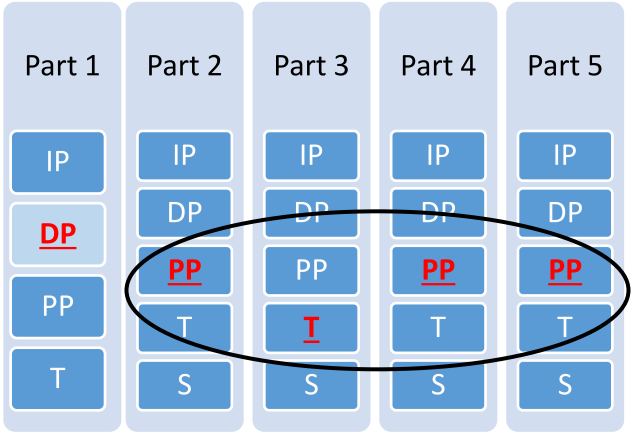


Figure 43. Linkage levels adapting up and down between testlets.

IP = Initial Precursor; DP = Distal Precursor; PP = Proximal Precursor; T = Target; S = Successor.

Following the spring 2015 administration, the mean percentage of testlets that adapted up a linkage level, stayed at the same linkage level, or adapted down a linkage level from the first testlet administered and the second was calculated over all students within a grade, content area, and complexity band. The aggregated results can be seen in and Table 48 and Table 49

For the majority of students across all grades who were assigned to the Foundational complexity band by First Contact survey, testlets did not adapt to a higher linkage level after the first assigned testlet. Consistent patterns were not as apparent for students who were assigned at Bands 1 and 2. Generally, there was a more even split between students assigned at Band 1 whose testlets did not adapt a linkage level and those students whose testlets did adapt up or down a linkage level between the first and second testlets. For students in Band 2, the distributions across the three categories were more variable. That is, for some combinations of grade, content area, and model, the percentage of students whose testlets did not adapt was greater than the percentage of students whose testlets did adapt up or down a level. In other combinations, the opposite pattern appeared. Further investigation is needed to evaluate reasons for these different patterns. Finally, for the majority of students assigned to Band 3, linkage levels between first and second testlets either did not adapt or adapted up a level. As students cannot be assigned to the Successor linkage level for their first testlet, this finding is expected.

Overall, for students assigned to the Foundational and Band 3 complexity bands by the First Contact survey, results indicated that linkage levels tend not to adapt to a different level between the first and second testlet. The exception was linkage levels that adapted up to the Successor level for students assigned to Band 3. These results build on earlier findings from the pilot study (see Chapter III) and suggest that the First Contact complexity band assignment was an effective tool for assigning these students content at appropriate linkage levels. Results also indicated that linkage levels of students assigned to Bands 2 and 3 are more variable with respect to the direction in which they move between the first and second testlets. Several factors may help explain these results, including more variability in student characteristics within this group and content-based differences across grade and content areas. Further exploration is needed in this area.

Table 48. Adaption of Linkage Levels Between the First and Second Testlets for Students in the English Language Arts Year-End Model or End-of-Instruction Model

Grade or Course	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 (%)
3	15.2	84.8	29.7	44.9	25.3	71.1	18.1	10.8	86.1	8.7	5.2
4	27.7	72.3	14.5	53.7	31.9	32.1	53.1	14.8	38.4	58.3	3.3
5	31.5	68.5	33.0	53.4	13.7	75.1	19.7	5.1	93.2	4.9	1.8
6	26.7	73.3	39.7	34.3	26.0	62.9	29.3	7.8	58.6	35.7	5.7
7	24.4	75.6	32.9	45.3	21.8	57.1	36.0	7.0	52.5	37.1	10.4
8	29.7	70.3	21.4	55.7	22.8	41.4	50.2	8.4	78.2	19.0	2.8
9	12.0	88.0	32.6	40.3	27.1	54.9	32.8	12.3	67.3	26.5	6.2
10	11.4	88.6	10.3	46.2	43.6	17.2	67.1	15.7	44.0	46.1	9.9
11	18.8	81.2	12.5	45.9	41.7	43.4	37.8	18.8	66.8	27.3	6.0
Eng 2	25.5	74.5	34.5	47.7	17.7	18.2	74.5	7.2	50.7	42.3	7.0
Eng 3	53.2	46.8	66.7	24.1	9.3	68.9	29.2	1.9	72.9	25.4	1.7

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

Table 49. Adaption of Linkage Levels Between the First and Second Testlets for Students in the Math Year-End Model or End-of-Instruction Model

Grade or Course	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 (%)
3	10.4	89.6	15.3	60.2	24.5	27.3	48.6	24.1	43.8	48.4	7.8
4	8.9	91.1	48.2	37.0	14.7	54.9	37.8	7.3	42.2	48.7	9.1
5	23.3	76.7	7.4	54.1	38.5	11.1	58.5	30.4	53.1	44.5	2.4
6	20.5	79.5	19.2	51.4	29.3	37.0	45.4	17.7	50.8	41.9	7.3
7	15.5	84.5	16.5	55.7	27.9	50.4	38.4	11.2	40.4	53.9	5.7
8	31.5	68.5	21.2	49.9	28.9	13.4	65.7	20.9	40.5	49.2	10.4
9	16.3	83.7	12.0	60.0	28.0	10.6	72.5	16.8	26.4	64.9	8.7
10	16.6	83.4	1.9	41.3	56.8	6.9	53.1	39.9	44.1	44.1	11.8
11	14.2	85.8	3.1	42.5	54.4	4.5	56.9	38.6	20.7	68.2	11.1
Alg 1	34.7	65.3	58.8	22.8	18.4	81.5	9.5	9.1	57.0	25.6	17.4
Alg 2	33.3	66.7	71.4	14.3	14.3	25.0	45.0	30.0	46.2	30.8	23.1
Geom	53.1	46.9	80.0	20.0	0.0	51.9	48.1	0.0	53.8	42.3	3.8

Note: Foundational is the lowest complexity band, so testlets could not adapt down a linkage level. Alg = algebra; Geom = Geometry.

IV.5.B. Administration Errors

The routing algorithm was first used during the 2014–2015 operational assessment. Monitoring of testlet assignment uncovered several incidents that affected student assignment to tests, including misrouting errors due to changes in student data during the testing window and scoring errors, which may have indirectly affected routing because the thresholds are based on percentage of items answered correctly within a testlet. For more information regarding the incidents identified, see Appendix C.7.

Table 50 provides a summary of the number of students affected by each of the incidents, as delivered to states in the Incident Supplemental File. The number of students participating in the year-end model who were affected by each incident ranged from 0 to 3263. In cases for which misrouting was identified during the testing window, states were provided with lists of students affected and were given an option to revert each student's test back to the end of the last correctly completed test (i.e., the point at which routing failed) and complete the remaining testlets as intended.

Table 50. Number of Students Affected by Each 2015 Incident

Incident Code	Incident Description	Frequency
1	Item did not have a correct answer.	116
2	Auto-enrollment problem in Missouri.	0
3	Misrouting (administration of the wrong testlet) because of First Contact.	126
4	Misrouting (administration of the wrong testlet) because of PNP change.	134
5	Misrouting (administration of the wrong testlet) due to incorrect linkage level assignment.	52
6	Misrouting (administration of the wrong testlet) because testlets were administered out of order.	38
7	Misrouting (administration of the wrong testlet) – the same test was given more than once.	0
8	Multi-select multiple choice items were not scored by the system.	3
9	Incorrect key.	23

Incident Code	Incident Description	Frequency
10	Error in the administration of certain multi-select multiple choice items (multiple answers could not be selected).	3263
11	More than one correct answer on a multiple-choice type item.	0

Note: PNP = Personal Needs and Preferences.

All reported incidents were shared with the Technical Advisory Committee in May 2015, and their feedback was solicited regarding potential impact and next steps for remediation and correction. The Technical Advisory Committee recommended that a special circumstance incident file be prepared for states and delivered with the General Research File (GRF; see Chapter VII) to inform the states of all students affected by each issue. States were able to use this file to make determinations about potential invalidation of records at the student level based on state-specific accountability policies and practices.

In instances in which states made the decision to revert to an earlier point in the testing process due to incorrect testlet assignment, the Technical Advisory Committee recommended that the original responses not be included in standard-setting impact data or score reporting. They also recommended that students who had been affected by (a) scoring errors that were corrected by script or (b) misrouting that was not addressed by a state's decision to revert to an earlier point in the testing process should be included in standard-setting impact data files because the small number of students in this category was unlikely to have a large effect on the overall sample for standard-setting impact data.

IV.5.C. User Experience with Assessment Administration and KITE System

User experience with the 2014–2015 assessments was evaluated through a spring 2015 teacher survey disseminated to classroom teachers who had administered a DLM assessment during the 2014–2015 school year spring window. User experience with the KITE system is summarized in this section, and additional survey contents are reported in the Accessibility section below and in Chapter IX (Validity).

A total of 1792 teachers from states participating in the DLM year-end assessment model responded to the survey (estimated response rate of 12.7%). Most of the respondents reported that they had assessed a relatively small number of students during the testing window; 58.8% reported assessing four or fewer students. The self-reported numbers of students assessed for the year-end assessment model are summarized in Table 51.

Table 51. Self-Reported Number of Students Assessed, YE (N=1792)

Number of Students Assessed	**YE**	
	n	**%**
1	360	20.1
2	268	15.0
3	212	11.8
4	209	11.7
5	145	8.1
6	148	8.3
7	113	6.3
8	104	5.8
9	51	2.8
10	42	2.3
11	41	2.3
12	21	1.2
13	22	1.2
14	12	0.7
15 or more	44	2.5

The remainder of this section describes teachers' responses to the portions of the survey addressing educator experience with DLM assessments and the KITE Client software.

IV.5.C.i. Educator Experience

Respondents were asked to reflect on their own experience with the assessments and their comfort level and knowledge with regard to administering them. Most of the questions required respondents to rate results on a four-point scale: strongly disagree, disagree, agree, or strongly agree. Responses are summarized in Table 52. The first two questions (regarding comfort level with the administration of both computer-administered and teacher-administered testlets) were only displayed if respondents had previously disclosed that they had administered the appropriate kind of testlet.

Table 52. Teacher Response Regarding Test Administration (N=1128)

	SD		D		A		SA	
Statement	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
Confidence in ability to deliver computer-administered testlets	26	2.3	61	5.4	492	43.6	549	48.7
Confidence in ability to deliver teacher-administered testlets	21	2.5	66	8.0	384	46.3	358	43.2
Test administrator training prepared respondent for responsibilities of test administrator	160	10.1	364	22.9	852	53.6	214	13.5
Respondent knew how to use accessibility features, allowable supports, and options for flexibility	70	4.4	246	15.5	1030	64.8	244	15.3
Testlet Information Pages helped respondent to deliver the testlets	188	11.8	382	24.0	828	52.1	191	12.0

Note: SD = strongly disagree; D = disagree; A = agree; SA = strongly agree.

Teachers responded that they were very confident with administering either kind of testlet, with 92.3% reporting responses of agree or strongly agree for computer-administered testlets, and 89.5% reporting responses of agree or strongly agree for teacher-administered testlets. Respondents believed that the required test administrator training prepared them for their responsibilities as a test administrator, with 67.1% responding with agree or strongly agree. Additionally, most teachers responded that they knew how to use accessibility features, allowable supports, and options for flexibility (80.1%) and that the TIPs helped them to deliver the testlets (64.1%).

IV.5.C.ii. KITE System

Teachers were asked questions regarding the technology used to administer testlets, including the ease and use of the KITE Client and Educator Portal.

The software used for the actual administration of DLM testlets is KITE Client. Teachers were asked to consider their experiences with KITE Client and respond to each question on a five-point scale: very hard, somewhat hard, neither hard nor easy, somewhat easy, or very easy. Table 53 summarize teacher response to these questions.

Table 53. Ease of Using KITE Client (N = *632)*

	VH		SH		N		SE		VE	
Statement	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
Enter the site	28	1.8	125	8.1	266	17.2	469	30.4	655	42.4
Navigate within a testlet	26	1.7	73	4.7	260	16.9	490	31.8	692	44.9
Submit a completed testlet	12	0.8	36	2.3	208	13.5	449	29.1	836	54.3
Administer testlets on various devices	49	3.2	113	7.4	446	29.4	421	27.8	488	32.2

Note: *VH = very hard; SH = somewhat hard; N = neither hard nor easy; SE = somewhat easy; VE = very easy.*

Respondents found it to be either somewhat easy or very easy to enter the site (72.8%), to navigate within a testlet (76.7%), to submit a completed testlet (83.4%), and to administer testlets on various devices (60.0%).

Educator Portal is the software used to store and manage student data and to enter PNP and First Contact information. Teachers were asked to assess the ease of navigating and using Educator Portal for its intended purposes. The data are summarized in Table 54 on the same scale that was used to rate experience with the KITE Client.

Table 54. Ease of Using Educator Portal (N = *650)*

	VH		SH		N		SE		VE	
Statement	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
Navigate the site	90	5.7	384	24.2	421	26.5	440	27.7	255	16.0
Enter PNP and First Contact information	35	2.2	256	16.1	415	26.1	566	35.7	315	19.8
Manage student data	100	6.3	341	21.5	442	27.8	475	29.9	230	14.5
Manage your account	59	3.7	275	17.3	509	32.1	489	30.8	255	16.1

Note: *VH = very hard; SH = somewhat hard; N = neither hard nor easy; SE = somewhat easy; VE = very easy; PNP = Personal Needs and Preferences.*

Overall, respondents found it to be either somewhat easy or very easy to navigate the site (43.7%), to enter PNP and First Contact information (55.5%), to manage student data (44.4%), and to manage his or her account (46.9%).

Finally, respondents were asked to rate their overall experience with the KITE Client and Educator Portal on a four-point scale: Poor, Fair, Good, and Excellent. Results are summarized in Table 55.

Table 55. Overall Experience with KITE and Educator Portal (N = *631)*

	Poor		Fair		Good		Excellent	
	n	**%**	**n**	**%**	**n**	**%**	**n**	**%**
KITE Client	123	8.0	341	22.1	733	47.5	346	22.4
Educator Portal	181	11.7	489	31.7	678	44.0	194	12.6

The majority of respondents reported a positive experience with KITE; 47.5% of respondents ranked their experience as good, and 22.4% of respondents ranked their experience as excellent. A majority reported an overall positive experience with Educator Portal, with 44.0% ranking their experience as good and 12.6% ranking their experience as excellent.

IV.5.D. ACCESSIBILITY

Guidance around accessibility provided by DLM distinguishes between supports that: (a) can be used by selecting online features via the PNP, (b) require additional tools or materials, and (c) are provided by the test administrator outside the system. Table 56 shows selection rates for three categories of PNP supports, sorted by rate of use within each category.

The first category, Supports Activated by PNP, includes supports that are provided within KITE Client. This category of support includes features delivered online. Magnification, which allows educators to choose the amount of screen magnification during testing (×2, ×3, ×4, or ×5), was used by 8.2% of students. Without magnification, the font is Report School, size 22. Overlay color, used by 5.4% of students, allows educators to change the background color of the test from white to an alternate color (blue, green, pink, gray, or yellow). Color contrast allows educators to change the color scheme for the background and font and was used by 5.4% of students. Invert color choice allows educators to change the background color to black and font color to white, which was used by 4.2% of students. Read aloud (text-to-speech; TTS) was used by 1.1% of students. Read aloud (TTS) consists of synthetic spoken audio (read aloud with highlighting).

The second category, Supports Requiring Additional Tools/Materials, includes supports that are recorded in the PNP but provided outside of KITE Client and require additional tools or materials. Individualized manipulatives were used by 39% of students. Individualized manipulatives are familiar manipulatives that teachers use during instruction. Additional information about individualized manipulatives is provided in the TIP. A calculator was used by 24% of students. A calculator is permitted on math testlets unless it interferes with measurement of the tested construct in the testlet. The TIP for each math testlet specifies whether a calculator is permitted. A single-switch system, used by 5.7% of students, is an interface that emulates the Enter key on the keyboard. Educators set scanning settings for the single-switch system in the PNP. An alternate form – visual impairment was used by 2.0% of students who do not read braille but are blind or have a visual impairment that prevents interaction with the onscreen content. This option is available for some specific EEs and linkage levels. Alternate forms are not provided at every single EE and linkage level. Two-switch systems were used by 1.2% of students. Two-switch systems consist of two switches and a switch interface that are used to emulate the Tab key to move between choices and the Enter key to select the choice when highlighted. Uncontracted braille was used by 0.2% of students. Uncontracted braille forms are delivered at the state or district level and in braille-ready files or embossed files.

The third category, Supports Provided Outside the System, includes supports offered outside the KITE system that require actions by the test administrator. Human read aloud was used by 87% of students. In human read aloud, test administrators read the assessment aloud to students. Responses were entered by the test administrator for 46% of students, an option that is intended for use when students are unable to independently and accurately record their responses in the KITE system. Students indicated their responses through their typical response mode, and teachers keyed in those responses. Partner-assisted scanning was used by 7.7% of students. Test administrators translated the text for 1.8% of students who were English language learners or responded best to a language other than English. Test administrators signed test content for 1.7% of students who used American Sign Language, Exact English, or personalized sign systems.

*Table 56. Personal Needs and Preferences (PNP) Supports Selected for Students, Spring 2015, Year-End Model (*N *= 61,958)*

Support	*n*	%
Supports Activated by PNP		
Magnification	5,083	8.2
Overlay color	3,367	5.4
Color contrast	3,335	5.4
Invert color choice	2,632	4.2
Read aloud (text to speech)	661	1.1
Supports Requiring Additional Tools/Materials		
Individualized manipulatives	24,222	39.0
Calculator	14,655	24.0
Single-switch system	3,549	5.7
Alternate form – visual impairment	1,264	2.0
Two-switch system	769	1.2
Uncontracted braille	114	0.2
Supports Provided Outside the System		
Human read aloud	53,803	87.0
Test administrator enters responses for students	28,583	46.0
Partner-assisted scanning	4,780	7.7
Sign interpretation	1,136	1.8
Language translation	1,064	1.7

Note: During 2015, read aloud was not available in the test delivery engine.

Table 57 describes teacher responses to a survey about the student accessibility experience. Teachers were asked to respond to three items using a four-point Likert-type scale (strongly disagree, disagree, agree, strongly agree). The majority of teachers agreed that the student was able to effectively use accessibility features (66.9%), that accessibility features were similar to ones the student used for instruction (66.7%), and that allowable options for flexibility were necessary when administering the test to meet students' needs (77.9%). These data support the conclusions that the accessibility features of the DLM alternate assessment were effectively used by students, emulated accessibility features used during instruction, and met student needs for test administration.

Table 57. Teacher Report of Student Accessibility Experience (Year-End Model)

Statement	SD		D		A		SA	
	n	%	*n*	%	*n*	%	*n*	%
Student was able to effectively use accessibility features	337	13.6	486	19.6	1254	50.5	407	16.4
Accessibility features were similar to ones student uses for instruction	307	12.4	521	21.0	1457	58.7	199	8.0
Allowable options for flexibility were needed when administering test to meet student needs	152	5.3	488	16.9	1436	49.6	818	28.3

Note: SD = strongly disagree; D = disagree; A = agree; SA = strongly agree.

IV.6. Conclusion

The DLM system was designed to promote instructional relevance, responsiveness to individual student needs, and the detailed model of learning, the DLM map. The dynamic nature of the DLM test administration is reflected in the initial input through the First Contact survey and later, in the linkage level adaptations based on student prior performance. Assessment delivery options allow for necessary flexibility for student communication mode and linkage level while also being controlled to maximize standardization and support valid scores. Finally, the DLM system addresses differences in state policies by offering two types of administration models: the integrated approach, which uses instructionally embedded results to inform summative scores and the year-end approach, which uses only spring results to inform summative scores. To summarize, the DLM system aims to support necessary flexibility while maintaining standard approaches that support the assessment claims and goals (Chapter I).

V. MODELING

The Dynamic Learning Maps® (DLM®) project draws upon on a well-established research base in cognition and learning theory but relatively uncommon operational psychometric methods 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.

V.1. PSYCHOMETRIC BACKGROUND INFORMATION

At the core of the DLM project are learning map models, which are networks of sequenced learning targets that use a form of Bayesian Inference Network methods for statistical modeling. A map model 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 content area. Put in the vocabulary of traditional psychometric methods, a learning map model defines (1) a large set of discrete latent variables indicating students' learning status on key skills and concepts relevant to a large content domain, and (2) a series of pathways indicating which topics (represented by latent variables) are prerequisites for learning other topics.

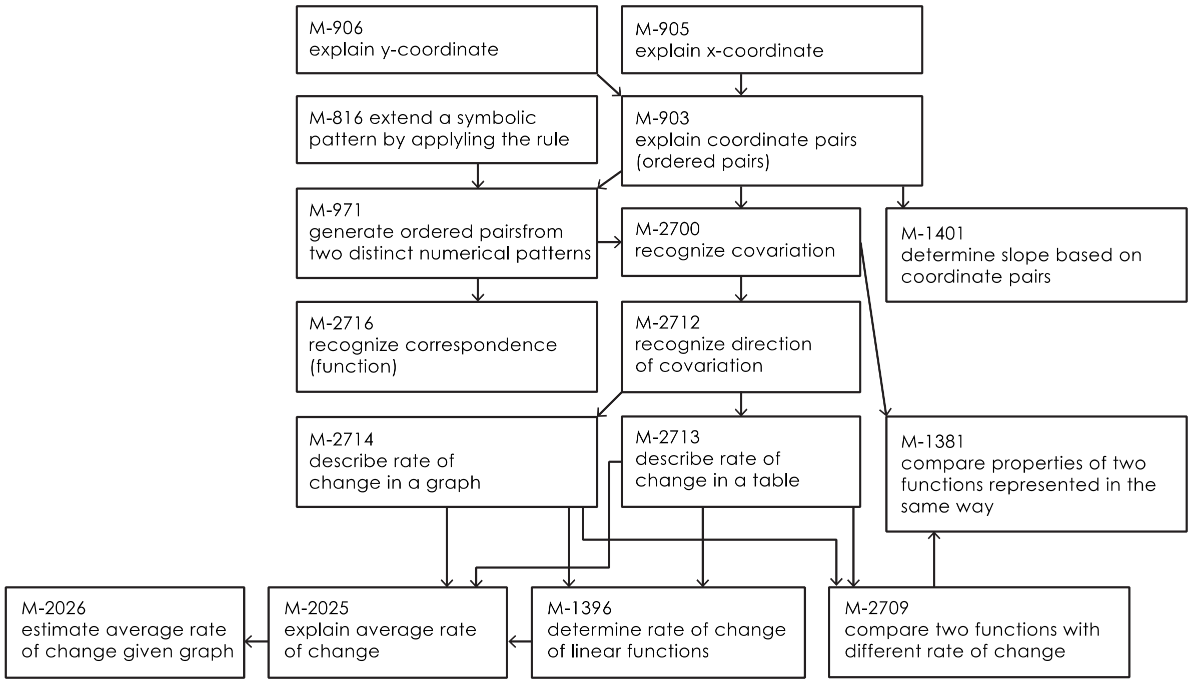
The language that is used to describe the component parts of a DLM map draws from Bayesian networks (e.g., Almond, Mislevy, Steinberg, Yan, & Williamson, 2015; Mislevy & Gitomer, 1995) and graphical modeling in computer science (e.g., Pearl, 1988). Therefore, what might be considered a skill or attribute in psychometrics—the latent variables that span a learning map model—are called nodes. The nodes, rather than items, are the key psychometric units measured in the DLM assessments.

Diagnostic classification models (e.g., Rupp & Templin, 2008; Rupp, Templin, & Henson, 2010), also known as cognitive diagnosis models (e.g., Leighton & Gierl, 2007), or multiple classification latent class models (Maris, 1999), are confirmatory latent class models that characterize the relationship of observed responses to a set of categorical latent variables. These latent variables are called attributes within the context of diagnostic classification models, and psychometrics in general, but are called nodes within the context of DLM modeling. Diagnostic classification models have primarily been used for educational measurement in which detailed information about test-takers is of interest, such as in assessing mathematics (e.g., Bradshaw, Izsák, Templin, & Jacobson, 2014), reading (e.g., Templin & Bradshaw, 2014), and science (e.g., Templin & Henson, 2008).

Because both diagnostic classification models and Bayesian inference networks can be used to describe latent variables in equivalent ways and to characterize relationships between observed responses, both bodies of literature are used to inform DLM modeling. Since the latent variables (called nodes) from Bayesian inference networks and the latent variables (called attributes) from diagnostic classification models are mathematically equivalent, this document blends research and terminology from the two measurement paradigms from which such methods have evolved. Prior to presenting the DLM psychometric model, an example from the DLM mathematics map is provided for context.

V.2. Guiding Example: A Section of the Map

To ground our psychometric model in the context of the nodes being measured and the connections between nodes, Figure 44 shows an excerpt from the DLM mathematics map. This excerpt shows the set of nodes that were developed from a section of the DLM mathematics map where each node is represented by a rectangular box (16 nodes in total). This schematic representation is common in Bayesian Networks where the latent nodes or attributes are represented by rectangular boxes. An excerpt from the text of the definition of slope is provided below the map segment.


M-906
explain y-coordinate
M-905
explain x-coordinate
M-816 extend a symbolic pattern by applying the rule
M-903
explain coordinate pairs (ordered pairs)
M-971
generate ordered pairsfrom two distinct numerical patterns
M-2700
recognize covariation
M-1401
determine slope based on coordinate pairs
M-2716
recognize correspondence (function)
M-2712
recognize direction of covariation
M-2714
describe rate of change in a graph
M-2713
describe rate of change in a table
M-1381
compare properties of two functions represented in the same way
M-2026
estimate average rate of change given graph
M-2025
explain average rate of change
M-1396
determine rate of change of linear functions
M-2709
compare two functions with different rate of change

"Slope involves the construction of a ratio as the measure of a given attribute" (Lobato & Thanheiser, 2002, p. 164). As such, a firm grasp of reasoning with ratios forms the foundation for understanding slope. One type of reasoning with ratios is covariational reasoning, that is, knowing how one quantity changes as another quantity changes (Adamson 2005), which is an essential component of algebraic thinking and understanding functions (Confrey & Smith, 1995).

Carlson and Oehrtman (2005) suggested a framework for thinking about covariation. Students first learn to coordinate two variables by recognizing that a change to one quantity yields a change in another quantity. Then students are ready to determine the direction of the relationship between the two variables. Then students are ready to coordinate the amounts of change in two related quantities.

Figure 44. DLM Mathematics Map Section Pertaining to Prerequisites for Slope

The lines connecting the boxes are called **edges** —the connections between nodes. Edges are also called directed paths or arcs, where one end of the line has an arrow and one end of the line has none. For example, take the connection between node M-903 and node M-1401, as shown near the top right of Figure 44. There is a directed arc between the two nodes where the line originates at node M-903 and terminates with an arrow at node M-1401. This depiction indicates that mastery of node M-903 (explain coordinate pairs/ordered pairs) is a prerequisite to mastery of node M-1401 (determine slope based on coordinate pairs). The graphical depiction of the nodes connected by the edges implies a mathematical model for the relationship between these nodes, described in more detail in the next section, which states that the probability of mastery of the latter node (M-1401) is conditional on the probability of mastery on the prerequisite node (M-903). Put another way, this connection implies that directly connected nodes (linked together by one edge) have direct relationships, nodes connected but not immediately by a single edge have indirect relationships, and nodes that are not connected by a single edge are conditionally independent.

Whereas the depiction in Figure 44 provides the nodes (which are the latent variables purported to be measured by DLM tests) and the connections between nodes, items that purport to measure each node are not shown. Specifications of relationships between items and nodes are called the **measurement model** in diagnostic classification models and broader psychometric fields or the **item model** in Bayesian inference networks. In the context of diagnostic classification models, Figure 44 would be called a **structural model path diagram** as it depicts only the latent attributes and the relationships between the latent nodes.

Figure 45 shows the items in the measurement model from a diagnostic classification model analysis reported by Bradshaw et al. (2014, p. 8). The DLM psychometric model can be described by using this example because it provides an educational context; however, such an example of a set of nodes could be part of any mathematics learning map model, like the DLM mathematics map. The model-related concepts apply to DLM maps in other content areas, as well.

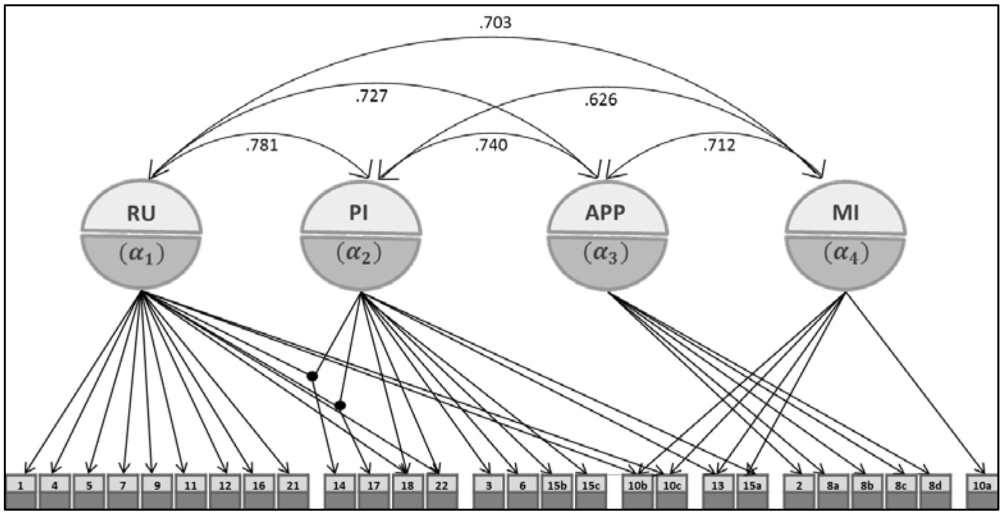


Figure 45. Diagnostic Model Path Diagram with Four Binary Nodes Measured by 27 Test Items.

Note: The bisected circles represent latent nodes from Bradshaw et al. (2014): RU (for Referent Unit), PI (for Partitioning and Iterating), APP (for Appropriateness), and MI (for Multiplicative Inference).

At the bottom of Figure 45, the items measuring each of the nodes are depicted as rectangular boxes where the arrows emanating from each node terminate. The boxes are also bisected to represent that items are dichotomously scored as correct/incorrect. The directed arcs depict the inference that the nodes explain variability in responses to the items. If items were to be represented in a network diagram (such as in Figure 44), they would appear similarly—as additional nodes items that are connected by directional arcs to the overall node they purport to measure. Simply as a matter of convention, items are not typically shown on network diagrams in the DLM project depictions.

To describe how the psychometric model works at the item level, Figure 46 shows an example item from Bradshaw et al. (2014). The item was written to measure two nodes: Referent Unit and Partitioning/Iterating. Using the terminology from diagnostic classification models, the relationship between items and nodes is delineated by an item-by-node *Q-matrix* indicating the nodes measured by each item. In general, for a given item, i, the Q-matrix vector would be represented as $\boldsymbol{q}_i = [q_{i1}, q_{i2}, \dots, q_{iA}]$. Similar to a factor pattern matrix in a confirmatory factor model, Q-matrix indicators are binary—either the item measures a node ($q_{ia} = 1$) or it does not ($q_{ia} = 0$). Therefore, for our example item, the Q-matrix vector would be $[1, 1, 0, 0]$, indicating the item measures the first and second node (Referent Unit and Partitioning/Iterating) but does not measure the third and fourth node (Appropriateness and Multiplicative Comparison).

Ms. Roland gave her students the following problem to solve:

> *Candice has 4/5 of a meter of cloth. She uses 1/8 of a meter for a project.*
> *How much cloth does she have left after the project?*

She had students use the number line so that they could draw the lengths. Which of the following diagrams shows the solution? Assume all intervals are subdivided equally.

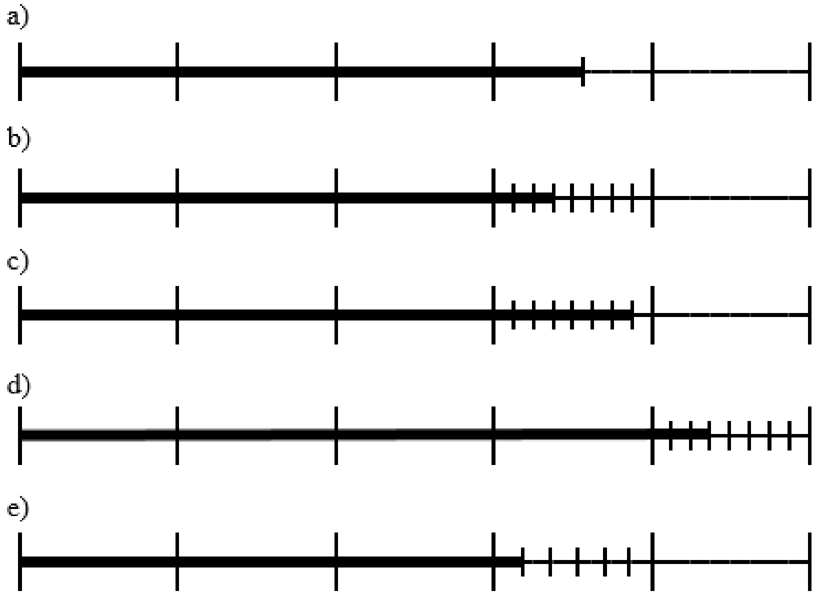


Figure 46. Example Item Measuring Two Nodes: Referent Unit and Partitioning/Iterating

V.3. DLM ITEM/MEASUREMENT MODEL

For each item that measures one or more nodes, there is a set of conditional item response probabilities displayed in a **conditional probability table** *(CPT).* Each combination of node statuses is shown in a row in the CPT, and each row has an estimated probability of correct response to the item, as illustrated in Figure 47 (from Bradshaw et al., 2014, p. 10).

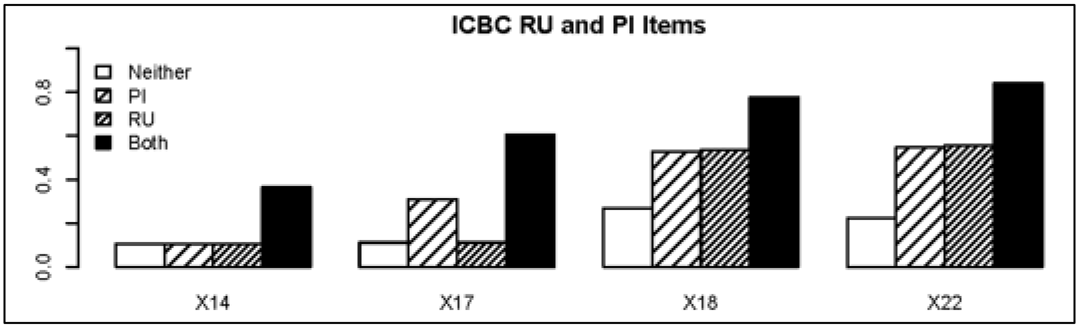


Figure 47. Example Conditional Item Response Probabilities (Item Characteristic Bar Charts) for Items 14, 17, 18, and 22

When an item measures two binary nodes, there are four possible statuses any examinee could have: (1) a master of both nodes, (2) a master of the first and a non-master of the second, (3) a master of the second and a non-master of the first, or (4) a non-master of both nodes. The figure describes these four conditional probabilities in what is called an Item Characteristic Bar Chart (ICBC), an analogous chart to an Item Characteristic Curve (ICC) used in item response models to show the conditional probability of a correct response to an item for a given value of a continuous latent trait or ability. Figure 47 shows the ICBC for the four items that measure both Referent Unit and Partitioning/Iterating (i.e., Items 14, 17, 18, 22). The four bars per item provide the conditional probability of a correct response for each of these items for a given pattern of mastery status on both Referent Unit and Partitioning/Iterating. For example, for Item 17, three statuses have a low probability of correct response: masters of neither node and masters of only Referent Unit have a correct response probability of approximately 0.10, while masters of Partitioning/Iterating have a probability of 0.30. Masters of both nodes have the highest probability of correct response (approximately 0.60). Similarly, in Bayesian inference networks, the statistical parameters describing the graph are conditional probabilities called conditional probability tables (CPTs). The CPTs in Bayesian inference networks and the values of the probabilities in ICBCs in diagnostic classification models are mathematically equivalent. That is, using the diagnostic classification model term, the measurement model parameters of Bayesian inference networks are contained within the CPTs.

When put into a more general latent class framework, diagnostic classification models/Bayesian networks constrain the conditional item response probabilities of a general latent class model so that latent classes with the same pattern of statuses on nodes measured by an item have the same conditional item response probability. For example, for Item 17, all four patterns containing mastery of both Referent Unit and Partitioning/Iterating (patterns [1,1,0,0], [1,1,0,1], [1,1,1,0], and [1,1,1,1]) have a conditional item response probability of 0.60, regardless of the mastery statuses for the third and fourth nodes. The resulting link between the Q-matrix, the node pattern statuses (representing each of the latent classes), and the observed item response show why diagnostic classification models and Bayesian inference networks are confirmatory mixture models.

Much of the psychometric development in the field of diagnostic classification models has focused on developing parameterizations that specify how nodes defined in the Q-matrix relate to item responses. Specifically, the Log-Linear Cognitive Diagnosis Model (LCDM; Henson, Templin, & Willse, 2009) is a general modeling approach based on the General Diagnostic Model (von Davier, 2005) and equivalent to the so-called Generalized DINA Model (de la Torre, 2011). Although similar in parameterization to exploratory classification models developed by Magidson and Vermunt (2001), the LCDM provides a confirmatory classification model for mapping latent nodes onto item responses. The LCDM parameterization allows for both non-compensatory and compensatory links between nodes and the items of a test, subsuming many models currently in use. Further, the logistic parameterization allows for inclusion of other effects (such as testlet effects), which may become important in future DLM modeling.

To demonstrate how the LCDM relates conditional item response probabilities to node mastery status, consider an item like the one presented in Figure 46 that measures two nodes: Node 1 (Referent Unit, denoted α_{e1}) and Node 2 (Partitioning/Iterating, denoted α_{e2}). For a dichotomous (binary) item response, the LCDM provides the log-odds or logit of a correct response (indicated by $X_{ei} = 1$) to item (i) by examinee (e) as:

$$\ln\left(\frac{P(X_{ei} = 1|\boldsymbol{\alpha}_e)}{P(X_{ei} = 0|\boldsymbol{\alpha}_e)}\right) = \lambda_{i,0} + \lambda_{i,1(1)}(\alpha_{e1}) + \lambda_{i,1(2)}(\alpha_{e2}) + \lambda_{i,2(1*2)}(\alpha_{e1}\alpha_{e2}). \quad (1)$$

There are four item parameters for this item as specified by the LCDM. The parameter $\lambda_{i,0}$ is the intercept and represents the predicted log-odds of a correct response for examinees in the reference group—examinees who have not mastered Referent Unit (Node 1; $\alpha_{e1} = 0$) or PI (Node 2; $\alpha_{e2} = 0$). The parameter $\lambda_{i,1(1)}$ is the simple or conditional main effect for mastery of Referent Unit, representing the **increase** in the log-odds of a correct response for examinees who have mastered Referent Unit ($\alpha_{e1} = 1$) but not Partitioning/Iterating ($\alpha_{e2} = 0$). Similarly, the parameter $\lambda_{i,1(2)}$ is the simple or conditional main effect for mastery of Partitioning/Iterating representing the *increase* in the log-odds of a correct response for examinees who have mastered Partitioning/Iterating ($\alpha_{e2} = 1$) but not Referent Unit ($\alpha_{e1} = 0$). Finally, the parameter $\lambda_{i,2(1*2)}$ is the interaction effect for mastery of Referent Unit and Partitioning/Iterating that represents the *change* in log-odds for examinees who have mastered both nodes ($\alpha_{e1} = 1$ and $\alpha_{e2} = 1$). For models where the latent trait is discrete, these two approaches are equivalent, yielding estimated parameters that are equivalent when transformed onto the appropriate metric used by each.

As the LCDM falls into a family of statistical models called finite mixture models (e.g., McLachlan & Peel, 2004; latent class models are also a member of this family), there are a few caveats that must be applied to the parameters of the model. Namely, each profile of attributes represents a unique latent class in a finite mixture model. In estimation of finite mixture models, the definition of each class may switch from one estimation to another (also between iterations within the same estimator). As the definition of each class defines an examinee's mastery status, a series of order constraints must be imposed on the LCDM parameters to ensure the nodes have the same meaning. In short, the constraints imposed are such that the item response probability monotonically increases as the overall number of nodes mastered increases. For main effects, this means all are constrained to be positive. More details on these constraints can be found in Henson et al. (2009) or Rupp et al. (2010).

V.4. Estimation of Student Mastery Probabilities

Once the LCDM item parameters have been calibrated, student mastery probabilities are then obtained for each node. For DLM scoring, student mastery probabilities are *Expected a Posteriori,* or EAP estimates (the most commonly used scoring method used in item response models; for a thorough treatment of the topic, see Rupp, Templin, & Henson, 2010, Chapter 10). For each node a and student e, EAP estimates of mastery probability $\hat{\alpha}_{ea}$ are obtained using the following formula:

$$\hat{\alpha}_{ea} = \frac{\prod_{i=1}^{I_a}\left[P(X_{ei}=1|\alpha_{ea}=1)^{X_{ei}}\left(1-P\left(X_{ei}=1|\alpha_{ea}=1\right)\right)^{1-X_{ei}}\right]^{q_{ia}} P(\alpha_{ea}=1|\boldsymbol{\alpha}_e)}{\sum_{m=0}^{1}\prod_{i=1}^{I_a}\left[P(X_{ei}=1|\alpha_{ea}=m)^{X_{ei}}\left(1-P\left(X_{ei}=1|\alpha_{ea}=m\right)\right)^{1-X_{ei}}\right]^{q_{ia}} P(\alpha_{ea}=m|\boldsymbol{\alpha}_e)} \quad (2)$$

Here, $P(X_{ei}=1|\alpha_{ea}=m)$ is the model-based probability of answering item i correct, conditional on student e having mastery status m for node α_{ea}. Mastery statuses can take two values: masters ($m=1$) and non-masters ($m=0$). The portion of the equation in brackets is raised to a power of q_{ia}, the Q-matrix indicator that item i measures node a. Since q_{ia} is binary, only items measuring node a directly contribute to the estimate of mastery probability. Finally, $P(\alpha_{ea}=m|\boldsymbol{\alpha}_e)$ represents the prior probability of mastery status m conditional on the statuses of nodes specified as precursor nodes in $\boldsymbol{\alpha}_e$ based on connections between nodes in the DLM map. For nodes without precursor connections, this probability represents the marginal probability that any student is a master of the node.

V.5. Additional DLM Categorizations: Essential Elements and Linkage Levels

Because the primary goal of the DLM Consortium is to assess what students with the most significant cognitive disabilities know and can do, alternate grade-level expectations called Essential Elements (EE) were created to more accurately reflect the skills and abilities that students in this population can demonstrate at the same grade level as students without disabilities. The EEs were derived from the CCSSs for each content area and strand/cluster, and they represent a similar skill development as do the CCSSs for each strand/cluster. Simply put: EEs are collections of nodes that are the focus of a given grade level. Each EEs has an associated cluster of nodes, roughly in order of increasing difficulty, that are called linkage levels. There are five linkage levels for each EE: Initial Precursor, Distal Precursor, Proximal Precursor, Target, and Successor. The relationship between node, linkage level, and Essential Element is summarized in Table 58.

Table 58. Organizational levels and their meanings

Organizational Level	Meaning
Essential Element	Collection of nodes related to a topic area within the CCSS
Linkage Level	Organization of nodes within an Essential Element, forming a stage in a learning progression
Node	Atomic knowledge, skill, or ability

DLM modeling uses the linkage levels of EEs for scoring. An example of an EE with sets of nodes labeled as linkage levels is given in Figure 48. The EE in the example is from third grade mathematics and is labeled M.EE.3.MD.4: "Measure length of objects using standard tools, such as rulers, yardsticks, and meter sticks." See Chapter III for more detail on the development of the linkage levels and how they relate to the DLM design.

In Figure 48, each node is shown as a red box. In the top right corner of the box, a letter code is given indicating the linkage level of the node (IP: Initial Precursor, DP: Distal Precursor; PP: Proximal Precursor; T: Target; S: Successor; and UN: Untested Node—a node not currently tested as part of DLM testlets).

For each linkage level embedded within each EE, DLM testlets were written with items measuring the listed linkage level nodes. Because of the DLM administration design, students seldom took testlets outside of a single linkage level within an EE. Students typically saw a single testlet within a given EE; consequently, data where students responded to testlets at adjacent linkage levels within an EE is sparse. Because direct evidence of connections between nodes at different linkage levels was not often collected, DLM node parameters could not be estimated. Instead, a linkage level model was used to estimate examinee proficiency. Measuring learning for students with the most significant cognitive disabilities requires a highly dimensional model, so each linkage level is treated as a dimension.

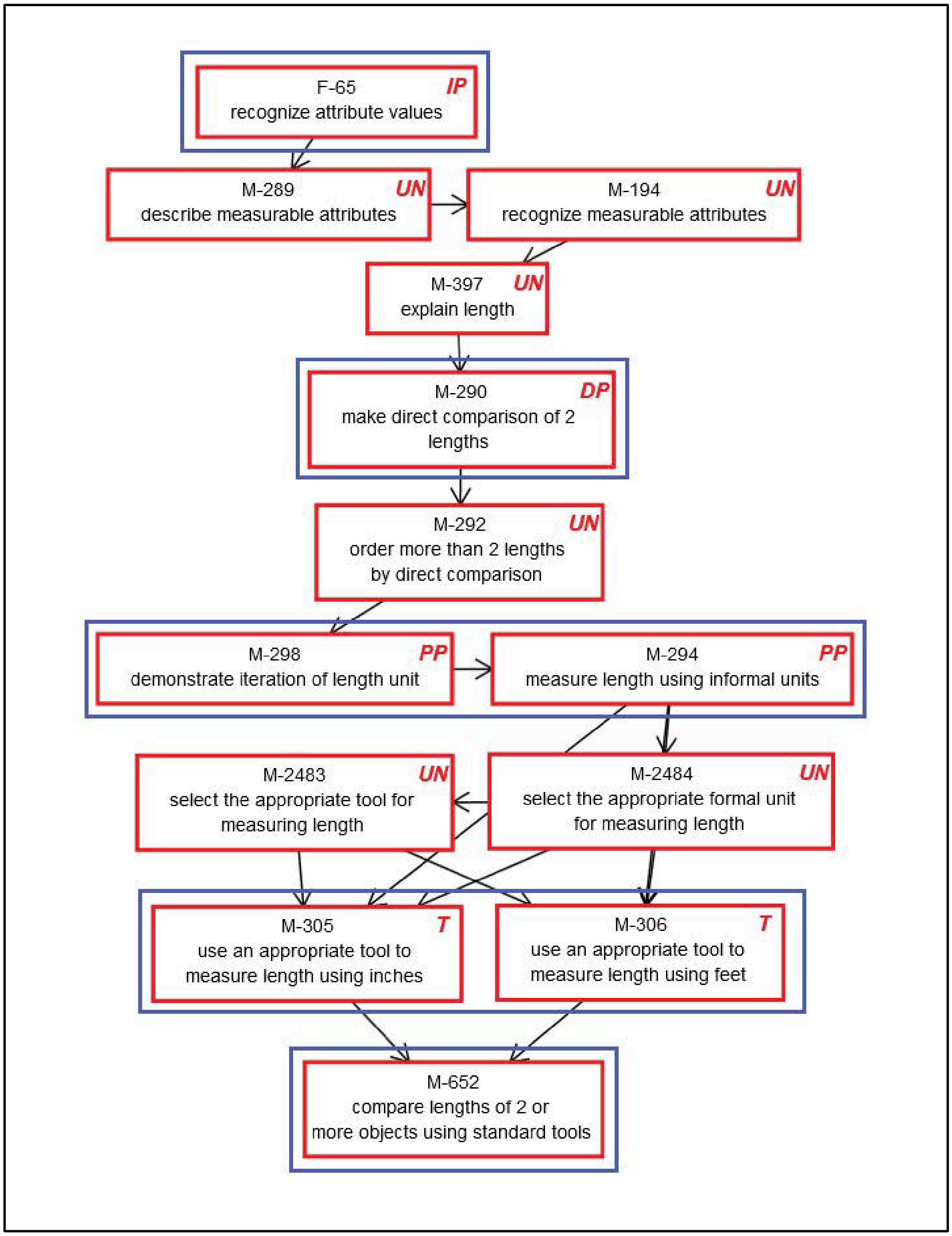

F-65 IP
recognize attribute values
M-289 UN
describe measurable attributes
M-194 UN
recognize measurable attributes
M-397 UN
explain length
M-290 DP
make direct comparison of 2 lengths
M-292 UN
order more than 2 lengths by direct comparison
M-298 PP
demonstrate iteration of length unit
M-294 PP
measure length using informal units
M-2483 UN
select the appropriate tool for measuring length
M-2484 UN
select the appropriate formal unit for measuring length
M-305 T
use an appropriate tool to measure length using inches
M-306 T
use an appropriate tool to measure length using feet
M-652
compare lengths of 2 or more objects using standard tools

Figure 48. Mini-map of nodes for Essential Element M.EE.3.MD.4 (third grade mathematics)

V.6. Linkage Level Model with Fungible Item Parameters

Difficulties were encountered in supporting a psychometric model that operated at the node level during the 2014-2015 year because of lower than expected field test participation and limitations of the operational administration design. In particular, many nodes were not assessed with sufficient numbers of items to produce accurate estimates, resulting in inaccurate mastery estimates. As a result, the calibration and scoring model was changed from being estimated at the node level to being estimated for the linkage level. Furthermore, because data from students taking testlets at multiple linkage levels within an EE were uncommon, simultaneous calibration of all linkage levels within an EE was not possible. Finally, because items were developed to a precise cognitive specification, all item intercept and main effect parameters from 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. These changes to the scoring model were discussed and approved by the DLM Technical Advisory Committee in a phone meeting on July 21, 2015.

The final DLM scoring model for the 2014-2015 administration was as follows. Each linkage level within each EE was considered the latent variable to be measured (the node). Students were therefore classified into two statuses for each linkage level of each EE: either master or non-master. All items in a linkage level were assumed to measure that linkage-level node, meaning the Q-matrix for the linkage level was a column of ones. As such, each item measured one latent variable, leaving two possible item parameters per item: an item intercept and an item main effect. As per the assumption of item fungibility, a single item intercept was calculated for all items within a linkage level, and all item main effects were set to be equal. Finally, the proportion of masters for the linkage level (the analogous map parameter) was estimated. In total, three parameters per linkage level were estimated by the final DLM calibration and scoring model: a fungible item intercept, a fungible item main effect, and the proportion of masters.

Then, for a student e, overall marginal likelihood function for any linkage level was:

$$L(\mathbf{X}_e) = \sum_{\alpha_{eLL(EE)}=0}^{1} \eta_{\alpha_{eLL(EE)}} \prod_{i=1}^{I_e} P\left(X_{ei} = 1 \middle| \alpha_{eLL(EE)}\right)^{X_{ei}} \left(1 - P\left(X_{ei} = 1 \middle| \alpha_{eLL(EE)}\right)\right)^{1-X_{ei}} \tag{3}$$

Here, $\eta_{\alpha_{eLL(EE)}}$ is the proportion of students with mastery status $\alpha_{eLL(EE)}$ for linkage level LL nested within Essential Element EE (Non-Masters = 0; Masters = 1); I_e is the number of items taken by student e for the linkage level; X_{ei} is the binary response of student e to item i; $\alpha_{eLL(EE)}$ is the mastery status indicator for student e; and linkage level LL within Essential Element EE and $P\left(X_{ei} = 1 \middle| \alpha_{eLL(EE)}\right)$ is given by:

$$P\left(X_{ei} = 1 \middle| \alpha_{eLL(EE)}\right) = \frac{\exp(\lambda_{0,LL(EE)} + \lambda_{1,LL(EE)}\alpha_{eLL(EE)})}{1 + \exp(\lambda_{0,LL(EE)} + \lambda_{1,LL(EE)}\alpha_{eLL(EE)})}. \tag{4}$$

In Equation (4), $\lambda_{0,LL(EE)}$ is the fungible item intercept for linkage level LL in Essential Element EE, and $\lambda_{1,LL}$ is the fungible item main effect for linkage level LL in Essential Element EE.

Although the connections between linkage levels were not modeled empirically, they were used in the scoring procedures. In particular, if through this procedure, a student was judged to have mastered a given linkage level within an EE, then the student was assumed to have mastered all lower levels within that EE. This scoring rule relies strongly on the expert opinion used to construct the DLM maps. Validation studies for the structure of the learning map model are currently being planned.

V.7. Model Calibration

The model specified in Equations (3) and (4) was estimated for each linkage level within each EE. Across all grade bands, there were 256 EEs, all with 5 linkage levels, so a total of $256 \times 5 = 1{,}280$ calibration models were run. Each model was estimated using marginal maximum likelihood using *Mplus 7.31* (Muthén & Muthén, 1998-2015).

V.8. DLM Scoring: Mastery Status Assignment

Following calibration, students were assigned mastery in one of two ways: (1) by having a posterior probability of mastery greater than or equal to 0.8 based on the EAP estimate of Equation (2), where the model from Equation (4) was substituted; or (2) having answered 80% of all items administered at the linkage level correctly. Because students often did not test at more than one linkage level within an EE, students who did not meet mastery status for any tested linkage level were assigned mastery status for the linkage level that was two levels below the highest level in which they were tested (unless the highest level tested was either the Initial Precursor or Distal Precursor levels, in which case, students were considered non-masters of all linkage levels within the EE). The scoring method was discussed and approved by the DLM Technical Advisory Committed during the phone call on July 21, 2015.

VI. STANDARD SETTING

The standard setting process for the Dynamic Learning Maps (DLM) Alternate Assessment System in English language arts (ELA) and mathematics consisted of the development of performance level descriptors, a four-day standard setting meeting, and follow-up evaluation of impact data and cut points. The purpose of the standard setting activities was to derive recommended cut points for placing students into four performance levels based on results from the 2014-2015 DLM assessments in ELA and mathematics. This chapter provides a brief description of the development of the rationale for the standard setting approach; the policy performance level descriptors; methods, preparation, procedures and results of the standard setting meeting; and follow-up evaluation of the impact data and cut points.[21] A more detailed description of the DLM standard setting activities and results can be found in the *2015 Year-End Model Standard Setting: English Language Arts and Mathematics* Technical Report #15-03 (Karvonen, Clark, & Nash, 2015). The chapter concludes with a full description of the development of grade and content-specific performance level descriptors (PLDs), which were developed after approval of the consortium cuts.

VI.1. Standard Setting Overview

The 2014-2015 school year was the first fully operational testing year for the DLM assessments in ELA and mathematics. The consortium operational testing window ended on June 12th, 2015, and the DLM staff conducted standard setting June 15 – 18, 2015 in Kansas City, Missouri. The standard setting event was a DLM Consortium-wide event with the purpose of establishing a set of cut points for each of the two testing models. Although state partners 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.

VI.1.A. 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, recommended practices for developing, implementing, evaluating, and documenting standard settings (Cizek, 1996; Hambleton, Pitoniak, & Copella, 2012), and the *Standards on Educational and Psychological Testing* (2014). The DLM standard setting approach used the DLM map and mastery classifications. The panel process drew from several established methods, including generalized holistic (Cizek & Bunch, 2006) and body of work (Kingston & Tiemann, 2012).

Because the DLM assessment is based on large, fine-grained learning map models and makes use of diagnostic classification modeling rather than traditional psychometric methods, DLM's

[21] There are two groups of states within the DLM Consortium that use different testing models: the integrated model (IM) and the year-end model (YE). The same standard setting methods were used for both models, but separate panels were convened consisting of representatives from either IM or YE states. There are separate technical reports and separate versions of this chapter for each model.

standard setting approach relied on aggregation of dichotomous classifications of node and linkage level mastery for each EE in the blueprint. Drawing from the generalized holistic and body of work methods, DLM used a profile approach to classify student mastery of linkage levels into performance levels. Profiles provided a holistic view of student performance by summarizing across the Essential Elements 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 Technical Advisory Committee (TAC) reviewed a preliminary description of the proposed methods. At the TAC's suggestion, the 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 49 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.

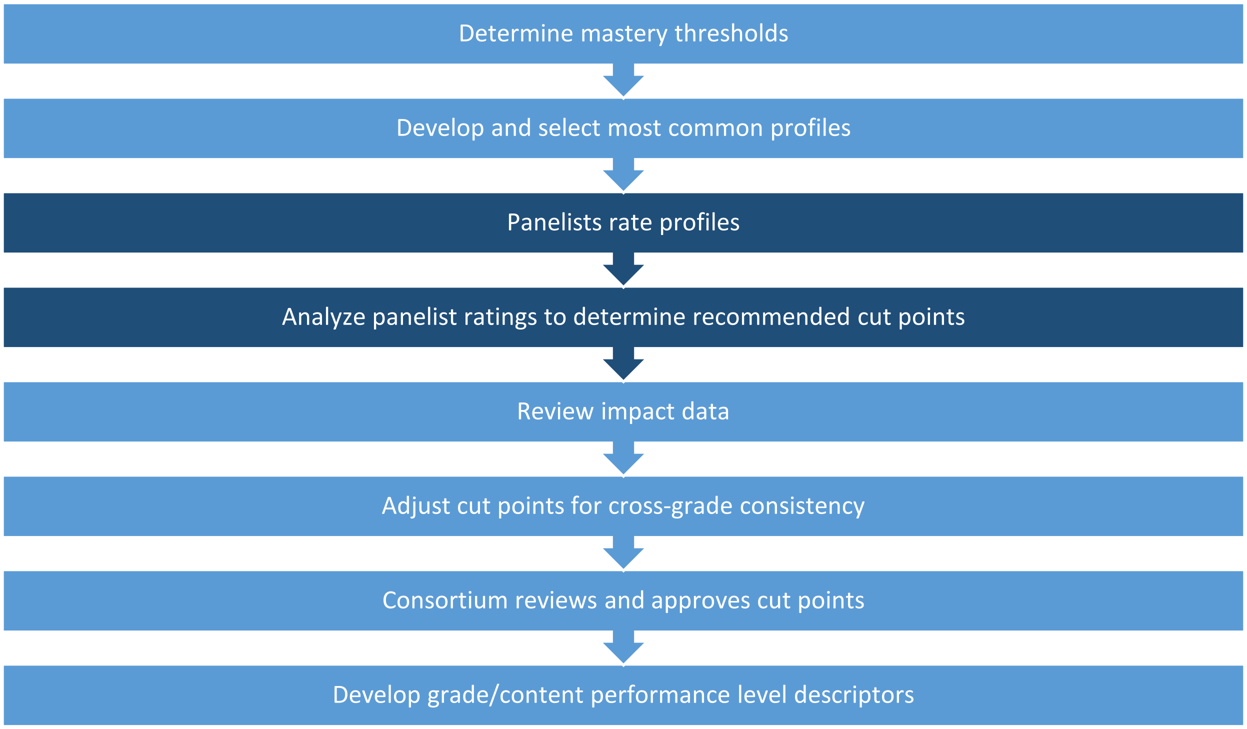

Determine mastery thresholds
Develop and select most common profiles
Panelists rate profiles
Analyze panelist ratings to determine recommended cut points
Review impact data
Adjust cut points for cross-grade consistency
Consortium reviews and approves cut points
Develop grade/content performance level descriptors

Figure 49. *Steps of the DLM standard setting process. Dark shading represents steps conducted at the standard setting meeting in June 2015.*

VI.1.B. Policy Performance Level Descriptors

Student scores are reported as performance levels, and performance level descriptors (PLDs) are used to inform the interpretation of those scores. The DLM state partners 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 state partners in attendance, who have 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, state partners 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 59 presents a summary of 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 59. Final performance level descriptors for 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.

VI.1.C. Profile Development

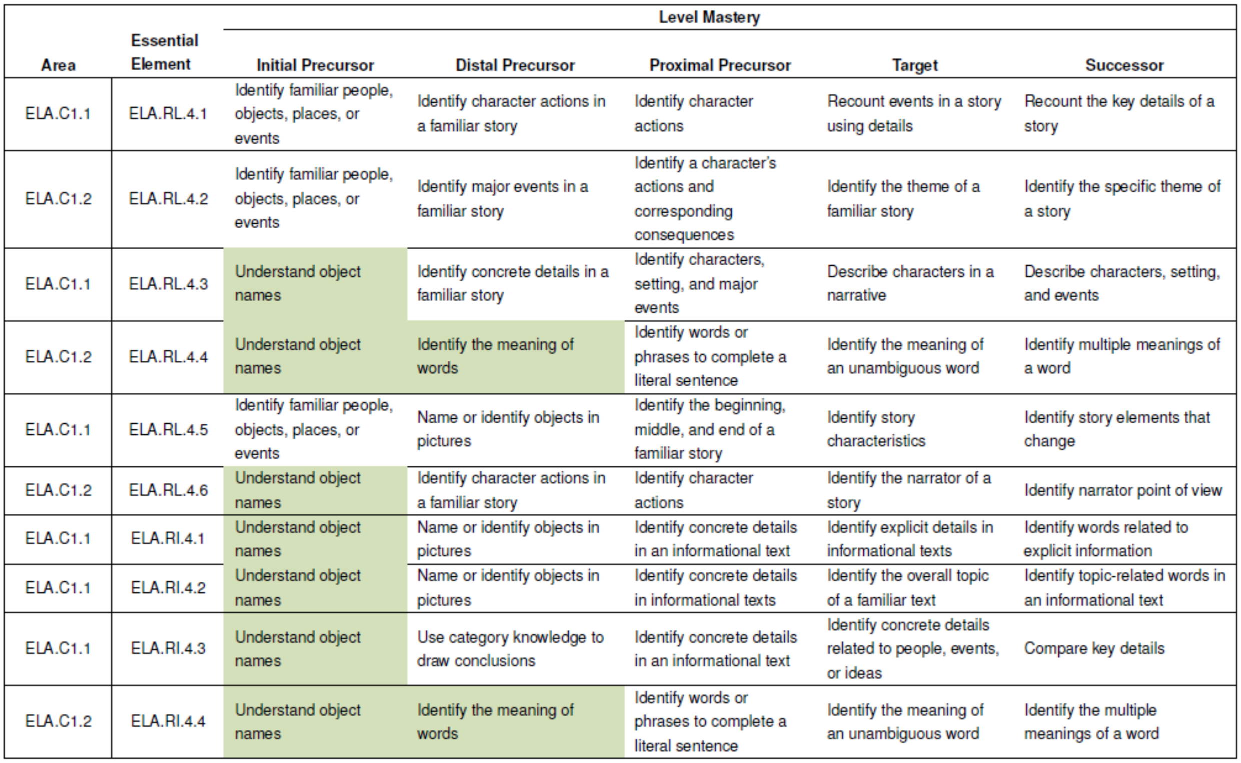
Prior to the standard setting meeting, student performance on nodes in the DLM map was aggregated to create profiles of student learning.

The first step to develop profiles required obtaining mastery classifications at the node level. Based on TAC and state input, an agreed-upon cut was applied to students' posterior probabilities from the diagnostic classification model (DCM) calibration. For each node, all students with a probability greater than or equal to .8 would receive a *node* mastery status of 1, or mastered. All students with a probability lower than .8 would receive a *node* mastery status of 0, or not mastered.[22]

[22] 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.

In the second step, the dichotomous node mastery statuses were then summed for every node the student was assessed on at the linkage level and then divided by the total number of nodes the student was assessed on at the linkage level to obtain the proportion of nodes mastered at the linkage level. This proportion represents *linkage level mastery*. Based on TAC and state partner input, the threshold used to determine *linkage level mastery* was set at .75. Similar to node mastery, a mastery status of 0 or 1 was obtained for each linkage level. Using 0.75 as the cutoff for *linkage level mastery*, all students with a proportion of nodes mastered greater than or equal to .75 would receive a *linkage level mastery* status of 1, or mastered. All students with a proportion of nodes mastered lower than .75 would receive a *linkage level mastery* status of 0, or not mastered.

Finally, the threshold values from step one and step two were applied to create profiles of student mastery, which summarize linkage level mastery by EE. Profiles were created using data for each content area, grade. Each profile listed all the linkage levels for all the EEs from the blueprint, along with the conceptual area for each, with shaded boxes indicating the mastered linkage levels. Figure 50 provides an example 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

Figure 50. Example standard setting profile for a hypothetical student. Green shading represents linkage level mastery.

Profiles were available for all students who participated in the spring window by May 15, 2015 (n_{YE} = 49,958, n_{EOI} = 1,877). 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 or course and content area. In instances where data was not available at a specific linkage level value, (e.g., no students mastered exactly 47 linkage levels for a grade and content area), profiles were based on simulated data. To simulate profiles, the DLM content 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.[23]

VI.1.D. Panelists

The 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 teachers as well as school and district administrators. Other subject matter experts, such as higher education institution faculty or state/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 54 panelists who participated in standard setting represented varying backgrounds. Tables Table 60 and Table 61 summarize their demographic information. Most of the selected panelists were classroom teachers. Panelists had a range of years of experience with mathematics, English language arts, and working students with the most significant cognitive disabilities.

Nearly half of the participants had experience with setting standards for other assessments (28). Some panelists already had experience with the DLM assessment, either from writing items (5) or externally reviewing items and testlets (19). Only two panelists reported having less than one year or no experience with alternate assessments: both were classroom teachers with at least 13 years of experience working with students with the most significant cognitive disabilities.[24]

[23] Further detail on specific procedures for preparing standard setting profiles may be found in Chapter 1 of Technical Report #15-03.

[24] Further detail on standard setting volunteers, selection process, and panel composition may be found in Chapter 3 of Technical Report #15-03.

Table 60. Panelist Demographic Characteristic

Demographic Category	Count
Gender	
Female	50
Male	4
Race	
African American	5
American Indian/Alaska Native	1
Asian	2
Hispanic/Latino	1
Native Hawaiian/Pacific Islander	0
White	42
Not Disclosed	4
Professional Role	
Classroom Teacher	37
Building Administrator	4
District Staff	5
University Faculty/Staff	1
Other	3
Total	**54**

Table 61. Panelist Years of Experience

Experience Type	*M*	Min	Max
Students with Significant Cognitive Disabilities	15.7	1.0	36.0
Mathematics	20.6	3.0	50.0
English Language Arts	16.6	1.0	35.0

VI.1.E. 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 where 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, the 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.[25]

VI.1.E.i. Training

Panelists were provided with training both before and during the standard setting workshop. Advance training was available online, on demand in the ten 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 maps, Essential Elements, claims and conceptual areas, linkage levels, and alignment
3. Accessibility by design, including the framework for the DLM Alternate Assessment System’s cognitive taxonomy and strategies for maximizing accessibility of the content; the use of the Access (Personal Needs and Preferences) Profile (PNP) 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
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

[25] Further information regarding all meeting procedures and fidelity of the final procedures to the planned procedures can be found in Technical Report #15-03 (Chapter 4, Appendix J).

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 related training before beginning the practice activity.

VI.1.E.ii. 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) where 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 round two ratings, built-in logistic regression functions calculated the probability of a profile being categorized in each performance level, conditioned on number of linkage levels mastered, and the most likely cut points for each performance level were identified. In instances where 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 approximate cut point based on the panelist recommendations.[26]

VI.1.E.iii. Pinpointing

Pinpointing rounds followed after 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 number of linkage levels mastered within the general ranges determined in range finding, not relying on conjunctive or compensatory judgments.

First, panelists reviewed profiles for the seven 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.

[26] Technical report #15-03 (Chapter 4) provides greater detail on range finding and pinpointing and includes details the number of linkage levels per grade and content area.

Once all panelists 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 rating commenced. Panelists were given the opportunity to adjust their independent ratings if they chose. Using the second round's ratings, built-in logistic regression functions calculated the probability of a profile being categorized in each performance level, conditioned on number of linkage levels mastered, and the most likely cut points for each performance level were identified. In instances where 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 the panelist recommendations.[5]

VI.1.E.iv. Panelist Evaluations of Panel-Recommended Cut Points

Across all panelists, panels, grades, and cut points (*N*=483), in 94.2% of cases panelists indicated that they were comfortable with the group-recommended cut point. Table 62 provides the panelist comfort with group recommended cut points.[7] Only 5.5% of responses (*n* = 26) indicated a discomfort with a group-recommended cut. Panelist comfort with all three recommended cut points was found for 12 out of 23 cut point panels (52.2%). 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.[27]

Table 62. Panelist Comfort with Group Recommended Cut Points

Content Area	***N*** **Panelists**	***N*** **Ratings** **(*n* Panelists x *n* Cut Points Evaluated)**	***n*** **"Yes" Ratings**	**% Agreement**
ELA	80	240	230	96
Math	81	243	225	93

VI.1.E.v. Adjusting 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, statistical adjustments were made to the panel-recommended cut points

[27] Technical report #15-03 (Chapter 5) provides greater detail on final independent evaluations of panel-recommended cut points.

in an effort to systematically smooth distributions within the system of cut points being considered. No adjustments were made for EOI because both the standards assessed and students taking these assessments were assumed to be very different from one course to another.[28]

VI.1.F. Results

The panel-recommended and statistically adjusted cut points as well as impact data and evaluation results are summarized.[29]

VI.1.F.i. Panel Recommended and Adjusted Cut Points

Table 63 includes a summary of the cut point recommendations reached by the panelists following the range-finding and pinpointing process.

Table 63. Final ELA and Math Panel Cut Point Recommendations

Content Area and Grade	Emerging/ Approaching	Approaching/ Target	Target/ Advanced	Minimum Required Linkage Levels
ELA				
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

[28] The specific steps applied to each subject within each grade level can be found in Technical Report #15-03 (Chapter 5).

[29] Additional detailed results are provided in Technical Report #15-03 (Chapter 5).

Content Area and Grade	Emerging/ Approaching	Approaching/ Target	Target/ Advanced	Minimum Required Linkage Levels
English 2	21	45	54	60
English 3	23	38	53	65
Math				
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
Algebra I	18	25	33	45
Algebra II	17	25	34	45
Geometry	14	20	30	40

To mitigate the effect of sampling error and issues related to a system of cut points across a series of grade levels, statistical adjustments were made to the panel-recommended cut points into systematically smooth distributions within the system of cut points being considered.

Table 64 summarizes the adjusted cut points that used the methods described above and the impact data for those adjusted cut points.

Table 64. Adjusted Cut-Point Recommendations

Content Area and Grade	Emerging/ Approaching	Approaching/ Target	Target/ Advanced	Minimum Required Linkage Levels
ELA				
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
English 2	21	45	54	60
English 3	23	38	53	65
Math				
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

Content Area and Grade	Emerging/ Approaching	Approaching/ Target	Target/ Advanced	Minimum Required Linkage Levels
Algebra I	18	25	33	40
Algebra II	17	25	34	45
Geometry	14	20	30	40

VI.1.F.ii. Final Impact Data

Figure 51, Figure 52, and Figure 53 display the results of the adjusted cut points in terms of impact for English language arts and mathematics and EOI courses, respectively.[30] Table 7 includes the demographic data for students included in the impact data.

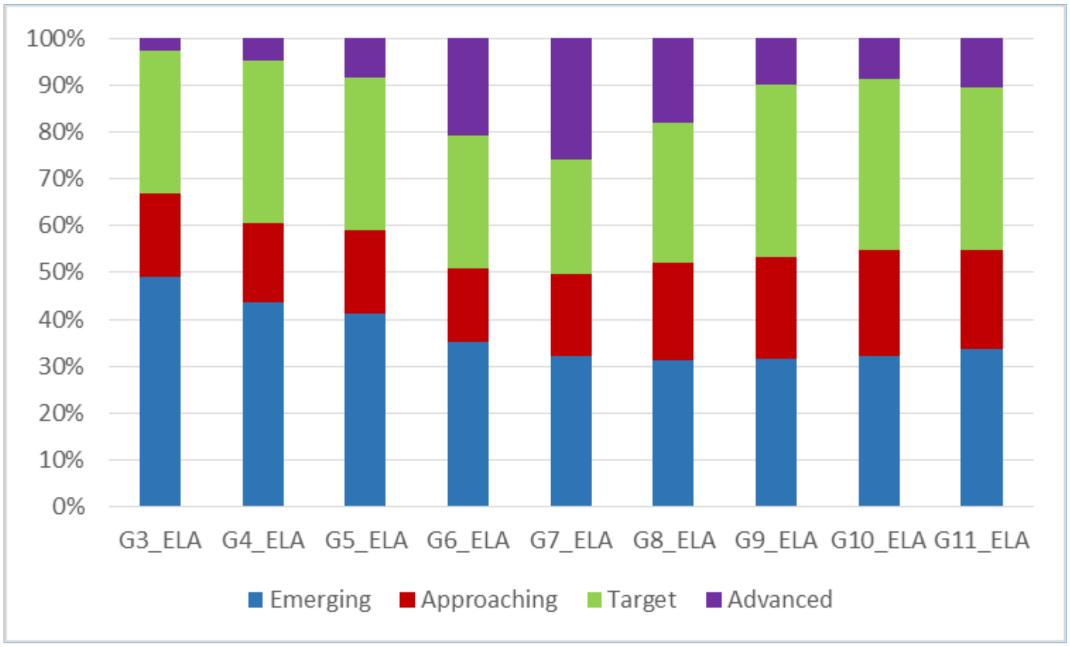


Figure 51. English language arts impact data using adjusted cut points.

[30] Technical Report #15-03 (Chapter 5) reports the frequency distributions for the panel-recommended cut points.

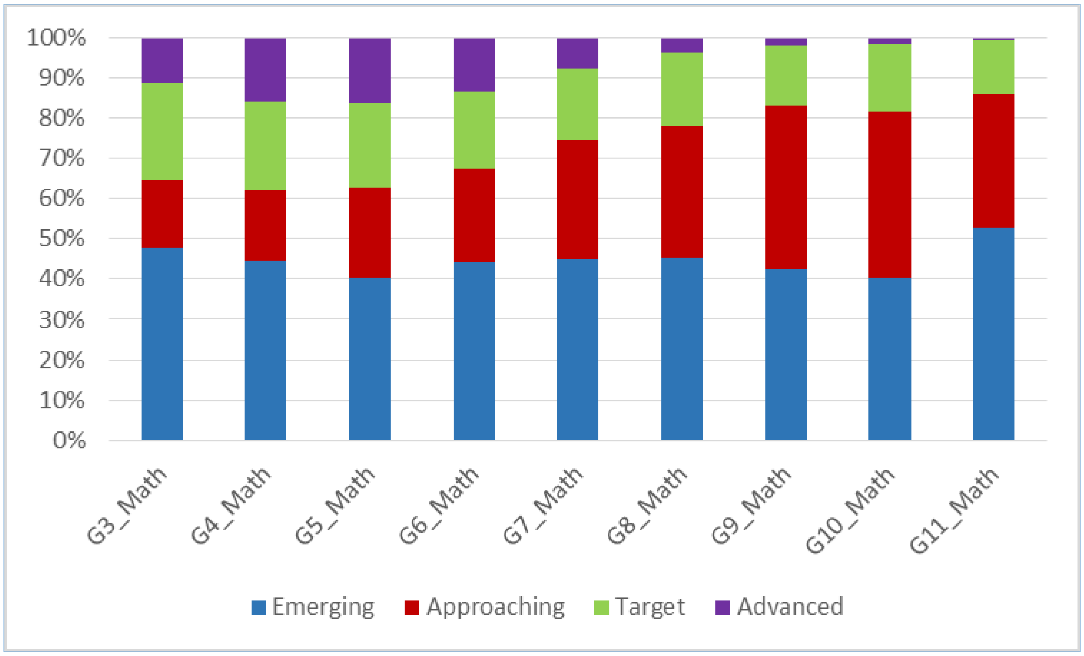


Figure 52. *Mathematics impact data using adjusted cut points.*

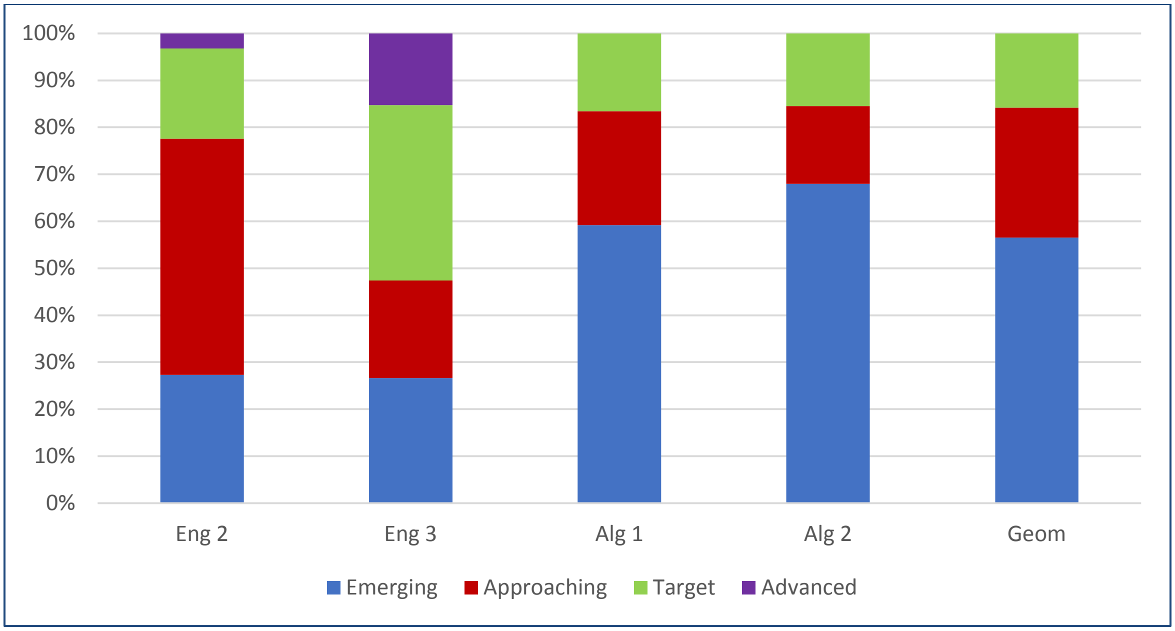


Figure 53. *EOI course impact data using adjusted cut points.*

Table 65. Demographic Information for Students Included in Impact Data

Demographic	*n*	%
Gender		
Female	18068	33.33
Male	35132	64.81
Missing	1004	1.85
Primary Disability		
Intellectual disability	6250	11.53
Autism	5097	9.40
Other health impairment	4526	8.35
Multiple disabilities	1455	2.68
Specific learning disability	637	1.18
(Other)	2429	4.48
Missing	33810	62.38
Comprehensive Race		
White	28459	52.5
African American	9309	17.17
Asian	1645	3.03
American Indian	3723	6.87
Alaska Native	244	0.45
Two or More Races	6741	12.44
Native Hawaiian/Pacific Islander	388	0.72
Missing	3695	6.82
Hispanic Ethnicity		
No	22776	42.02
Yes	4790	8.84
Missing	26638	49.14
ESOL Participation		

Demographic	*n*	%
Not an ESOL eligible student and not an ESOL monitored student	51547	95.15
ESOL eligible or monitored student	2630	4.85
Missing	27	<. 01
ELA Band		
Foundational	9306	17.17
Band1	15324	28.27
Band2	19636	36.23
Band3	9937	18.33
Missing	1	<. 01
Math Band		
Foundational	10282	18.97
Band1	16433	30.32
Band2	20398	37.63
Band3	7089	13.08
Missing	2	< .01
Total	54204	

VI.1.F.iii. 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.[31]

The panel-recommended cut points, adjusted cut points, and associated impact data for both sets of cut points were presented to the TAC and consortium states for review. The TAC accepted the DLM adjustment method and resulting adjusted cut points. Following the states' review process and discussion with the DLM team, the states voted to accept the DLM-recommended adjusted cut points as the final consortium cut points with no further adjustment.

[31] The TAC chair memorandum and TAC resolution are provided in Technical Report #15-03 (Appendix L).

VI.1.G. GRADE LEVEL/CONTENT PERFORMANCE LEVEL DESCRIPTORS

Based on the general approach to standard setting, which relied on mastery profiles to anchor panelists' content-based judgments, grade and content-specific performance level descriptors (PLDs) were not used during standard setting. Instead, they emerged based on the final cut points and were syntheses of content from the more fine-grained linkage level descriptors. Grade and content specific performance level descriptors 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 the 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 content teams as they developed language for grade and content 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. Content 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 in order to consider grain size of descriptive language and variety of formats for publication. In addition to the draft lists generated by standard setting panelists, content teams used the following materials as they drafted specific language for each grade and content specific PLD:

- The DLM test blueprint
- The cut points set at standard setting for each grade
- Sample mastery profiles used as part of standard setting
- Essential Element Concept Maps (EECMs) for each Essential Element (EE) included on the blueprint for each grade level
- Linkage Level descriptions and associated sections of the DLM maps for every EE
- For the math team: *The Standards of Mathematical Practice*

Content teams reviewed the EEs, EECMs, 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 content teams reviewed the draft skill lists created by standard setting panelists and final cut points approved by the consortium. Content teams then used to 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 content 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 maps, the synthesis of standard setting panelist judgments and content team judgments provided the basis for descriptions to describe the typical performance of students showing mastery at each performance level. As content teams drafted PLDs for each grade, they reviewed the descriptors in relationship to each other and the underlying DLM map to ensure that there was differentiation in skills from one grade to the next. In very few cases, where panelists recommended skill placement that was inconsistent with development of content knowledge as represented in the DLM maps, content teams adjusted the placement of skills. This was only done in cases where 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.

The DLM staff prepared initial drafts of the grade and content specific descriptions for grade 3. Project staff reviewed these drafts internally. Additional drafts were prepared for grade 4 and 5. The DLM state partners reviewed a combination of grades 3, 4 and 5 at the December 2015 consortium governance meeting. Project staff asked state partners 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 state partners focused on utility for educators and parents and structuring the descriptions to make them more user-friendly. The primary responses to state partner feedback were to:

- review technical language in existing drafts and simplify wherever possible,
- organize each grade and content 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, and
- organize descriptions consistently within and across grades so that related skills were described in the same sequence within each level in a grade.

The DLM staff delivered drafts of all grade and content specific descriptions to state partners for review in February 2016. After the review period ended, content 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. Final versions of the grade and content PLDs are available on the DLM website (http://dynamiclearningmaps.org/content/assessment-results). Appendix D.1 contains examples of grade and content PLDs.

VII. ASSESSMENT RESULTS

Following from the discussion of the standard-setting process in Chapter VI, Chapter VII reports the 2014–2015 operational results of the Dynamic Learning Maps® (DLM®) alternate assessment. This chapter presents student participation data, final results in terms of the percent of students at each performance level (impact), and subgroup performance by gender, race, ethnicity, and English language learner (ELL) status. This chapter also reports the distribution of students by the highest linkage level mastered. Finally, this chapter and Appendix E describe all types of score reports, data files, and interpretive guides.

VII.1. Student Participation

The 2014–2015 spring summative assessments were administered to a total of 52,760 students, including states administering End-of-Instruction (EOI) courses and districts affiliated with the Bureau of Indian Education (BIE), as shown in Table 66. The 623,289 assessment sessions were administered by 13,841 educators in 7,993 schools and 2,470 school districts.

Table 66. Student Participation by State or Agency

State	Students
Alaska	692
BIE-Affiliated Districts	15
Colorado	5,476
Illinois	11,915
Mississippi	3,772
New Hampshire	793
New Jersey	9,900
Oklahoma	6,003
Utah	4,432
West Virginia	2,571
Wisconsin	7,191
Total	52,760

Note: Results are for the 2014–2015 spring administration.

In grades 3 through 8, over 6,500 students participated in each grade (see Table 67). In high school, the largest number of students participated in grade 11 and the smallest number participated in grade 12. The differences in grade-level participation can be traced to differing state-level policies about the grade in which students are assessed in high school.

Table 67. Student Participation by Grade

Grade	Students
3	6,575
4	6,774
5	6,932
6	7,190
7	6,991
8	6,918
9	2,703
10	2,805
11	5,269
12	603

Note: Results are for the 2014–2015 spring administration.

Table 68 summarizes the demographic characteristics of students who participated in the spring 2014-2015 administration. The majority of participants were male (64%) and white (52%). Only 5% of students were labeled as being eligible for or monitored for ELL services.

Table 68. Demographic Characteristics of Participants

Subgroup	*n*	%
Gender		
Female	17,569	33.30
Male	34,225	64.87
Missing	966	1.83
Race		
White	27,694	52.49
African American	9,089	17.23
Asian	1,615	3.06
American Indian	3,592	6.81
Alaska Native	225	0.43
Two or more races	6,591	12.49
Native Hawaiian or Pacific Islander	371	0.70
Missing	3,583	6.79
Hispanic Ethnicity		
No	22,067	41.83
Yes	4,681	8.87
Missing	26,012	49.30
English Language Learner (ELL) Participation		
Not ELL eligible or monitored	50,147	95.05
ELL eligible or monitored	2,613	4.95
Missing	0	0.00

VII.2. STUDENT PERFORMANCE

Student performance on DLM assessments is interpreted using cut points, determined during standard setting (see Chapter VI), which separate student scores into four performance levels. A student receives a performance level based on the total number of linkage levels mastered across the assessed Essential Elements (EEs).

As previously described (see Chapter VI), students were considered masters of a linkage level if (1) their posterior probability from the diagnostic classification model (DCM) was greater than or equal to .8, or (2) the proportion of items that they responded to correctly within the linkage level was greater than or equal to .8. If the student did not demonstrate mastery at the level assessed, mastery was assigned two linkage levels below the level assessed. In addition, students were considered masters of all linkage levels below the level at which they demonstrated mastery.

Mastery status values were aggregated within and across EEs to obtain the total number of linkage levels the student mastered. Although the total number of mastered linkage levels is not a raw or scale score, the number of linkage levels mastered across EEs assessed was the metric used for setting performance level cut points.

For the 2014–2015 administration, student performance was reported using the four performance levels approved by the DLM Consortium:

- 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.

VII.2.A. OVERALL PERFORMANCE

Results of the standard setting are described in detail in Chapter VI.

Table 69 reports the impact data (i.e., the percent of students at each performance level) from the 2014–2015 spring administration for ELA and mathematics.

The percent of students who achieved at the Target or Advanced performance levels in ELA was over 30% in grade 3 and over 40% for all other grades. For the English 2 EOI course, just over 22% of students met or exceeded Target expectations, while a larger portion (approximately 53%) achieved at the Target or Advanced levels in English 3. In mathematics, over 30% achieved at the Target or Advanced levels in grades 3-6; in grade 7 and above, the percent of students who met or exceeded the Target performance level ranged from 13% to 25%.

For the Algebra 1, Algebra 2, and Geometry EOI courses, approximately 15% reached the Target performance level and no students achieved at the Advanced level.

Table 69. Percent of Students by Grade and Performance Level

Grade/Course	Emerging (%)	Approaching (%)	Target (%)	Advanced (%)	Target/Advanced (%)
ELA					
3	48.7	17.4	31.1	2.9	33.9
4	42.9	17.0	35.3	4.8	40.1
5	40.7	17.7	33.0	8.5	41.5
6	34.8	16.2	27.9	21.1	49.0
7	32.8	17.3	24.6	25.3	49.9
8	31.9	19.6	30.9	17.7	48.5
9	31.6	21.9	36.7	9.7	46.5
10	32.2	22.2	36.5	9.1	45.6
11	34.1	21.7	34.1	10.2	44.3
English 2	27.3	50.3	19.2	3.2	22.4
English 3	26.6	20.8	37.3	15.3	52.6
Mathematics					
3	47.5	16.8	24.1	11.5	35.6
4	44.2	17.3	22.3	16.2	38.5
5	42.8	20.9	20.4	15.8	36.2
6	44.5	22.8	19.2	13.5	32.7
7	45.6	29.4	16.4	8.6	25.0
8	45.9	32.6	17.2	4.3	21.5
9	42.3	40.8	15.5	1.4	16.9
10	40.5	39.0	19.2	1.4	20.6
11	54.1	32.8	12.9	0.2	13.2
Algebra 1	59.2	24.2	16.6	0.0	16.6
Algebra 2	68.0	16.5	15.5	0.0	15.5
Geometry	56.5	27.7	15.8	0.0	15.8

VII.2.B. SUBGROUP PERFORMANCE

Impact data for subgroups, including groups based on gender, race, ethnicity, and ELL status, was computed to set a baseline for the evaluation of achievement gaps in future years.

The distribution of students across performance levels was examined by demographic subgroup. Table 70 and Table 71 summarize the disaggregated frequency distributions for ELA and mathematics, respectively, collapsed across all assessed grades. Although states 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 70. Students at Each ELA Performance Level by Demographic Group

	Emerging		Approaching		Target		Advanced		Not Assessed*	
Subgroup	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Gender										
Female	6,294	35.82	3,211	18.28	5,373	30.58	2,124	12.09	567	3.23
Male	12,457	36.40	6,079	17.76	10,326	30.17	4,275	12.49	1,088	3.18
Missing	332	34.37	157	16.25	295	30.54	151	15.63	31	3.21
Race										
White	9,785	35.33	4,955	17.89	8,590	31.02	3,640	13.14	724	2.61
African American	3,110	34.22	1,608	17.69	2,748	30.23	1,093	12.03	530	5.83
Asian	837	51.83	295	18.27	348	21.55	117	7.24	18	1.11
American Indian	1,128	31.40	585	16.29	1,097	30.54	509	14.17	273	7.60
Alaska Native	108	48.00	52	23.11	57	25.33	8	3.56	n/a	n/a
Two or more races	2,760	41.88	1,285	19.50	1,884	28.58	623	9.45	39	0.59
Native Hawaiian or Pacific Islander	171	46.09	67	18.06	90	24.26	32	8.63	11	2.96
Missing	1,184	33.04	600	16.75	1,180	32.93	528	14.74	91	2.54
Hispanic Ethnicity										
No	7,109	32.22	3,786	17.16	7,015	31.79	3,097	14.03	1,060	4.80
Yes	1,669	35.65	895	19.12	1,471	31.42	572	12.22	74	1.58

	Emerging		Approaching		Target		Advanced		Not Assessed*	
Subgroup	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Missing	10,305	39.62	4,766	18.32	7,508	28.86	2,881	11.08	552	2.12
English Language Learner (ELL) Participation										
Not ELL eligible or monitored	18,013	35.92	8,942	17.83	15,262	30.43	6,292	12.55	1,638	3.27
ELL eligible or monitored	163	35.43	82	17.83	149	32.39	50	10.87	16	3.48
Missing	164	33.81	92	18.97	170	35.05	56	11.55	3	0.62

Note: N = *52,760.*
**The student was not assessed on any Essential Elements in that content area.*

Table 71. Students at Each Math Performance Level by Demographic Group

	Emerging		Approaching		Target		Advanced		Not Assessed*	
Subgroup	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Gender										
Female	8,256	46.99	4,551	25.90	3,123	17.78	1,480	8.42	159	0.91
Male	15,111	44.15	8,464	24.73	6,506	19.01	3,754	10.97	390	1.14
Missing	388	40.17	215	22.26	203	21.01	133	13.77	27	2.80
Race										
White	12,311	44.45	7,221	26.07	5,196	18.76	2,689	9.71	277	1.00
African American	4,028	44.32	2,233	24.57	1,744	19.19	997	10.97	87	0.96
Asian	961	59.50	310	19.20	227	14.06	106	6.56	11	0.68
American Indian	1,364	37.97	781	21.74	751	20.91	588	16.37	108	3.01
Alaska Native	121	53.78	64	28.44	29	12.89	10	4.44	1	0.44
Two or more races	3,349	50.81	1,690	25.64	1,093	16.58	426	6.46	33	0.50
Native Hawaiian or Pacific Islander	204	54.99	84	22.64	48	12.94	33	8.89	2	0.54
Missing	1,417	39.55	847	23.64	744	20.76	518	14.46	57	1.59

	Emerging		Approaching		Target		Advanced		Not Assessed*	
Subgroup	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Hispanic Ethnicity										
No	9,143	41.43	5,683	25.75	4,472	20.27	2,487	11.27	282	1.28
Yes	1,993	42.58	1,178	25.17	928	19.82	541	11.56	41	0.88
Missing	12,619	48.51	6,369	24.48	4,432	17.04	2,339	8.99	253	0.97
English Language Learner (ELL) Participation										
Not ELL eligible or monitored	22,492	44.85	12,590	25.11	9,355	18.66	5,146	10.26	564	1.12
ELL eligible or monitored	190	41.30	113	24.57	104	22.61	51	11.09	2	0.43
Missing	210	43.30	129	26.60	98	20.21	46	9.48	2	0.41

Note: N = 52,760

**The student was not assessed on any Essential Elements in that content area.*

VII.2.C. Linkage Level Mastery

This section of the chapter summarizes the average distribution of students by linkage level mastered across all EEs in the grade and content area. For each EE, a student can demonstrate mastery of any of the five linkage levels. If the student does not master any of the linkage levels, the student's score report will indicate no evidence of mastery for the EE.

Table 72 and Table 73 report the average distribution of students according to linkage level mastered across all EEs for ELA and mathematics, respectively. For ELA, the average percent of students who mastered the Target or Successor linkage level ranged from approximately 26% in grade 3 to 38% in English 3. For mathematics, the average percent of students who mastered the Target or Successor linkage level ranged from approximately 6% in grade 11 to 24% in Algebra 1.

Table 72. Average Percent of Students in Each Grade by ELA Linkage Level Mastered

	Linkage Level					
Grade/Course	**No Evidence (%)**	**IP (%)**	**DP (%)**	**PP (%)**	**T (%)**	**S (%)**
3 (*n* = 6,684)	20.4	13.9	17.0	22.3	13.4	13.0
4 (*n* = 6,922)	20.4	10.7	14.0	22.6	14.1	18.2
5 (*n* = 7,064)	19.8	12.2	16.2	19.3	13.4	19.1
6 (*n* = 7,358)	20.1	11.9	13.7	20.9	14.1	19.3
7 (*n* = 7,135)	19.1	11.4	13.3	21.7	12.4	22.1
8 (*n* = 7,067)	20.6	10.4	12.7	22.9	14.5	18.9
9 (*n* = 2,546)	21.7	11.8	11.4	28.5	15.3	11.3
10 (*n* = 2,273)	23.2	12.8	10.6	24.7	14.0	14.7
11 (*n* = 5,099)	22.4	12.5	12.4	25.1	13.4	14.2
English 2 (*n* = 1,275)	16.9	11.7	10.2	31.7	18.0	11.5
English 3 (*n* = 251)	7.8	7.7	10.3	36.5	20.7	17.0

Note: IP = Initial Precursor; DP = Distal Precursor; PP = Proximal Precursor; T = Target; S = Successor.

Table 73. Average Percent of Students in Each Grade by Mathematics Linkage Level Mastered

	Linkage Level					
Grade/Course	**No Evidence (%)**	**IP (%)**	**DP (%)**	**PP (%)**	**T (%)**	**S (%)**
3 (*n* = 6,684)	32.4	23.9	16.5	14.6	6.8	5.8
4 (*n* = 6,922)	27.0	22.0	17.7	16.7	8.8	7.8
5 (*n* = 7,064)	30.4	22.4	17.7	15.2	7.8	6.5
6 (*n* = 7,358)	28.5	19.3	18.4	15.4	9.1	9.3
7 (*n* = 7,135)	25.9	22.8	20.1	16.5	8.6	6.1
8 (*n* = 7,067)	30.5	20.5	17.8	15.3	11.2	4.7
9 (*n* = 2,546)	25.7	27.3	23.0	15.3	5.2	3.5
10 (*n* = 2,273)	33.5	25.8	20.0	12.8	5.7	2.2
11 (*n* = 5,099)	39.8	31.3	19.1	4.3	4.3	1.2
Algebra 1 (*n* = 1,323)	21.2	16.5	16.9	21.2	11.9	12.3
Geometry (*n* = 142)	8.9	21.6	23.9	22.2	9.8	13.6
Algebra 2 (*n* = 54)	16.3	22.4	19.9	19.6	13.4	8.4

Note: IP = Initial Precursor; DP = Distal Precursor; PP = Proximal Precursor; T = Target; S = Successor.

VII.3. SCORE REPORTS

Assessment results were provided to all DLM member states to be reported to parents/guardians and to educators at state and local education agencies. Individual reports were provided to educators and parents/guardians. Several aggregated reports were provided to state and local education agencies.

VII.3.A. INDIVIDUAL REPORTS

Individual student score reports were developed through a series of focus groups 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 state partners and revised based on multiple rounds of feedback. Refined prototypes were shared with parents/guardians, advocates, and educators through additional focus groups (Clark, Karvonen, Kingston, Anderson, & Wells-Moreaux, 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 on each conceptual area and for the subject overall, 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.

The performance levels reported on the Performance Profile are Emerging, Approaching the Target, At Target, and Advanced. These labels, which reflect a student's overall performance, were determined through a standard-setting process in summer 2015. The Performance Profile also reports the percent of skills, or linkage levels, the student mastered within each conceptual area. Bulleted lists of the skills mastered follow the results reported for the conceptual area. The Learning Profile shows each EE separated into the five linkage levels: Initial Precursor, Distal Precursor, Proximal Precursor, Target, and Successor. Shading indicates the levels that the student mastered. Sample individual student score reports are provided in Appendix F.

VII.3.B. AGGREGATED 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 aggregated reports are provided in Appendix F.

VII.3.C. INTERPRETATION RESOURCES

At the onset of DLM, 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. Achievement (or performance) level descriptors provide useful information about student achievement.
3. Inferences regarding student achievement, progress, and growth 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 were provided to aid score interpretation:

- The "Parent Interpretive Guide" was designed to provide definition and context to student score reports.
- Parent/guardian letter templates were developed within the DLM Consortium to be used by educators and state superintendents to introduce the student reports to parents/guardians.
- The "Teacher Interpretive Guide" was designed to support educators' discussions and build understanding for parents/guardians and other stakeholders.
- The "Scoring and Reporting Guide for Administrators" targeted building and district-level administrators.

- All of the resources listed above were compiled on a webpage, “Scoring and Reporting Resources” (http://dynamiclearningmaps.org/srr/ye). This page also contained an overview of scoring, score-report delivery, and data files. The overview was intended for state education agency staff who would be receiving DLM assessment results but did not have a lot of familiarity with the assessment. Finally, the resources page hosted score report prototypes for individual score reports and class, school, district, and state aggregated reports.

VII.3.C.i. Parent Interpretive Guide

The “Parent Interpretive Guide” (see Appendix E.3) 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 Essential Elements. The guide goes on to describe how EEs detail what the individual student should know and be able to do at a particular grade level. In addition, the guide clarifies that students took assessments in ELA and mathematics and that this report describes how the student performed on the assessment.

Since the Performance Profile section reports overall results in terms of the four performance levels, the sample report explains these performance level descriptions. The sample report clarifies that *At Target* means the student has met the alternate achievement standards in a given subject area at grade level. The Performance Profile goes on to define conceptual areas and relates the student performance to those conceptual areas. Finally, the Performance Profile describes specific academic skills that the student demonstrated on the assessment within the context of grade-level academic content.

The sample report also provides additional information about the Learning Profile. The sample report shows how this section identifies what the student can do to build on the skills and knowledge demonstrated in the assessment and progress toward more complex grade-level skills. The Learning Profile uses colored shading to illustrate which skills the student mastered and which skills were assessed but not mastered. Finally, the sample Learning Profile clarifies the target for performance using a bull’s eye symbol to mark the Target performance level.

VII.3.C.ii. Parent Letters

The DLM Consortium developed templates for explanatory letters that educators and state superintendents could use to introduce parents/guardians to the student reports (see Appendix F). These letters provided context for the reports, including what the DLM assessment is, when it was administered, and what results tell 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’ IEP teams have established.

VII.3.C.iii. Teacher Interpretive Guide

An interpretive guide was provided for educators who would discuss results with parents/guardians or other stakeholders. The guide, "Talking to Parents about the DLM Score Reports," walked educators through directions for getting ready for a parent/guardian meeting, discussing the score report, and finding additional information. See Appendix F for the complete guide.

VII.3.C.iv. Scoring and Reporting Guide for Administrators

The guide designed for principals and district administrators covered each type of report provided for DLM assessments and explained how reports would be distributed. The guide explained the contents of each report and provided hints about interpretation. See Appendix F for the complete guide.

VII.4. Data Files

Three data files, made available to DLM states, summarized results from the 2014–2015 year. 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. The Date/Time Supplemental File provided date/time stamps for the start and end times of each student test session for each EE assessed. Finally, the Incident File listed students who were affected by one of the known problems with testlet assignments during the spring 2015 window.

The GRF, the Date/Time Supplemental File, and the Incident File organized information into columns with student records in rows. If combined, the number of columns was too large for some software to read. Therefore, the GRF and supplemental files were provided separately and followed different structures. The file structures for each of these files were located on the online scoring and reporting resource page. For more information, see "File Structure Data Dictionary" in Appendix F.

A sample GRF with ten fictitious records was provided to DLM state members during the 2014–2015 year to assist in the preparation of software and data systems within each state. A "Guide to Scores & Reports" was also provided (see Appendix E.7). The structures of the GRF and supplemental files were also discussed on several partner calls to orient state members to their formats.

After standard setting, student performance levels for each subject were added to the GRF and files were distributed to state partners. Each member state determined how the DLM performance levels translated into their own definitions of proficient or not proficient. Individual states applied their accountability measures to the GRF to determine proficient and non-proficient status for accountability purposes.

VII.4.A. QUALITY CONTROL PROCEDURES FOR DATA FILES AND SCORE REPORTS

Quality control procedures were implemented for all three data file types. To ensure that formatting and the order of columns were identical, column names in each file were compared with the data dictionary that was provided to states. Additional file-specific checks were conducted to ensure accuracy of all data files.

Upon its creation, each state's GRF was checked against a variety of sources to ensure that the information provided was accurate and complete. The students listed in the GRF were checked against those listed in Educator Portal, the online site where educators enroll and roster students. Each state's GRF was also checked to ensure it only included students belonging to that specific state.

For 10% of the EEs, the values in the EE columns of the GRF were recalculated manually from the original scoring file to check against the values reported in the GRF. All performance levels were manually recalculated based on EE mastery status values and compared to the printed GRF levels. Records for Oklahoma high school students enrolled in multiple EOI courses were checked to ensure that each row contained EE values for only one course.

The Date/Time Supplemental File was compared with various sources as well. The student and EE values in the Date/Time Supplemental File were compared with the GRF to ensure accuracy and completeness of the records. Start and end dates contained in the Date/Time Supplemental File were checked to ensure that all dates reported fell within the testing window (March 16 to June 12, 2015). Additionally, EEs listed in the Date/Time Supplemental File for a given student were compared with the EEs in the GRF to verify that they were the same.

In addition to the data files, individual student and aggregated score reports were generated and checked for quality. Given the large number of score reports generated, a random sample of approximately 1–2% of the score reports generated were checked.

For this sample, both the Performance Profile and the Learning Profile portions of the individual student score reports were checked for accuracy. Performance Profiles were checked to make sure the correct performance level displayed and matched with the value in the GRF. The percent of skills mastered in the Performance Profile was compared against the GRF and the Learning Profile portion of the student score report to ensure that all three contained the same values. Additionally, the number of conceptual areas listed in the Performance Profile were compared with the blueprint. For each EE on the student's Learning Profile, the highest linkage level mastered was 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 was compared against the number listed in the blueprint for that subject and grade or course. Demographic information in the header of the Performance Profile and Learning Profile was checked to ensure that it matched values in the GRF. Formatting and text within each report was given an editorial review as well.

Aggregated reports underwent 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) were present and that no extraneous students were listed. Performance levels (for class and school reports) and the number of students with a given performance level (for district and state reports) were checked against the GRF.

Once all reports were checked, all files to be disseminated to states underwent a final set of checks to ensure that all files were present. This last set of checks involved higher level assurances that the correct number of district files were present for each type of report according to the expected number calculated from the GRF, that file naming conventions were followed, that all types of data files were present, and that all student reports had been generated.

All errors identified during quality-control checks were corrected prior to distribution of data files and score reports to states.

VIII. RELIABILITY

The Dynamic Learning Maps (DLM) Alternate Assessment System uses non-traditional psychometric models (diagnostic classification models) to produce student score reports. As such, evidence for the reliability of scores[32] is based on methods that are commensurate with the models used to produce score reports. As details on modeling are found in Chapter V, this chapter discusses the methods used to estimate reliability, the factors that are likely to affect the variability in reliability results, and an overall summary of reliability results.

The reliability information presented in this chapter adheres to guidance given in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014). Simulation studies 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 DLM reliability evidence reported here supports the "interpretation for each intended score use," as Standard 2.0 dictates (AERA et al., 2014, p. 42). The "appropriate evidence of reliability/precision" (AERA et al., 2014, p. 42) was assembled using a nontraditional 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. Information about alignment with individual standards is provided throughout this chapter.

VIII.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 due to chance from one test administration to another. 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. When a test is perfectly reliable (i.e., an alpha 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, any variation in test scores comes solely from random error.

Many traditional measures of reliability exist; their differences are due to assumptions each makes about the nature of the data from a test. For instance, the Spearman–Brown reliability formula assumes items are parallel, having equal amounts of information about the trait and 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

[32] The term "results" is typically used in place of "scores" to highlight the fact that DLM assessment results are not based on scale scores. For ease of reading, the term "score" is used in this chapter.

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 the assumptions of the test are met.

As the DLM Alternate Assessment System uses a different type of psychometric approach than is commonly found in contemporary testing programs, the reliability evidence reported may, at first, look different from that reported when test scores are produced using traditional psychometric techniques such as classical test theory or item response theory. Consistent with traditional reliability approaches, however, is the meaning of all indices reported for DLM assessments: 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.

As the name suggests, diagnostic classification models are models that produce classifications as probability estimates for student test takers. For the DLM system, the classification estimates are based on the set of areas and levels within areas in which each student was tested. In DLM terms, each area is called an Essential Element (EE). Each EE is divided into five linkage levels of complexity: Initial Precursor (IP), Distal Precursor (DP), Proximal Precursor (PP), Target (T), and Successor (S).

An example of an EE with sets of nodes labeled as linkage levels is given in Figure 54. The EE in the example is from third-grade mathematics and is labeled M.EE.3.MD.4 “Measure length of objects using standard tools, such as rulers, yardsticks, and meter sticks.” See Chapter III for more detail on the development of the linkage levels and how they relate to the DLM design.

In Figure 54, each node is shown as a red box. In the top-right corner of the box, a letter code is given, indicating the linkage level of the node (IP, DP, PP, T, S, and UN: Untested Node, a node not currently tested as part of DLM testlets).

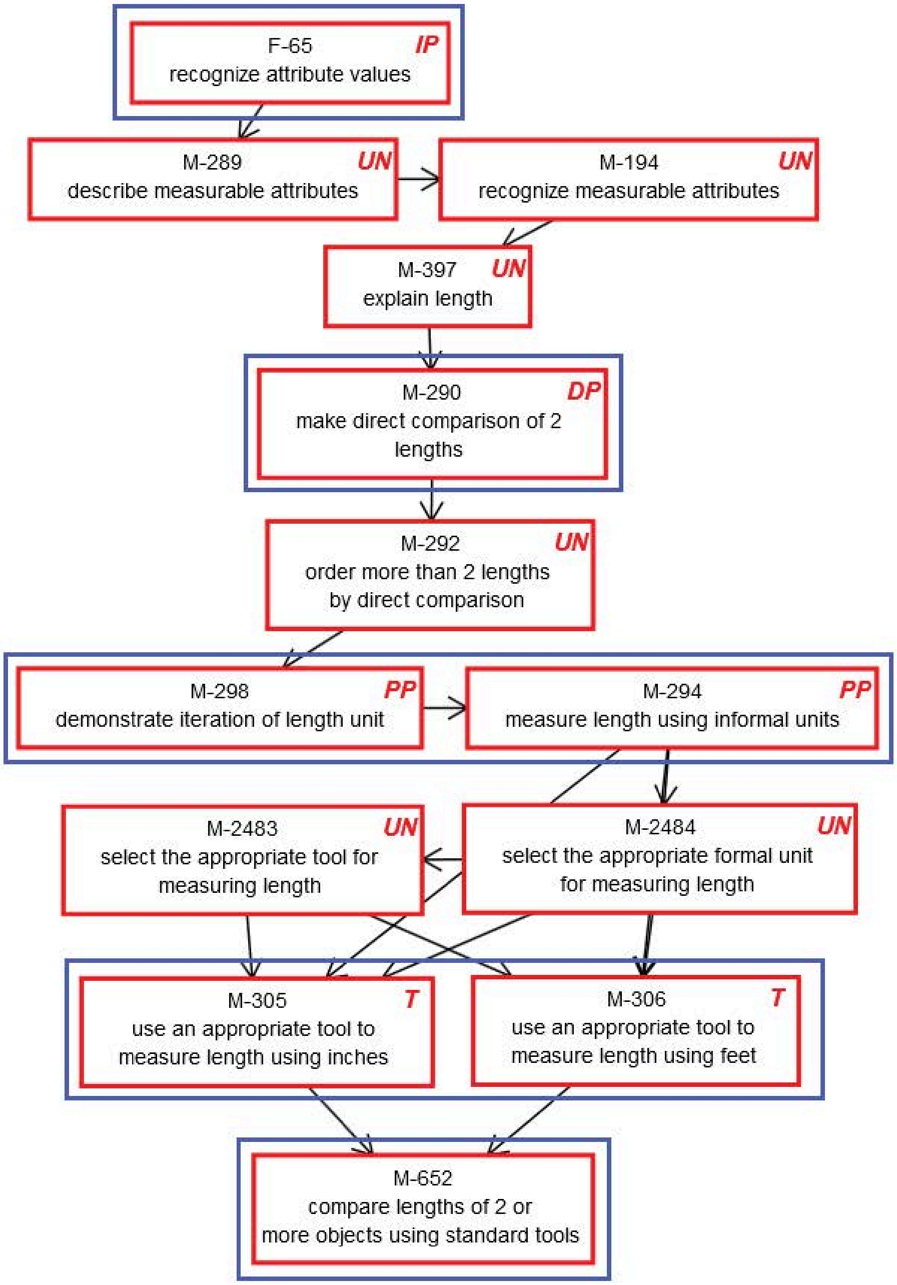

F-65 IP
recognize attribute values
M-289 UN
describe measurable attributes
M-194 UN
recognize measurable attributes
M-397 UN
explain length
M-290 DP
make direct comparison of 2 lengths
M-292 UN
order more than 2 lengths by direct comparison
M-298 PP
demonstrate iteration of length unit
M-294 PP
measure length using informal units
M-2483 UN
select the appropriate tool for measuring length
M-2484 UN
select the appropriate formal unit for measuring length
M-305 T
use an appropriate tool to measure length using inches
M-306 T
use an appropriate tool to measure length using feet
M-652
compare lengths of 2 or more objects using standard tools

Figure 54. Mini-map of nodes for EE M.EE.3.MD.4 (third-grade mathematics).

For each linkage level embedded within each EE, DLM testlets were written with items measuring the listed linkage-level nodes. Because of the DLM administration design, students did not take testlets outside of a single linkage level within an EE. Students typically saw a single testlet within a given EE; consequently, data obtained when students responded to testlets at adjacent linkage levels within an EE are sparse. Because direct evidence of connections between nodes at different linkage levels was not often collected, DLM node parameters could not be estimated. Instead, a linkage-level model was used to estimate examinee proficiency.

The diagnostic classification models used in psychometric analyses of student test data produced student-level classifications for each linkage level for which a student was tested. Because students often did not test at more than one linkage level within an EE, students who did not meet mastery status for any tested linkage level were assigned mastery status for the linkage level two levels below the highest level in which they were tested (unless the highest level tested was either the IP or DP levels, in which case students were considered nonmasters of all linkage levels within the EE). The classification results were then aggregated across all linkage levels to produce a count of the total number of linkage levels mastered across both content areas (English language arts (ELA) and mathematics); this total was then compared with cut points that resulted in a final proficiency-level classification.

Reliability evidence is provided at three levels of testing: (a) the number of linkage levels mastered within a content area (labeled content-area reliability; provided for ELA and mathematics); (b) the number of linkage levels mastered within each EE (labeled EE reliability); and (c) the classification accuracy of each linkage level within each EE (labeled linkage-level reliability). As described in the next section, reliability evidence comes from simulation studies in which model-specific test data are generated for students with known levels of attribute mastery. For content-area reliability and EE reliability, a correlation between the true and estimated numbers of linkage levels mastered is reported, along with a summary of the number of linkage levels correctly classified. The correlation estimate mirrors estimates of reliability from contemporary measures such as the Guttman-Cronbach Alpha. For linkage-level reliability, the reliability evidence is in the form of correct classification rates (raw and chance corrected) and a tetrachoric correlation between true and estimated linkage-level mastery statuses. The tetrachoric correlation—a correlation that ranges from –1 to 1—is provided as the classification status. This correlation is a discrete categorization whereby the traditional correlation (the Pearson correlation coefficient) would likely lead to results where the correlation's lower and upper bounds were different from –1 and 1.

With the classification methods of diagnostic classification models based on discrete statuses of an examinee, reliability-estimation methods based on item response theory estimates of ability 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 (e.g., Henson & Douglas, 2005; Templin & Bradshaw, 2013).

VIII.1.A. Methods of Obtaining Reliability Evidence

Standard 2.1: "The range of replications over which reliability/precision is being evaluated should be clearly stated, along with a rationale for the choice of this definition, given the testing situation" (AERA et al., 2014, p. 42).

As the DLM psychometric model yields complex results for multiple sources of information (content area, EE, and linkage levels), reliability methods were based on simulation. Simulation has a long history of use in deriving reliability evidence; large testing programs such as the Graduate Record Examination report reliability results based on simulation (e.g., Educational Testing Service, 2016). With respect to diagnostic classification models, simulation-based reliability has been used in a number of studies (e.g., Bradshaw, Izsák, Templin, & Jacobson, 2014; Templin & Bradshaw, 2013). For any simulation-based reliability method, the approach is to generate simulated examinees with known characteristics, simulate test data using calibrated-model parameters, score the test data using calibrated-model parameters, and finally compare estimated examinee characteristics with those characteristics known to be true in the simulation. For DLM, the known characteristics of the simulated examinees are the set of linkage levels the examinee has mastered and not mastered.

The use of simulation is necessitated by two factors: the assessment blueprint and the classification-based results that such administrations give. The method provides 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. In the end, reliability simulation replicates DLM versions of scores from actual examinees based upon the actual set of items each examinee has taken. Therefore, this simulation provides a replication of the administered items for the examinees.

The simulation used to estimate reliabilities for DLM versions of scores and classifications takes into consideration the unique design and administration of DLM assessments in the initial operational year. Specifically, students tested only at the end of the academic year, with minimal items taken per EE as specified by the blueprint.

VIII.1.A.i. Reliability Sampling Procedure

The simulation design for the reliability estimates developed a resampling design to mirror the trends existing in the DLM assessment data. In accordance with Standard 2.1, the sampling design used the entire set of operational testing data to generate simulated examinees. Using this process guarantees that the simulation takes on characteristics of the DLM operational test data that are likely to affect the reliability results. For one simulated examinee, the process was as follows:

1. Draw with replacement the student record of one student from the spring 2015 operational testing data. Use the student's originally scored pattern of linkage-level mastery and nonmastery as the true values for the simulated student data.
2. Simulate a new set of item responses to the set of items administered to the student in the operational testlet. Item responses are simulated from calibrated-model parameters for the items of the testlet, conditional on the profile of linkage-level mastery or nonmastery for the student.
3. Score the simulated-item parameters using the operational DLM scoring procedure (described in Chapter V), producing estimates of linkage-level mastery or nonmastery for the simulated student.
4. Compare the estimated linkage-level mastery or nonmastery to the known values from step 2 for all linkage levels for which the student was administered items. Note that the comparison does not include the additional assumption of mastery of linkage levels at least two levels below the highest level tested (as this assumption is not testable).
5. Repeat steps 1 through 4 for 2,000,000 simulated students.

Figure 55 shows the simulation process as a flow chart.

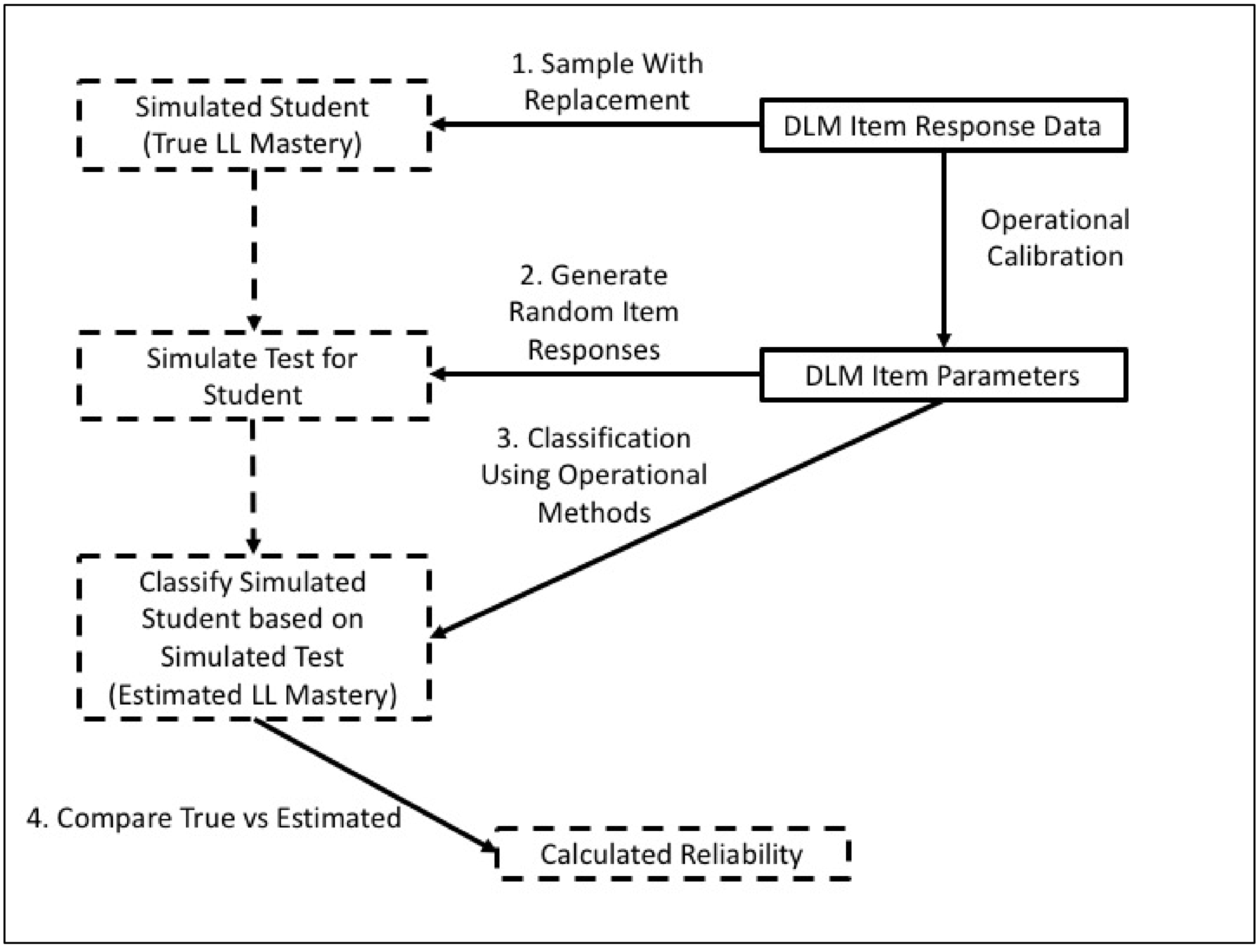

1. Sample With Replacement
Simulated Student (True LL Mastery)
DLM Item Response Data
Operational Calibration
2. Generate Random Item Responses
Simulate Test for Student
DLM Item Parameters
3. Classification Using Operational Methods
Classify Simulated Student based on Simulated Test (Estimated LL Mastery)
4. Compare True vs Estimated
Calculated Reliability

Figure 55. *Simulation process for creating reliability evidence. LL = linkage level.*

VIII.1.B. RELIABILITY EVIDENCE

Standard 2.2: "The evidence provided for the reliability/precision of the scores should be consistent with the domain of replications associated with the testing procedures, and with the intended interpretations for use of the test scores" (AERA et al., 2014, p. 42).

Standard 2.5: "Reliability estimation procedures should be consistent with the structure of the test" (AERA et al., 2014, p. 43).

Standard 2.12: "If a test is proposed for use in several grades or over a range of ages, and if separate norms are provided for each grade or each age range, reliability/precision data should be provided for each age or grade-level subgroup, not just for all grades or ages combined" (AERA et al., 2014, p. 45).

Standard 2.16: "When a test or combination of measures is used to make classification decisions, estimates should be provided of the percentage of test takers who would be classified in the same way on two [or more] replications of the procedure" (AERA et al., 2014, p. 46).

Standard 2.19: "Each method of quantifying the reliability/precision of scores should be described clearly and expressed in terms of statistics appropriate to the method" (AERA et al., 2014, p. 47).

Reliability evidence is given for three levels of data, each important in the DLM testing design: (a) content-area reliability, (b) EE reliability, and (c) linkage-level reliability. With 255 EEs, each with five linkage levels, a total of 1,275 analyses were conducted to summarize reliability. Therefore, the reported evidence will be summarized in this chapter. Full reporting of reliability evidence for all 1,275 linkage levels and 255 EEs is provided in an online appendix (http://dynamiclearningmaps.org/reliabevid). The full set of evidence is provided in accordance with Standard 2.12.

Reporting reliability at three levels ensures that the simulation and resulting reliability evidence were conducted in accordance with Standard 2.2. Additionally, providing reliability evidence for each of the three levels of data ensures that these reliability-estimation procedures meet Standard 2.5.

VIII.1.B.i. Content-Area (Performance-Level) Reliability Evidence

Content-area reliability provides evidence for reliability for the number of linkage levels mastered across all EEs for a given content area (ELA or mathematics) and grade level. As students are assessed on multiple linkage levels within a content area, content-area reliability evidence is similar to reliability evidence for testing programs that use summative tests to describe content-area performance. That is, the number of linkage levels mastered within a content area can be thought of as being analogous to the number of items answered correctly in a different type of testing program.

Table 74 shows this information across all grades and content areas. The content-area reliability evidence takes the true and estimated number of linkage levels mastered across all tested levels for a given content area. Reliability is reported with two summary numbers: the Pearson correlation between the true and estimated number of linkage levels mastered within a content area, and the correct classification rate for which linkage levels were mastered as averaged across all simulated students. Classification-rate information is provided in accordance with Standard 2.16. The two summary statistics included in Table 74 also meet Standard 2.19.

Table 74. Linkage Levels Mastered Correlation and Average Student Correct Classification by Content Area

Grade	Content Area	Linkage Levels Mastered Correlation	Average Student Correct Classification
3	M	.863	.854
3	ELA	.921	.847
4	M	.869	.849
4	ELA	.935	.856
5	M	.866	.849
5	ELA	.950	.873
6	M	.861	.856
6	ELA	.942	.871
7	M	.869	.857
7	ELA	.948	.860
8	M	.837	.846
8	ELA	.920	.825
High School	M	.861	.897
High School	ELA	.925	.840

Note: M = Mathematics. ELA = English language arts.

From the table, it is evident that content-area reliability, as demonstrated by the correlation between true and estimated number of linkage levels mastered, ranges from .837 to .950. As such, the DLM scoring procedure of reporting the number of linkage levels mastered has adequate reliability.

VIII.1.B.ii. Essential-Element Reliability Evidence

Moving from content areas to Essential Elements (EE), the reliability evidence shifts slightly. That is, because EEs are collections of linkage levels with an implied order, the number of linkage levels mastered per EE, rather than for each content area is examined. If one considers content-area scores as total scores from an entire test, then EEs are the more fine-grained strands within the content area itself.

The following three statistics are used to summarize reliability evidence at the linkage level:

1. The correct classification rate for the number of linkage levels mastered within an EE.
2. The correct classification kappa for the number of linkage levels mastered within an EE (i.e., a chance-corrected classification rate labeled kappa that represents the proportion of error reduced above chance). Values of kappa above .6 indicate substantial-to-perfect agreement between simulated and estimated numbers of linkage levels mastered within an EE (Landis & Koch, 1977).
3. The Pearson correlation between true and estimated numbers of linkage levels mastered within an EE.

As there are 255 EEs, the summaries reported herein are based on the proportion of EEs that fall within a given range of an index value. Results are given in both tabular and graphical form. Table 75 and Figure 56 provide proportions of EEs falling within prespecified ranges of values for the three reliability summary statistics (i.e., correct classification rate, kappa, and correlation). Kappa values for 13 EEs could not be computed due to diagonal cells with 0s, that is, whenever the number of linkage levels mastered was the same for all students. Proportions in Table 75 and Figure 56 are based on kappa values that could be calculated.

Table 75. Reliability Summaries Across All EEs: Proportion of EEs Falling Within a Specified Index Range

Reliability Index	**Index Range**								
	< .60	**.60–.64**	**.65–.69**	**.70–.74**	**.75–.79**	**.80–.84**	**.85–.89**	**.90–.94**	**.95–1.0**
Correct Classification Rate	0	.004	.017	.029	.103	.298	.335	.186	.029
Kappa	.318	.128	.186	.169	.095	.070	.021	.004	.008
Correlation	.219	.128	.202	.169	.140	.091	.021	.021	.008

Note: Kappa proportions are based on kappa values that could be calculated.

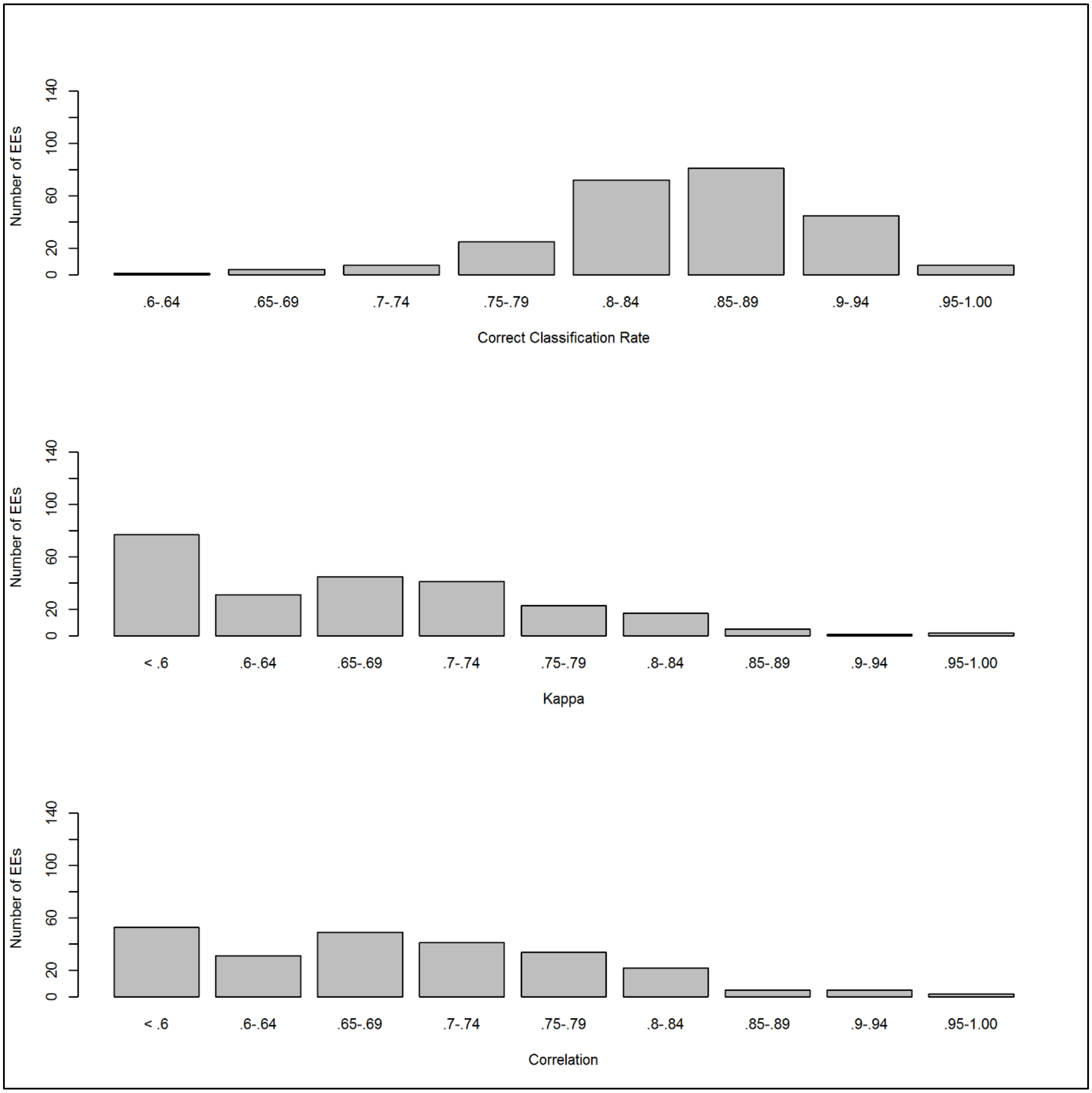


Figure 56. Number of linkage levels mastered within EE reliability summaries. Kappa proportions are based on kappa values that could be calculated.

In general, the reliability summaries for number of linkage levels mastered within EEs show good levels of reliability. However, roughly one-third of kappa values fell below .6, which may be due to students taking a small number of items per EE.

VIII.1.B.iii. Linkage-Level Reliability Evidence

Evidence at the linkage level comes from the comparison of true and estimated mastery statuses for each of the 1,275 linkage levels in the operational DLM assessment[33]. As an example, Table 76 shows a simulated table from the PP linkage level of the previously shown EE, M.EE.MD.3.4 (see Figure 54).

Table 76. True and Estimated Mastery Status from Reliability Simulation for Proximal-Precursor Linkage Level of EE M.EE.MD.3.4

	Estimated Mastery Status	
	Nonmaster	**Master**
True Mastery Status	574	235
	83	592

The summary statistics reported are all based on tables like this one: the comparison of true and estimated mastery statuses across all simulated examinees. As with any contingency table, a number of summary statistics are possible.

For each statistic, figures are given comparing the results of all 1,275 linkage levels. Three summary statistics are presented:

1. The correct classification rate (i.e., the proportion of simulated examinees where estimated mastery status was the same as true mastery status).
2. The correct classification kappa (i.e., a chance-corrected classification rate that represents the proportion of error reduced above chance). Values of kappa above .6 indicates substantial to perfect agreement between simulated and estimated linkage-level mastery status (Landis & Koch, 1977).
3. The tetrachoric correlation between estimated and true classification status (i.e., ranges between 0 and 1; interpreted like a Pearson correlation).

As there are 1,275 total linkage levels across all 255 EEs, the summaries reported herein are based on the proportion of linkage levels that fall within a given range of an index value. Results are given in both tabular and graphical form. Table 77 and Figure 57 provide proportions of linkage levels falling within prespecified ranges of values for the three reliability

[33] The linkage level reliability evidence presented here focuses on consistency of measurement given student responses to items. For more information on how students were assigned an initial linkage level during assessment, see Chapter 3 – Pilot Administration: Initialization and Chapter 4 – Adaptive Delivery.

summary statistics (i.e., correct classification rate, kappa, and correlation). Similar to the EE results, kappa values for 154 linkage levels could not be computed whenever all students were labeled as masters of the linkage level. Proportions in Table 77 and Figure 57 are based on kappa values that were calculated. Moreover, some tetrachoric correlation values could not be computed due to dependencies in the true- and expected-data contingency tables.

The reliability summaries for classification accuracy for linkage levels show fairly good levels of reliability.

Table 77. Reliability Summaries Across All Linkage Levels: Proportion of Linkage Levels Falling a Within Specified Index Range

Reliability Index	**Index Range**								
	< .60	**.60–.64**	**.65–.69**	**.70–.74**	**.75–.79**	**.80–.84**	**.85–.89**	**.90–.94**	**.95–1.0**
Correct Classification Rate	.034	.015	.025	.045	.123	.189	.222	.180	.166
Kappa	.472	.118	.104	.076	.072	.051	.035	.019	.054
Correlation	.108	.024	.045	.058	.077	.121	.139	.155	.272

Note: Kappa and tetrachoric correlation proportions are based on values that were able to be computed.

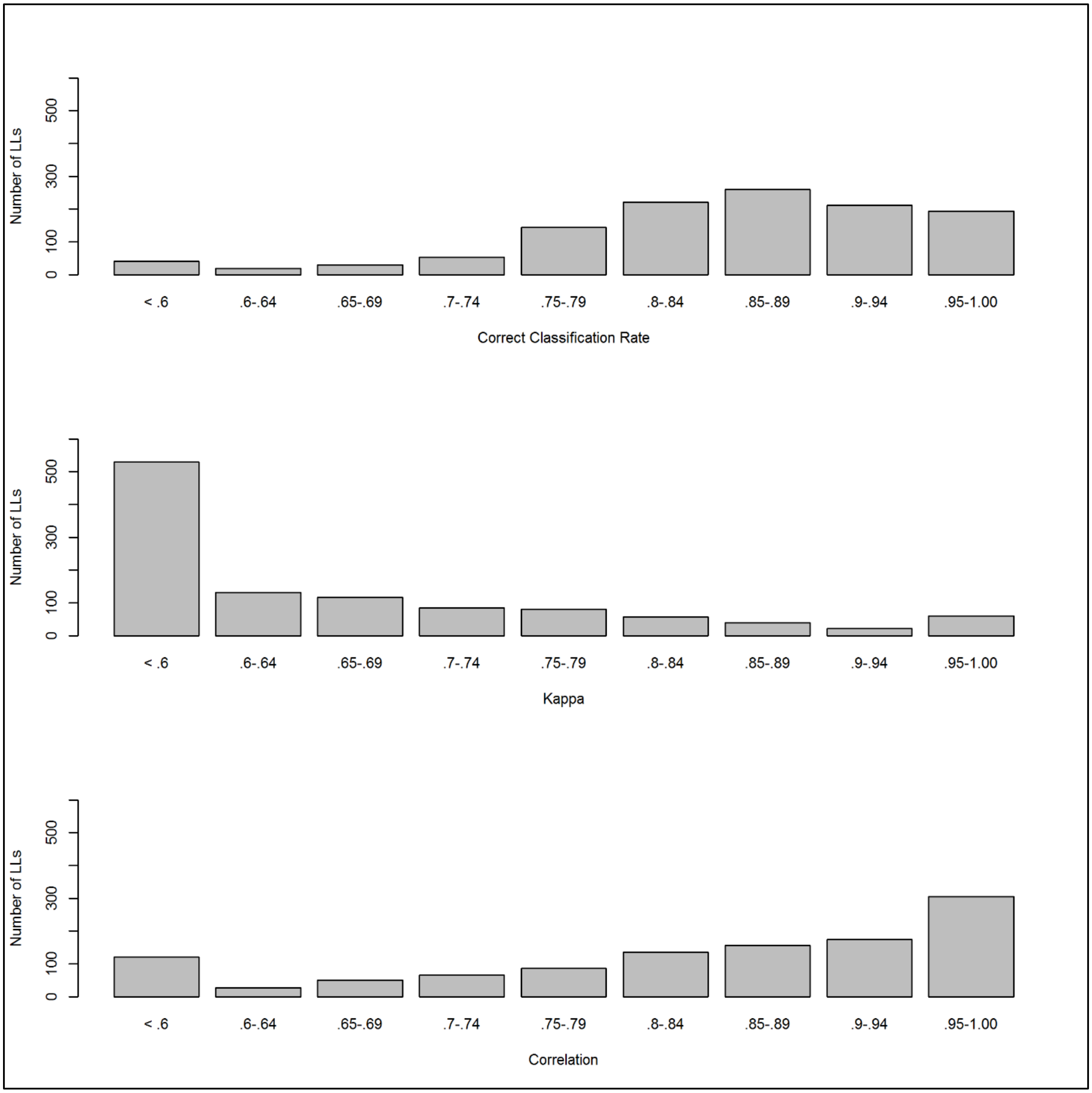


Figure 57. Linkage-level reliability summaries. Kappa and tetrachoric correlation proportions are based on values that were able to be computed.

In summary, reliability measures for the DLM assessment system addressed the standards set forth by AERA et al., 2014. The methods used were consistent with assumptions of DCM and yielded evidence to support the argument for internal consistency of the program.

IX. VALIDITY STUDIES

The preceding chapters provide evidence in support of the overall validity argument for scores produced by the Dynamic Learning Maps® (DLM®) Alternate Assessment System. Chapter IX presents additional evidence. The special studies presented here were conducted throughout the assessment development, administration, and evaluation processes. These studies address four of the critical sources of evidence as described in *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014): evidence based on (a) test content, (b) response processes, (c) internal structure, and (d) consequences of testing. Each study addresses assumptions related to the theory of action, specifically, related to the four propositions for score interpretation and use. These propositions and score purposes are discussed in depth in the Evaluation Summary section of Chapter XI where the DLM Alternate Assessment System's overall validity framework is laid out alongside evidence sources.

IX.1. Evidence Based on Test Content

Evidence based on test content relates to the evidence "obtained from an analysis of the relationship between the content of the test and the construct it is intended to measure" (AERA et al., 2014, p. 14). The interpretation and use of DLM scores depends on the validation of the model of learning and cognition underlying the system and its alignments to and applications through content standards, items, and full tests. The validity studies presented in this section focus on the alignment of test content to content standards via the DLM maps (which underlie the assessment system) and preliminary evidence of student opportunity to learn the assessed content.

IX.1.A. External Alignment Study

ACERI Partners conducted an external alignment study on the 2014–2015 DLM operational assessment system (Flowers, Wakeman, McCord, & Taub, 2016). The purpose of the study was to investigate the relationships between the content structures in the DLM Alternate Assessment System and assessment items. A modification of Links for Academic Learning alignment methodology (Flowers, Wakeman, Browder, & Karvonen, 2009) was used to evaluate the coherence of the DLM Alternate Assessment System. The alignment study focused on the following relationships (as illustrated by the corresponding numbers in Figure 58 below):

1. College and Career Ready (CCR) Standards and Essential Elements (EEs)
2. an EE and its Target level node(s)
3. the vertical articulation of the linkage levels associated with an EE
4. DLM map nodes within a linkage level and assessment items

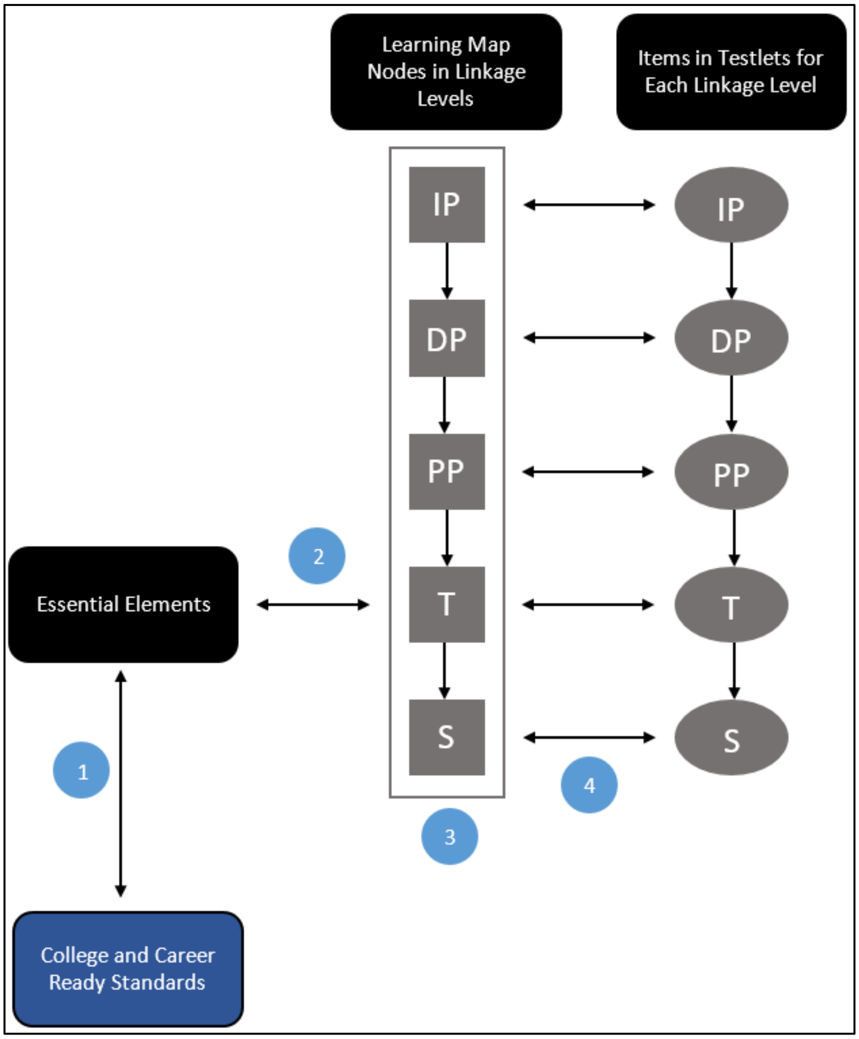

Learning Map Nodes in Linkage Levels
Items in Testlets for Each Linkage Level
IP
DP
PP
T
S
IP
DP
PP
T
S
Essential Elements
College and Career Ready Standards
1
2
3
4

Figure 58. Relationships investigated in the external alignment study.

A sample of English language arts (ELA) and mathematics 2014–2015 operational testlets from grade 3 through high school were examined. In ELA, a total of 175 testlets and 910 items were examined for alignment. In mathematics, 180 testlets and 835 items were evaluated. Items and testlets were sampled from the spring pool. Confidence intervals (90% CI) were calculated and the lower limit and upper limit for each interval were reported for sampled pools.

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 target (CCR, EE, Target level node, and linkage levels) and the linked target (EE, Target level node, linkage level, and 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*. Where panelists identified a relationship that did not meet criteria for alignment (e.g., *no link* for content centrality) additional feedback was provided. When evaluating items, panelists also identified the category for the highest cognitive process dimension required of the student when responding to the item, using the DLM cognitive process dimension taxonomy.

The following sections provide a brief summary of findings from the external alignment study. Full results are provided in the separate technical report (Flowers et al., 2016).

IX.1.A.i. Alignment of College and Career Ready Standards and Essential Elements

All EEs identified in the test blueprints were included in these analyses. The results of content centrality and performance centrality ratings are shown in Table 78 and Table 79. Across all content areas and testlet pools, 81% to 93% of the EEs were rated as maintaining fidelity to the content in the grade-level CCR standards, as indicated by the bold font. 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.

Table 78. Content Centrality of CCR Standards to Essential Elements

Pool	**EE**	**No**		**Far**		**Near**		**Met**		**CI (90%)**	
Content Centrality	***N***	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	**LL (%)**	**UL (%)**
ELA – YE	136	25	18%	93	68%	18	13%	111	82%	75%	87%
ELA – EOI	38	6	16%	23	61%	9	24%	32	85%	71%	93%
Math – YE	145	28	19%	106	73%	11	8%	117	81%	74%	86%
Math – EOI	41	3	7%	33	80%	5	12%	38	93%	82%	98%

Note: YE = Year End. EOI = End of Instruction. CI = confidence interval. LL = lower limit. UL = upper limit. Bold indicates acceptable level of alignment.

Table 79. Performance Centrality of CCR Standards to Essential Elements

	EE	None		Some		All		Met[1]		CI: 90%	
Performance Centrality	***N***	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	**LL**	**UL**
ELA – YE	111	0	0%	95	86%	16	14%	111	100%	97%	100%
ELA – EOI	34	6	18%	22	65%	6	18%	28	82%	68%	92%
Math – YE	117	20	17%	77	66%	20	17%	97	83%	76%	88%
Math – EOI	38	3	8%	28	74%	7	18%	35	92%	81%	98%

Note: YE = Year End. EOI = End of Instruction. CI = confidence interval. LL = lower limit. UL = upper limit. Bold indicates acceptable level of alignment.

At the grade level, content centrality for ELA grades 4 and 5 was slightly below the 80% threshold for acceptable alignment and math grades 3, 5, 6, and 8 fell below the 80% threshold. The most common reason for ratings of no content centrality was that panelists believed the EE was a mismatch to the skill in the identified CCR standard (17 of 25 EEs). For performance centrality, all of the ELA EEs retained some or all of the performance expected in the CCR standard. In math, the 80% threshold was met in all grades except 4, 6, and 7. All of the end-of-instruction subject areas (e.g., Algebra 1, English I), met the 80% criteria for both content centrality and performance centrality.

IX.1.A.ii. Alignment of Essential Element and Target Level Node(s)

Statistics for content and performance centrality on the alignment of EEs to Target level node(s) are displayed in Table 80 and Table 81. The number of EEs in Table 80 and Table 81 is different from Table 78 and Table 79 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. Together the *far* and *near* findings suggest a strong alignment between EEs and Target level nodes.

Table 80. Content Centrality of EEs to Target Level Node(s)

	EE	No		Far		Near		Met		CI (90%)	
Content Centrality	***N***	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	**LL**	**UL**
ELA – YE	148	0	0%	11	7%	137	93%	148	100 %	98%	100%

	EE	**No**		**Far**		**Near**		**Met**		**CI (90%)**	
Content Centrality	***N***	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	**LL**	**UL**
ELA – EOI	38	0	0%	0	0%	38	100%	38	100%	92%	100%
Math – YE	219	0	0%	54	25%	165	75%	219	100%	99%	100%
Math – EOI	49	0	0%	21	43%	28	57%	49	100%	94%	100%

Note: YE = Year End. EOI = End of Instruction. CI = confidence interval. LL = lower limit. UL = upper limit. Bold indicates acceptable level of alignment.

Table 81. Performance Centrality of EEs to Target Level Node(s)

	EE	**None**		**Some**		**All**		**Met**		**CI: 90%**	
Performance Centrality	***N***	***N***	**%**	***N***	**%**	***N***	**%**	***N***	**%**	**LL**	**UL**
ELA – YE	148	0	0%	20	14%	128	86%	148	100%	98%	100%
ELA – EOI	38	0	0%	5	13%	33	87%	38	100%	92%	100%
Math – YE	219	0	0%	50	23%	169	77%	219	100%	99%	100%
Math – EOI	49	0	0%	8	16%	41	84%	49	100%	94%	100%

Note: YE = Year End. EOI = End of Instruction. CI = confidence interval. LL = lower limit. UL = upper limit. Bold indicates acceptable level of alignment.

IX.1.A.iii. Vertical Articulation of Linkage Levels for each Essential Element

Panelists evaluated linkage levels (small, ordered collections of nodes associated with each EE) to see if they reflected a progression of knowledge, skills, and understandings. Results of the vertical articulation of the linkage levels for the EEs at each grade level are reported in Table 82 and Table 83. For ELA, a total of 147 linkage levels were reviewed by panelists, and 120 (82%) were rated as showing a clear progression from Initial Precursor to Successor linkage levels. The low rating for seventh grade was due to panelists reporting that the Initial Precursor linkage level was not clearly part of the progression in the ordered nodes. For math, 103 linkage levels were reviewed, and 99 linkage levels (96%) were rated as demonstrating a clear progression in the ordered nodes. Based on the comments provided by ELA panelists, in most cases of weak progression, they judged the Initial Precursor level to be inappropriate content for the progression. Based on observation of panel discussions during part of this rating process, it is possible that they perceived the distance between nodes selected for assessment at the Initial Precursor and Distal Precursor linkage levels to be too far.

Table 82. Vertical Articulation of Linkage Levels for Essential Elements in ELA

	Total	**Clear Progression**	
Grade	***N***	***n***	**%**
3	17	15	88
4	17	14	82
5	19	15	79
6	19	15	79
7	18	10	56
8	20	17	85
9-10	19	17	89
11-12	18	17	94
All	147	120	82

Table 83. Vertical Articulation of Linkage Levels for Essential Elements in Math

	Total	**Clear Progression**	
Grade	***N***	***n***	**%**
3	11	10	91
4	16	15	94
5	15	14	93
6	11	10	91
7	10	10	100
8	14	14	100
9	8	8	100
10	9	9	100
11	9	9	100
All	103	99	96

IX.1.A.iv. Alignment of Learning Map Nodes within a Linkage Level and Assessment Items

Content and performance centrality ratings for the nodes corresponding to the assessment items are reported in Table 84 and Table 85. Almost all items were rated as having *far* or *near* content centrality to the corresponding node, ranging from 97% to 100%. Similarly, the performance centrality ratings indicated that almost all items maintained the performance expectations found in the corresponding linkage level node.

Table 84. Content Centrality of Linkage Level Nodes to Assessment Items

Pool	EE	No		Far		Near		Met[a]		CI	
	N	*N*	%	*N*	%	*N*	%	*N*	%	*LL*	*UL*
ELA – YE	669	21	3%	34	5%	614	92%	648	97%	96%	98%
ELA – EOI	241	8	3%	15	6%	218	90%	233	97%	94%	98%
Math – YE	622	2	<.5%	11	2%	609	98%	620	100%	99%	100%
Math – EOI	213			2	<.5%	211	100%	213	100%	99%	100%

Note: YE = Year End. EOI = End of Instruction. CI = confidence interval. LL = lower limit. UL = upper limit. [a]Met is the total number of items and percentage rated as far or near.

Table 85. Performance Centrality of Linkage Level Nodes to Assessment Items

Pool	EE	No		Some		All		Met[a]		CI	
	N	*N*	%	*N*	%	*N*	%	*N*	%	*LL*	*UL*
ELA – YE	669	16	2%	66	10%	586	88%	652	97%	96%	98%
ELA – EOI	241	6	2%	25	10%	210	87%	235	97%	95%	99%
Math – YE	622	2	<.5%	14	2%	606	97%	620	100%	99%	100%
Math – EOI	213			1	<.5%	212	100%	213	100%	99%	100%

Note: YE = Year End. EOI = End of Instruction.CI = confidence interval. LL = lower limit. UL = upper limit. [a]Met is the total number of items and percentage rated some or all.

The percentages of DLM cognitive process dimension for ELA and math items are reported in Table 86. Most ELA items were rated at the Respond through Understand levels, while most math items received ratings from the Remember through the Analyze cognitive process dimension levels. Most items were located in the middle of the cognitive process dimension distribution. These results suggest that the items cover a wide range of cognitive complexity and provide opportunities for student with the most significant cognitive disabilities to demonstrate knowledge of appropriately challenging content.

Table 86. Cognitive Process Dimension for ELA and Math Items

	ELA-YE (*N*=669)	ELA-EOI (*N*=241)	Math-YE (*N*=622)	Math-EOI (*N*=213)
Pre-Intentional	0%	0%	0%	0%
Attend	0%	0%	0%	0%
Respond	40%	31%	0%	0%
Replicate	0%	0%	0%	0%
Remember	14%	9%	22%	12%
Understand	45%	56%	42%	48%
Apply	0%	4%	21%	23%
Analyze	0%	0%	13%	14%
Evaluate	0%	0%	2%	4%
Create	0%	0%	0%	0%

Note: YE = Year End. EOI = End of Instruction.

Panelist ratings were compared against the categories identified by DLM item writers. With nine categories in the taxonomy that are potentially appropriate for items, exact and adjacent agreements were calculated. Exact agreement ranged from 70% to 90% of items and adjacent agreement from 76% to 94% of items.

Overall, the external alignment study provides evidence of the DLM Alternate Assessment System components that connect the Common Core State Standards to the assessment items, via EEs and nodes in linkage levels. The external alignment study provides substantial content-related evidence to support the DLM Consortium's claims about what students know and can do in ELA and math. Areas for further investigation and action based on the findings are addressed in Chapter XI.

IX.1.B. Opportunity to Learn

After completing administration of the spring 2015 operational assessments, test administrators were invited to complete a survey about the assessment administration process. All educators who had administered a DLM assessment during the spring 2015 window (*N* = 14,145) were invited to respond to the survey. State partners announced the availability of the survey and encouraged test administrators' participation. A total of 1,792 test administrators responded, yielding an overall response rate of 12.7%.

The survey served several purposes.[34] Two items provided very preliminary information about the relationship between the learning opportunities that students had prior to testing and the test content (testlets) that they encountered on the assessment. The surveys asked test administrators to indicate whether they judged that the test content, across all testlets, was aligned with their instruction. Table 10 reports the results. Overall, the frequency distribution ranged from no testlets matching instruction to all seven testlets matching in both math and ELA. The results underscore the need for improvement of the match between tested content and instructed content. More specific measures of instructional alignment are planned.

Table 87. Number of Testlets that Matched Instruction

	ELA		**Math**	
Number of Testlets	***n***	**%**	***n***	**%**
0	195	7.6	170	6.9
1	364	14.3	481	19.4
2	426	16.7	487	19.6
3	433	17.0	445	17.9
4	392	15.4	353	14.2
5	328	12.9	254	10.2
6	217	8.5	181	7.3
7	196	7.7	109	4.4

IX.2. Evidence Based on Response Processes

The study of the response processes of test takers provides evidence regarding the fit between the test construct and the nature of how students actually experience test content (AERA, APA, & NCME, 2014). Both theoretical and empirical evidence is appropriate and should come from both the individual test taker and external observation. The interpretation and use of DLM scores depends in part on the validation of whether the cognitive processes that students are engaged in when taking the test match the claims made about the test construct. This category of evidence includes studies on student and test administrator behaviors during testlet administration. Because testlets must be administered with fidelity in order to support the ability of students to respond based on their knowledge of the construct, evidence of fidelity is

[34] Results for other items are reported in Chapter IV and later in this chapter.

also included in this section. Finally, test administrator feedback on students' abilities to respond to testlets during the spring 2015 administration is provided in this section.

IX.2.A. Response Process Studies

Two cognitive lab studies were conducted to better understand response processes. The first focused on students' experience engaging with test content for various item types in computer-administered testlets. The second focused on test administrators' interpretations of student behavior and responses to questions about their students' responses in teacher-administered testlets.

IX.2.A.i. Student Cognitive Labs

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 (3-8 and HS) 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 four types of technology-enhanced items, including drag-and-drop, click-to-place, select-text, and multi-select multiple-choice item types. The first three item types were designed specifically for DLM assessments and are delivered through a user interface designed for this population. The drag-and-drop and click-to-place item types are used for sorting. The difference between them is that the drag-and-drop format requires continuous selection (clicking and dragging) while click-to-place items require clicking on the origin and then clicking on the intended destination. The latter item type is accessible for switch users, but one theory was that non-switch users would also find clicking without dragging to be easier since the process was less demanding on fine motor skills. Both the drag-and-drop and click-to-place items were built to require a similar response process: sorting objects into categories. To facilitate comparisons with drag-and-drop and click-to-place items, multi-select multiple-choice items were also constructed to access a response process requiring the student to select the answer options that matched a category. 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.

To avoid relying on items that might be too difficult and therefore inappropriate for use in cognitive labs (Johnstone, Altman, & Moore, 2011), the labs used four-item testlets with content that did not rely on prior academic knowledge. For example, while students who might be candidates for cognitive labs are highly likely to know their shapes, completing an item with shapes did not require an understanding of specific shapes (see Figure 59). Figure 60 shows a select-text item that was similarly constructed to minimize the need for prior knowledge.

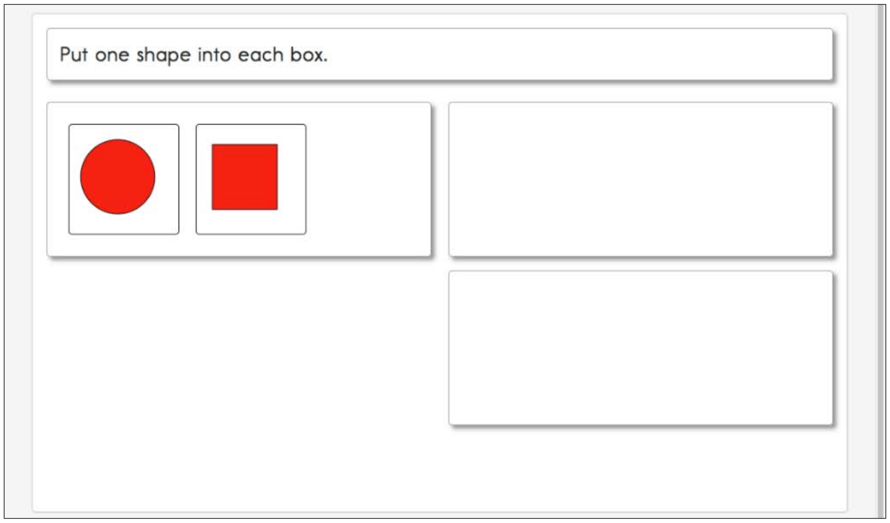
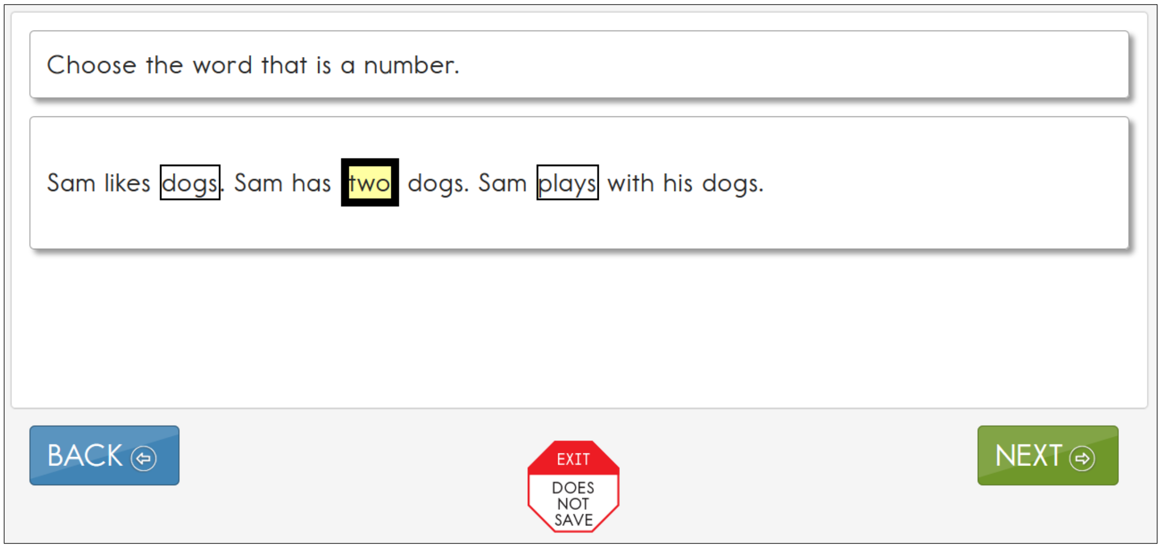

Put one shape into each box.

Figure 59. Sample drag-and-drop item.



Choose the word that is a number.

Sam likes dogs. Sam has two dogs. Sam plays with his dogs.

Figure 60. Sample select-text item.

Each testlet contained one type of item. For select-text and drag-and-drop item types, 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 testlet assignments were counterbalanced. Fifteen students completed drag-and-drop testlets, eleven completed click-to-place testlets, eight completed select-text testlets, and eleven completed multi-select multiple-choice 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 (e.g., for drag-and-drop items, the process includes initial item selection, manipulation, and item placement) 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. 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. (2011; 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 here 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 drag-and-drop and click-to-place item types were demonstrated when students had difficulty selecting the desired object, difficulty maintaining continuous selection, difficulty with group selection, or difficulty with number of objects. In general, students tended to have more difficulty with click-to-place items than drag-and-drop items, and more frequently needed assistance to complete them (see Table 88).

Table 88. Sources of Challenge in Response to Drag-and-Drop and Click-to-Place Item Types

Source of Challenge	**Drag and Drop** **(*N* = 15 students, 60 items)**		**Click to Place** **(*N* = 11 students,** **44 items)**	
	n	**%**	***n***	**%**
Difficulty with object selection	6	10.0	16	37.2
Difficulty with continuous selection	7	11.5	–*	–
Difficulty with group selection	6	10.0	26	60.5
Difficulty with number of objects	2	3.0	10	23.3
Needed assistance to complete	7	11.5	26	60.5

*Note: *Click-to-place items do not require continuous selection.*

Sources of challenge in responding to multi-select multiple-choice 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 multi-select multiple-choice items is shown in Table 89. On 41% of the items, students had difficulty with the concept of making multiple selections.

Table 89. Sources of Challenge in Response to Multi-select Multiple-Choice 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 multi-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 90 summarizes student responses to post-hoc interview questions. Students more often liked drag-and-drop and select-text items, perceived them as easy, and understood the response process required. Students viewed multi-select multiple-choice items less positively and reported the most difficulty with click-to-place items. Student interview responses were consistent with evaluations of item effectiveness based on sources of challenge noted by the observers.

Table 90. Affirmative Student Responses to Post-Hoc Interview Questions

Question	Drag and Drop (*n* = 15)		Click to Place (*n* = 11)		Multiple Select (*n* = 11)		Select Text (*n* = 8)	
	n	%	*n*	%	*n*	%	*n*	%
Did you like it?	15	100.0	7	63.6	9	81.8	8	100.0
Was it easy?	15	100.0	8	72.7	10	90.9	8	100.0
Was it hard?	1	6.0	1	9.0	1	9.0	1	12.5
Did you know what to do?	14	93.3	6	54.5	8	72.7	8	100.0

IX.2.A.ii. Teacher Cognitive Labs

Teacher 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 teacher-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 teacher administered at all linkage levels.

Cognitive labs were conducted in spring 2015 with 15 teachers in five schools across two states. Teachers completed think-aloud procedures while preparing for and administering teacher-administered testlets in reading, writing, and math. They were first presented with the Testlet Information Page (TIP), which is a short document that provides background information needed to prepare to administer the testlet. For example, a TIP may contain instructions about materials needed, guidelines for material substitution, instructions about alternate text to be read aloud when describing pictures to students with visual impairments, and an indication that calculator use is appropriate on a specific math testlet.

Teachers were asked to think out loud as they read through the TIP. Next, the teacher gathered the materials needed for the assessment and administered the testlet. Probes were sometimes used during the process to ask about teacher interpretation of the on-screen instructions and the rationale behind decisions they made during administration. When the testlet was finished, teachers 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.

- Teachers were easily able to understand student intent when the student indicated a response by picking up objects and handing them to the teacher.
- In a case where the student touched the object rather than handing it to the teacher, the teacher 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 teacher 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 teacher 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 teacher 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 teacher 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.

Additional data collection is anticipated, particularly as instructions to test administrators are refined during future phases of test development.

IX.2.B. Evaluation of Test Administration

Two studies were conducted to better understand response processes and test administration procedures. Data were collected during test administration observations and cognitive labs at participating schools during the 2014–2015 academic year.

IX.2.B.i. Observations of Test Administration

Test administration observations were conducted to further understand response processes for students. Observations were conducted in multiple states during field testing in spring 2014 and operational assessments in spring 2015 and 2015–2016.[35] The student's typical test administration process was observed, with the student's actual test administrator. School administrations were also observed for the full range of students eligible for DLM assessments (students with the most significant cognitive disabilities).

The DLM Consortium used a test-administration observation protocol to gather information about how educators in the consortium states deliver testlets to students with the most significant cognitive disabilities. This protocol gave observers a standardized way to describe the way a DLM testlet was administered—no matter their role or experience with DLM assessments. The test-administration observation protocol captured data about student actions

[35] The timing of this chapter's development allowed for inclusion of observational data collected through the first half of the 2015–2016 school year. Although the manual covers the 2014–2015 administration, the 2015–2016 observation data are included for completeness.

(navigation, responding, etc.), educator assistance, variations from standard administration, engagement, and barriers to engagement. Test-administration observations were collected by DLM project staff, as well as state education agency and local education agency staff. The observations protocol was only used for descriptive purposes; it was not used to evaluate or coach the educator or to monitor student performance. Most items were a direct report of what was observed, for instance, how the test administrator set up for the assessment, and what the test administrator and student said and did. One section asked observers to make judgments about the student's engagement during the session.

During computer-administered testlets, the intent was that students could interact independently with a computer, using special devices such as alternate keyboards, touch screens, or switches as necessary. In teacher-administered testlets, the test administrator was responsible for setting up the assessment, delivering it to the student, and recording responses in the KITE™ system. The test-administration protocol contained different questions specific to each type of testlet.

Test administration observations were collected in five states beginning in 2014 and continuing through February 2016. The numbers of observations collected by state are shown in Table 91.

*Table 91. Teacher Observations by State (*N *= 147)*

State	*n*	%
Alaska	5	3.4
Iowa	45	30.6
Kansas	1	0.7
Mississippi	1	0.7
Missouri	95	64.6

Of the 147 test-administration observations collected, 117 (79.6%) were of computer-delivered assessments and 30 (20.4%) were of teacher-administered testlets. Of the 147 observations, 70 (47.6%) were of ELA reading testlets, 32 (21.8%) were of ELA writing testlets, 40 (27.2%) were of math testlets, and one (0.7%) was of a science[36] testlet. Most testlets were administered in students' usual classrooms (81.6%).

To investigate the assumptions that underlie the claims of the validity argument, several parts of the test-administration observation protocol corresponded to assumptions. One assumption addressed is that educators allow students to engage with the system as independently as they are able. For computer-administered testlets, related evidence is summarized in Table 92, with behaviors identified as supporting, neutral, or non-supporting. For example, clarifying

[36] DLM science testlets were field-tested during one of these observations.

directions (26% of observations) removes student confusion over 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 (such as hand-over-hand guidance) is a clear indicator that the teacher directly influenced the student's answer choice.

Table 92. Test Administrator Actions During Computer-Administered Testlets (N = 117)

Evidence	Action	n	%
Supporting	Used verbal prompts to direct the student's attention	65	55.6
	Clarified directions	30	25.6
Neutral	Navigated one or more screens for the student	85	72.6
	Repeated question(s) before student responded	76	65.0
	Defined vocabulary used in the testlet	34	29.1
	Repeated question(s) after student responded	11	9.4
	Asked the student to clarify one or more responses	10	8.5
Non-supporting	Used physical prompts	30	25.6
	Reduced number of choices available to student	6	15.1

Note: Respondent 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 73% of the observations is not necessarily an indication that 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 assessment, navigating for students who are able to do so independently would be counter to the assumption that students are able to interact with the system as intended. The observation protocol did not capture the reason the test administrator chose to navigate, and the reason was not always obviously inferred just from watching.

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 (see Table 93). Independent response selection was observed in 39% of the cases and the use of eye gaze (one unique form of independent selection that was recorded separately) was seen in 21% of the observations. Verbal prompts for navigation and response selection are strategies that are within the realm of allowable flexibility during test administration. These strategies, which are commonly used during direct instruction for students with significant cognitive disabilities, would be 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 93. Student Actions during Computer-Administered Testlets (N = 117)

Action	n	%
Navigated the screens independently	19	16.2
Navigated the screens with verbal prompts	8	6.8
Selected answers independently	45	38.5
Selected answers with verbal prompts	53	45.3
Indicated answers using eye gaze	24	20.5
Indicated answers using materials outside of KITE	4	3.4
Used manipulatives	30	25.6

Note: Respondent could select multiple responses to this question.

Another assumption in the validity argument is that students are able to respond to tasks irrespective of a sensory, mobility, health, communication, or behavioral constraint. This assumption was evaluated by having observers note whether there was difficulty with accessibility supports (including lack of appropriate available supports) during observations of teacher-administered testlets. Of the 30 observations of teacher-administered testlets, observers noted difficulty in two cases (6.7%). For computer-delivered testlets, evidence to evaluate this assumption was observed by noting students' abilities to indicate responses to items using multiple response modes such as sign language, eye gaze, and using manipulatives or materials outside of KITE. A summary of the frequencies of these behaviors is shown in Table 94. Additional evidence for this assumption was gathered by observing whether students were able to complete testlets. Of the 147 test-administration observations collected, in 132 cases (89.8%) students completed the testlet.

Another assumption underlying the validity argument is that test administrators enter student responses with fidelity. Observers rated whether test administrators accurately captured student responses. In order to record student responses with fidelity, test administrators needed to observe multiple modes of communication, such as verbal, gesture, and eye gaze. Table 94 summarizes students' response modes for teacher-administered testlets.

Table 94. Primary Response Mode for Teacher-Administered Testlet (N = 30)

Response mode	*n*	%
Verbal	7	23.3
Gesture	12	40.0
Eye gaze	2	6.7
Other	6	20.0
No response	5	16.7

Note: Respondent could select multiple responses to this question.

Across all observations and student response modes, test administrators recorded responses with fidelity in 93.3% of observations.

Computer-administered testlets provided another opportunity to confirm fidelity of response entry when test administrators entered responses on behalf of students. This is a support recorded on the Personal Needs & Preferences 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 75 of 98 observations of computer-administered testlets, the test administrator entered responses on the student's behalf. In 98.6% of those cases, observers indicated that the entered response matched the student's response. This evidence supports the assumption that test administrators entered student responses with fidelity.

IX.2.B.ii. Test Administrator Feedback Studies

Test administrators provided feedback after administering field tests and operational assessments. Survey data that inform evaluations of assumptions regarding response processes include test administrator perceptions of student ability to respond as intended, free of barriers, and test administrator perceptions of the ease of administering teacher-administered testlets. Perceptions of student response come from the spring 2015 test administrator survey[37]. Feedback on the process of delivering a particular subtype of teacher-administered testlet (writing testlets) came from 305 test administrators who responded to surveys after field-testing writing assessments prior to the spring 2015 window.

The spring 2015 test administrator survey included three items about students' ability to respond. Test administrators were asked to rate statements from *strongly disagree* to *strongly*

[37] Recruitment and response information for this survey was provided earlier in this chapter.

agree. Results are presented in Table 95. The majority of test administrators agreed or strongly agreed that their students (1) responded to items to the best of their knowledge ability, (2) were able to respond regardless of disability, behavior, or health concerns, and (3) had access to all necessary supports to participate.

Table 95. Test Administrator Perceptions of Student Experience with Testlets, Spring 2015

	Strongly Disagree		Disagree		Agree		Strongly Agree	
Statement	*n*	%	*n*	%	*n*	%	*n*	%
Student responded to items to the best of his/her knowledge and ability	271	9.3	358	12.3	1535	52.6	752	25.8
Student was able to respond regardless of his/her disability, behavior, or health concerns	503	17.3	522	18.0	1398	48.2	479	16.5
Student had access to all necessary supports to participate	206	7.1	276	9.5	1644	56.6	777	26.8

Educators who administered DLM writing testlets as field tests in early 2014-15 (referred to as Phases B and C) were invited to complete surveys about their experience. Writing testlets are similar to other DLM teacher-administered testlets in reading and mathematics, where the test administrator engages in a scripted activity with a student outside the KITE system and then enters observations and ratings of the student’s behavior into KITE. Data collected from these surveys related to response processes included educator judgment of ease of administration, appropriateness of testlet answer options, and potential barriers related to student use of writing tools.

During Phase B, 108 educators in eight states administered writing testlets to their students in grades 4, 8, and 11, and then responded to the survey. During Phase C, 197 educators from seven states administered writing testlets in grades 3-8 and high school and responded to the survey. In both Phase B and Phase C (46% and 53%, respectively), more students were offered emergent writing testlets than conventional (17% and 16% for Phase B and C, respectively). Some test administrators were unable to make the judgement, marking “Not Sure”; however, fewer administrators marked “Not Sure” in Phase C (9%) than in Phase B (37%).

Respondents were asked to consider one student to whom they administered a writing testlet in order to collect information on ease of administration and appropriateness of answer options in the writing testlets. Table 96 reports results of the test administrator survey about the ease of administration. Educators who did not understand how to deliver the testlet could introduce construct-irrelevant variance into the assessment administration. Most educators rated the ease of administration as *somewhat easy* or *very easy* during Phases B and C. Table 97 includes a summary of educator ratings of the match between the answer options and observed student behavior in response to items in the writing testlets. Most educators reported a match between answer options and student behaviors on some or all of the items.

Table 96. Percent of Test Administrators Rating Ease of Testlet Administration

Ease of Administration	Phase B	Phase C
Not at all Easy	24	24
Somewhat Easy	40	50
Very Easy	21	20
No Response	14	6

Table 97. Percent of Test Administrators Reporting Match Between Answer Options and Student Response

Portion of Items	Phase B	Phase C
All of the Items	30	23
Some of the Items	43	58
None of the Items	18	17

IX.3. EVIDENCE BASED ON INTERNAL STRUCTURE

Analyses that address the internal structure of an assessment indicate the degree to which "relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA et al., 2014, p. 16). 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).

IX.3.A. EVALUATION OF ITEM-LEVEL BIAS

Differential item functioning (DIF) addresses the broad problem created when some test items are "asked in such a way that certain groups of examinees who are knowledgeable about the intended concepts are prevented from showing what they know" (Camilli & Shepard, 1994, p. 1). Studies that use 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 DIF does not always indicate a weakness in the test item, it can help point to construct-irrelevant variance or unexpected multidimensionality, thereby contributing to an overall argument for validity and fairness.

IX.3.A.i. Method

The initial DIF analysis for items in the DLM alternate assessment was conducted using data collected during the spring 2015 administration. Because 2014–2015 was the first operational year of DLM assessments and DIF analyses were dependent upon the amount of data collected for each item, the initial DIF analyses examined only performance for male and female subgroups. As additional data is collected in subsequent operational years, the scope of DIF analyses will be expanded to include additional items, subgroups, and approaches to detecting DIF.

Items were selected for inclusion in the initial DIF analyses based on minimum sample size requirements for the two groups. Within the DLM population, the number of female students responding to items is smaller than the number of male students by a ratio of approximately 1:2; therefore, a threshold for item inclusion was imposed whereby the female 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 initial DIF analyses described here because they are scored at the option level rather than item level. Only operational content meeting sample size thresholds was included in the initial DIF analyses.

Using the above criteria for inclusion, 2,096 multi-EE items (39%) were selected for inclusion in the analysis. The number of items evaluated for evidence of DIF by grade and content area ranged from 83 in grade 9 ELA to 144 in grade 7 math. Sample sizes for multi-EE items were between 258 and 3,091.

For each item, logistic regression was used to predict the probability of a correct response given group membership and total linkage levels mastered by the student in the content area. The logistic regression equation for each item included a matching variable comprised of the student’s total linkage levels mastered in the content area of the item and a group membership variable, with females coded zero as the focal group and males coded one as the reference group. An interaction term was included to evaluate whether non-uniform DIF was present for each item (Swaminathan & Rogers, 1990), which, when present, is indicative that the item functions differently as a result of the interaction between total linkage levels mastered and gender. Said another way, when non-uniform DIF is present, the gender group with the highest probability of a correct response to the item differs along the range of total linkage levels mastered, whereby one group is favored at the low end of the spectrum and the other group is favored at the high end of the spectrum.

Three logistic regression models were fitted for each item:

$$M_0: \mathit{logit}(\pi_i) = \alpha + \beta X + \gamma_I + \delta_i X$$

$$M_1: \mathit{logit}(\pi_i) = \alpha + \beta X + \gamma_I$$

$$M_2: \mathit{logit}(\pi_i) = \alpha + \beta X$$

where π_i is the probability of a correct response to the item for group i, X is the matching criterion, α is the intercept, β is the slope, γ_I is the group-specific parameter, and $\delta_I X$ is the interaction term.

Due to the number of items being 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 the gender 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_2 to M_1 or M_0, to account for the impact of the addition of the group and interaction terms to the equation. All effect-size values are reported using both the Zumbo & Thomas (1997) and Jodoin & Gierl (2001) indices for reflecting a negligible, moderate, or large effect. The Zumbo & Thomas thresholds for classifying DIF effect size are based off of Cohen's (1992) guidelines for identifying a small, medium, or large effect, with corresponding thresholds of 0.13 and 0.26 for distinguishing negligible, moderate, and large effects. The Jodoin & Gierl approach expanded on the Zumbo & Thomas effect-size classification by basing the effect-size thresholds for the Simultaneous Item Bias Test procedure (Li & Stout, 1996), which like logistic regression, also allows for the detection of both uniform and non-uniform DIF, and makes use of classification guidelines that are based on the widely accepted ETS Mantel-Haenszel classification guidelines. The Jodoin & Gierl threshold values for distinguishing negligible, moderate, and large DIF are 0.035 and 0.07, whereby items with an effect size less than 0.035 are classified as having negligible DIF, and so on. Similar to the ETS method, negligible effect is classified with an A, moderate effect with a B, and large effect with a C.

Jodoin & Gierl (2001) also examined Type I error and power rates in a simulation study examining DIF detection using the logistic regression approach. Two of their conditions featured a 1:2 ratio of sample size between the focal and reference groups. As with equivalent sample-size groups, the authors found that power increased and Type I error rates decreased as sample size increased for the unequal sample size groups. Decreased power to detect DIF items was observed when sample size discrepancies reached a ratio of 1:4.

IX.3.A.ii. Results

Uniform DIF Model. A total of 138 items were flagged for evidence of uniform DIF when comparing M_1 to M_2. Table 98 summarizes the number of items flagged for evidence of uniform DIF by content area and grade for each model. The percent flagged for each grade and content area ranged from 1 to 17.

Table 98. Items Flagged for Evidence of Uniform DIF

Content Area	Grade	n	*N*	%
ELA	3	7	104	7
	4	9	108	8
	5	12	130	9
	6	11	107	10
	7	8	90	9
	8	11	99	11
	9	11	83	13
	10	4	90	4
	11	16	95	17
Math	3	9	126	7
	4	6	139	4
	5	5	139	4
	6	5	141	4
	7	7	144	5
	8	8	135	6
	9	4	128	3
	10	1	128	1
	11	4	110	4

Using the Zumbo and Thomas (1997) effect-size classification criteria, all 138 items were found to have a negligible change in effect size after adding the gender term to the regression equation.

Using the Jodoin & Gierl (2001) effect-size classification criteria, two items were found to have a moderate or large effect size and the remaining 136 items were found to have a negligible change in effect size after adding the gender term to the regression equation.

Information about the flagged items with a moderate and large change in effect size after adding in the gender term is summarized in Table 99. One ELA item had a large effect size value, as represented by a value of C. One math item had a moderate effect size value, as represented by a value of B. The γ values in the table indicate which group was favored on the item after holding total linkage levels mastered constant, with negative values indicating that the reference group (males) had a higher probability of success on the item.

Table 99. Items Flagged for Uniform DIF with Moderate or Large Effect Size

Content Area	Grade	Item	EE	χ^2	*p* value	γ	R^2	Z & T effect size	J & G effect size
ELA	4	30807	RL.4.5	6.16	0.013	-2.04	0.09	A	C
Math	9	24871	HS.A-AAE.1	22.62	< 0.000	-1.07	0.04	A	B

Combined Model. A total of 258 items were flagged for evidence of DIF when both the gender and interaction terms were included in the regression equation. Table 100 summarizes the number of items flagged for either uniform or non-uniform DIF by content area and grade for each model.

Table 100. Items Flagged for Evidence of DIF for the Combined Model

Content Area	Grade	Count of Flagged Items	*n*	%
ELA	3	12	104	12
	4	13	108	12
	5	13	130	10
	6	8	107	7
	7	10	90	11
	8	16	99	16
	9	10	83	12
	10	4	90	4
	11	17	95	18
Math	3	19	126	15
	4	14	139	10
	5	13	139	9
	6	23	141	16
	7	19	144	13
	8	33	135	16
	9	9	128	7
	10	10	128	8
	11	15	110	14

Using the Zumbo and Thomas (1997) effect-size classification criteria, one item was found to have a moderate change in effect size. The remaining 257 items were found to have a negligible change in effect size after adding the gender and interaction terms to the regression equation.

Using the Jodoin & Gierl (2001) effect-size classification criteria, two items were found to have a moderate change in effect size, one item was found to have a large change in effect size, and the remaining 255 items were found to have a negligible change in effect size after adding the gender and interaction terms to the regression equation.

Information about the flagged items with a moderate or large change in effect size is summarized in Table 101. Two ELA items and one math item had moderate or large changes in effect-size values, as represented by a value of B or C respectively. Only one item was classified as beyond a negligible change in effect size by both criteria.

Table 101. Items Flagged for DIF with Moderate or Large Effect Size

Grade	Item	EE	χ^2	*p* value	γ	$\delta_i X$	R^2	Z & T *	J & G*
				ELA					
4	30807	RL.4.5	10.30	0.006	0.48	-0.11	0.15	B	C
10	26222	RL.9-10.4	6.06	0.048	1.70	-0.05	0.04	A	B
				Math					
9	24871	HS.A.SSE.1	22.63	0.000	-1.11	0.00	0.04	A	B

*Note: * Effect size measure*

Appendix F includes plots (Figures A–C) labeled by the item ID, which display the best fitting regression line for each gender group, along with jittered plots representing the total linkage levels mastered for individuals in each gender group.

Next Steps. After additional data is collected during the 2016 operational year, items flagged for evidence of DIF with either a moderate or large effect size change will be given the priority for further analysis by content and psychometric teams. Depending on their review, items may be subject to further analysis (e.g., cognitive labs, panel reviews). Decisions to revise or remove items or testlets will not be made based on results of flagging alone.

IX.4. Evidence Based on Relations to Other Variables

According to *Standards for Educational and Psychological Testing*, "analyses of the relationship of test scores to variables external to the test provide another important source of validity evidence" (AERA et al., 2014, p. 16). Test administrator perception of testlet difficulty serves as one source of external evidence for the DLM Alternate Assessment System.

Prior to administering testlets, educators complete the First Contact survey, which is a survey of learner characteristics[38]. Responses to the survey determine the complexity band the student is

[38] More information on the First Contact survey and student classification to complexity bands can be found in Chapter III.

placed into, which is then used to assign the linkage level for the first testlet administered to the student during operational assessment. Field Test 3 was the first opportunity for students to be assessed at a single linkage level based on responses to the First Contact survey, which provided an opportunity to evaluate the relationship between the DLM-assigned complexity band and test administrators' perception of testlet difficulty.

Following the administration of testlets in Field Test 3, test administrators were asked to report their views on testlet difficulty for individual students. Responses were then evaluated by complexity band. Across complexity bands 1, 2, and 3, test administrators reported the difficulty of most testlets was *about right* for the student. However, test administrators indicated testlets administered at the Foundational (lowest) band were too hard for many students. Table 102 summarizes the reported difficulty levels by student complexity band. These findings provide evidence that most test administrators believe students receive content of appropriate difficulty as assigned by the student's complexity band, which is based on First Contact survey responses.

Table 102. Test Administrator-Reported Testlet Difficulty

Complexity Band	**Too Easy**		**About Right**		**Too Hard**	
	n	%	*n*	%	*n*	%
ELA Foundational	11	4	119	46	128	50
ELA Band 1	56	12	298	62	122	26
ELA Band 2	92	18	379	73	50	10
ELA Band 3	64	23	191	67	27	10
Math Foundational	17	5	182	54	137	41
Math Band 1	86	17	340	67	85	17
Math Band 2	142	24	383	66	58	10
Math Band 3	53	30	113	63	13	7

IX.5. EVIDENCE BASED ON CONSEQUENCES OF TESTING

Validity evidence must include the evaluation of the overall "soundness of these proposed interpretations for their intended uses" (AERA et al., 2014, p. 19). In order to establish sound score interpretations and delimit score use, score reports must be useful and provide relevant information for teachers that informs instructional choices and goal setting. Teachers must use horizontal and vertical recommendations to plan subsequent instruction, and scores can only be interpreted and used for purposes called out in the theory of action as part of the validity argument. Evidence that the DLM Consortium developed score reports and interpretive resources to support intended uses and interpretations is provided in Chapter VII.

As educators and students become familiar with a new assessment during the first operational year, there is limited potential for consequential evidence. For 2014–2015, two sources of evidence were collected. Results are presented on a multi-stage research effort on score report design and interpretation. Baseline data are also reported for a longitudinal test administrator survey.

IX.5.A. DLM Score Report Design and Use

The DLM Consortium embarked on a series of studies to inform the development of, and evaluate the effectiveness of, individual student score reports. First, focus groups were conducted in five states with parents of children with disabilities (Nitsch, 2013) to learn about parent perceptions of alternate assessments based on alternate achievement standards (AA-AAS) and parent need for information about student performance. Parents rated themselves as having relatively little knowledge of AA-AAS and some indicated they had not received AA-AAS score reports from their schools. Parents tended to perceive the purpose of AA-AAS as to fulfill a legislative mandate and to drive decisions about the school (including educator evaluation and determination of resources) rather than to provide information about their child or measure things relevant to their child's learning. Concerns about the information parents received on AA-AAS results included lack of understanding of how scores were determined or how the content was related to academic content standards, unfamiliar terminology, focus on deficits more so than progress, and lack of information about how results could be used to change instruction or provide different supports to their child.

In 2014, additional focus groups were conducted with parents, advocates, and educators (Clark et al., 2015). Participants evaluated prototype score reports. Prototypes were refined between waves of feedback, with the goal of maximizing the clarity of the contents and supporting accurate interpretations. Preliminary evidence supported educators' ability to interpret the reports' contents. Parents appreciated the emphasis on strengths rather than deficits but expressed concern about educators' ability to communicate about the contents. Participant feedback led to many of the features seen in the 2014–2015 score reports, including narrative statements and linkage level descriptors for every EE (see DLM System Design, below, for more information about report contents).

Building on the previous research that informed score report design (Nitsch, 2013) and refinement (Clark et al., 2015), the purpose of this study was to evaluate educators' interpretations and use of DLM individual student score reports. Specific research questions included:

1. How do participants read and interpret the information in reports?
2. How do participants explain results to parents?

3. What resources do participants use to support their interpretation and use of report contents?
4. How do participants use report contents for educational planning and instruction?

IX.5.A.i. Methods

As described in Chapter VII, the Performance Profile aggregates linkage level mastery information for reporting on each conceptual area and for the subject overall. The Learning Profile shows rows for each EE and columns that correspond to the five linkage levels (Initial Precursor, Distal Precursor, Proximal Precursor, Target, and Successor). Table 103 summarizes the components of the Performance Profile and Learning Profile that make up the individual student score report. These components were part of the coding scheme used for data analysis and are referred to by number throughout the results section.

Table 103. Components of the DLM 2014–2015 Individual Student Score Report

Performance Profile	Learning Profile
1) Overall performance level: a) narrative b) graphic c) performance level descriptors 2) Conceptual areas: bar graphs with subtitles 3) Mastery list: a) Conceptual area headings b) Introductory statement c) Bulleted statements	4) Learning Profile narrative 5) Conceptual area and Essential Element codes 6) Mastery information: a) Mastered (green) b) No evidence of mastery (blue) c) Untested (no shading)

Results were based on individual interviews and paired interviews conducted with teachers in one state. Protocols were slightly different for individual and paired interviews, but both versions were semi-structured.

The individual interview protocol began with general questions about the participant's background with DLM assessments and previous experience with the score reports. Then the participant was presented with the first score report and asked what it said about the student. Participants were asked to think aloud while they read the contents. Probes were used for clarification of responses and to ensure participants attended to each part of the report (e.g., to point them back to a section they skipped). After interpreting each section of the report (i.e., Performance Profile and Learning Profile), the participant was asked how they might say things differently when explaining the report to a parent. The same process (initial interpretation and reinterpretation for a parent) was followed for a second, contrasting report. The interview concluded with an opportunity for the participant to make recommendations about resources that other teachers would need to support their interpretation and use of DLM score reports.

The paired interview began with the same general background questions as the individual interview but also included a question about the participants' history of collaboration. The pair was then presented with a score report and asked to talk aloud about their interpretation of its contents. The primary focus of the interview was the use of the report to plan for instruction, including long-term educational planning and mid-year adjustments to instruction. Participants engaged in unstructured dialog about the contents and probes were used during the dialog as needed for clarification and elaboration to cover both major categories of use (instruction and IEP planning). After repeating the process with a second, contrasting report, the interview concluded with an opportunity for recommendations about resources to support score report interpretation and use.

Both types of interviews used 2014–2015 score reports with realistic student results but fictitious student identifiers. Sample score reports were prepared in both subjects (ELA and math) and across elementary, middle, and high school grades. Samples were also selected within each subject and grade band to provide contrasting patterns of student performance.

Each interview incorporated two sample reports. The choice of specific reports for each interview were based on the participant's familiarity with the grade band and subject. For example, a middle school educator who was responsible for both ELA and math might be presented with an ELA grade 6 report for a high-achieving student and a math grade 7 report for a low-achieving student. There was no intentional sequence in which report was presented first.

Educators were all from one campus in an urban area in a Midwestern state. The school exclusively serves students with intellectual and multiple disabilities from sixth grade through age 21. Participants taught in secondary grades (grades 6-8, grades 9-10, or grades 11-12). All of them taught two or more academic subjects. Their years of teaching experience ranged from 1 to 26 years. Four educators participated in individual interviews and four more participated in two paired interviews.

Individual interviews were coded using a two-step process. First, the researcher reviewed each transcript to mark responses related to the primary research questions (i.e., reading and interpretation, explanation to parents, resources to support interpretation, and uses of report contents). During the second step, the researcher added codes to identify the part of the report the participant was referring to. Thematic codes were also used to identify processes or elements associated with the primary codes. For example, within responses coded as reading and interpretation, statements were also coded to indicate the types of behaviors (e.g., paraphrase, question about contents, misinterpretation). A tentative list of codes was developed prior to analysis, based on review of the literature. Codes were added and refined as new ideas emerged from the data. Paired interviews relied on many of the same codes as individual interviews, but the emphasis was primarily on uses of the contents rather than interpretation.

Since the results presented in this manual are preliminary, they are descriptive with regard to the themes, not quantified for dominant patterns.

IX.5.A.ii. Results

Reading and Interpretation. Participants varied in the parts of the report that they tended to rely on for information. Results are described with numeric references back to the report component listed in Table 103.

Since the interview imposed minimal structure on the order in which participants reviewed the report and the emphasis they placed on each section, each participant's preferences for information were clear in the think-aloud portion of the interview, even before discussing the report contents. For example:

- Anna[39] walked systematically through each major section of the entire report, starting with the Performance Profile narrative (1a) to characterize the student's overall performance, describing conceptual areas (2) as general strengths and weaknesses, and using the mastery list (3) to reflect on skills seen during the assessment. In the Learning Profile she emphasized the mastery information (6) and did not use the narrative (5).
- Liz briefly mentioned the numbers in the Performance Profile narrative (1a) and spoke briefly about all parts of the Performance Profile but had a strong preference for the mastery information (6) in the Learning Profile.
- Margaret primarily relied on the conceptual areas (2) and looked to the mastery list bullets (3c) to identify examples of the skills in each area, especially when talking to parents. When thinking about instruction, she gravitated to the mastery information (6) in the Learning Profile.

In general, participants paid little attention to narrative statements (1a, 4) and only one briefly mentioned the performance level graphic (2). The Performance Profile mastery statements (3) and Learning Profile mastery table (6) were emphasized the most. More detail about interpretation of the Learning Profile is provided in the Report Use section below.

As participants talked through the report contents, most of their comments were verbatim or near verbatim language from the report. Minimal paraphrasing was occasionally used when interpreting results for parents:

> *I basically sort of explained the [performance] levels first . . . so I said emergent is they're just starting out with this skill. They may not have a good understanding. And then I said approaching target, they have some understanding. And then I said target is right where we want them.*

Statements about report contents were also evaluated for signs of misinterpretation or misunderstanding. Since most statements were verbatim or near verbatim, there were few opportunities for misinterpretation. One type of misinterpretation came from inappropriately

[39] All names are pseudonyms.

applying terms from one part of the report to results in other sections. For example, in one case a student was described as "emerging" (a performance level descriptor) in one of the conceptual areas although there are no performance levels assigned to conceptual areas. In another case, the student was described as having "mastered" a conceptual area although mastery judgments are only made at the linkage level. Both of these misstatements were attempts to give a qualitative label to a percent of skills mastered in a conceptual area.

One participant misinterpreted the percent values reported for conceptual areas when talking to parents. Instead of describing percent of skills mastered, she interpreted percent as it is often used in monitoring instruction and setting instructional goals for students with the most significant cognitive disabilities: percent accuracy or percent correct over repeated trials.

> *So it's like constructs understanding [Conceptual Area]—he can identify concrete details in an informational text [linkage level]. But reminding the parent that that was only like a 20 percent. . . . But it seems that oh, my child can identify that. Then you're like, well, but if we look back here, again, remember, that was one out of five times. So it's still only with 20 percent accuracy, which is—you want 80 percent. So definitely make sure they understand that like a target child, that goal is about 80 percent for their classmates.*

The most extreme misconception was seen for one participant who asked many questions that reflected his confusion. Some of his challenge was in relating the score report contents to the assessment design and administration. He could not recall how testlets were assigned or the relationship between the linkage level tested and where mastery would be reported. He also wanted to see information in the Performance Profile (i.e., which skills were not mastered) without realizing it was in the Learning Profile. He reported using the Performance Profile bulleted mastery list with parents and the Learning Profile to think about instruction.

IX.5.A.iii. Interpreting Reports for Parents

Each participant indicated that they were selective about the parts of the report they chose to discuss with parents. Most commonly mentioned were the Conceptual Area (CA) bar graphs (2), bulleted mastery list statements (3a), and the entire Learning Profile. For example, one teacher used the CA bar graphs to explain the student's general strengths and weaknesses before discussing more specific skills from the bulleted list as examples from specific CAs. Those who preferred to discuss the Learning Profile with parents pointed out that it allowed them to focus on current mastery **and** areas for instruction, whether that be to reteach something that was not mastered or move to another skill after mastering a previous one. The participant who reported less discussion of the report with parents said she focused only on the CA bar graphs and referenced a couple of skills from the Learning Profile. Her rationale was that parents' best level of understanding was in the CAs. She sent the report home with them and invited them to ask her questions after they looked it over on their own.

Although the mastery list (3) and the Learning Profile (6) contained very similar information, some teachers preferred one over the other. Those who preferred the bulleted mastery list tied the CA headings (3a) back to the bar graphs to help anchor their conversation with the parent. When discussing results that did not resonate with parents (e.g., the student demonstrated mastery of a skill the parent thought was implausible or did not demonstrate mastery of a skill the parent believed the student possessed), another strategy was to refer to the introductory statements (3b) to remind the parent that the report was explaining evidence of mastery from the DLM assessments and that there were multiple ways the student might demonstrate the skill.

As participants described the ways in which they talked with parents about report contents, it became clear that they added contextual information to support parents' understanding. For example, one teacher drew connections to the reports for the general education assessments and content standards, since many parents were familiar with those for other children in their family. Another strategy was to explain why the assessment was challenging that year (e.g., that the assessment was still relatively new, or that they expected the student to improve after becoming more familiar with working in a computer-based environment).

When discussing specific mastery statements or linkage levels from the Learning Profile, another contextualizing strategy was to describe what the skill looked like for that student, either during assessment or during instruction. One participant modeled how she would talk to a parent about an EE that had no evidence of mastery on the Learning Profile:

> *I even have parents with some intellectual needs. I would actually say it to them that your student—you see these highlighted areas right here in the blue? These areas were the areas where they're struggling, right here, and these areas are the areas that they did really well, and we want to focus on those areas where they were struggling, and right here—so understanding function of the objects—okay, what does that mean? So let's say, we need [the student] to understand that when she goes over and turns that light on—so understanding what that means, we're going to work on that.*

Yet describing skills to parents was difficult when teachers themselves did not understand the linkage level statement. Two types of challenges were noted. First, academic vocabulary was seen as a barrier to talking with parents about the report. One participant, commented on the word "subitizing" in a linkage level descriptor:

> *I had that word and we were like what does that mean? We had to get on our phone and look it up to see what it meant, and it was like I can't even teach it if I don't know what it means, and how does a parent understand it if we don't know what it means?*

A second challenge occurred when two similar linkage level statements were difficult to distinguish from one another. One participant illustrated this challenge as she talked through her understanding of *Match pictures with representations of real objects* and *Match pictures with real objects*:

> *That says matching pictures with representation of real objects. That's interesting. Match a picture with a real object. . . . I might have a parent ask me why did they do well*

> *here and they didn't do well here? Why did they not do well there and they did well here? . . . So, these are two different areas. This one is in the—I'm going to get this wrong. One is in reading . . . reading, and yes, and this one is . . . reading information, right. Okay, yes. I know, but I'm missing it, but okay, yes, yes. So this is in the story itself. This is in the story itself. So when she's reading the story and understanding, she's getting that information. Okay. She's able to match pictures with, yes, okay. And this is just absolute picture, just like, identifying. Okay. All right.*

IX.5.B. Teacher Resources

All teachers in this preliminary study were from the same campus. The campus had an instructional facilitator and built-in time for both structured professional development sessions and professional learning community meetings. All of the participants credited those resources with helping them interpret and use the score reports. For example, they had a one-hour professional development session on how to read the score reports. In the professional learning community meetings, they planned for assessment, shared materials and resources, and helped one another with interpretation of linkage levels. Several participants mentioned talking with the student's teacher from the previous year (whether from within their school or at another school) to better understand how a student was demonstrating a skill that was listed as mastered on the score report.

IX.5.B.i. Report Use for Planning Instruction

Participants described a range of uses of the report contents beyond sharing the results with parents. For this manual, uses are roughly grouped into planning for instruction and IEP development.

Planning for Instruction. A consistent finding across interviews was teachers' use of the Learning Profile to guide instruction. This included looking to the next linkage level beyond the highest level mastered for a given EE and planning to instruct next on that level. However, where students were assessed and did not show mastery, or where teachers thought the student's mastery was limited, teachers indicated they would reteach a skill that the student had already mastered.

Some participants provided evidence of more sophisticated evaluation and planning, particularly by looking at connections across linkage levels and EEs to think about larger instructional goals.

> *Because he's mastered the Level 3 which is the precursor—so we want him to get up to the target, so I would start teaching for the target for the student, tying it back into the precursor stuff that he can do so that we're not working on stuff that he already knows.*
>
> *So if we can connect those two Elements there, we know that we can start up here with them on this one and I'd have to explain that to a parent and then I would want to know where he's at with this. Once we teach him how to do that, how fast is he going to pick*

> *that up to doing the real-world problems with numbers and if he can do real-world problems up here with numbers, can he do it the same way here? This is adding and subtracting—so this is multiplying, so it would be different, but how is it different there and the same there.*

Sometimes an apparently inconsistent or unusual pattern of performance raised questions for the teacher. The typical response was a desire to assess further using their routine classroom methods to understand possible reasons for the inconsistency:

> *He can combine and partition sets, which should lead to multiplying. I don't understand why he can do multiplying in one but not combining in another. I guess I would want to take a look at that one and see how those lead to each other because combining and portioning are the same I guess for both multiplication and adding and subtracting.*

When planning for instruction in an area the student had not mastered, the teacher sometimes relied on understanding of the DLM assessment content. One common instructional strategy for students with the most significant cognitive disabilities is to first teach a skill in a familiar context and then work on transferring the skill to novel situations. One participant describing instruction on *Identify the end of a familiar routine* offered this example related to a reading testlet:

> *What type of routine for it? I know that on the assessments that was really hard for me to think of what type of routine are we using . . . because the example has you doing stuff out of a book and that's the routine is what's in the book but then how do you end that routine? . . . Well what do we do at the end of math? It all depends on the day. . . . Okay when we are getting ready to go on the bus, what's the last thing that you do? You buckle yourself in. Okay. That type of thing for familiarity.*

There were a few other ways in which teachers mentioned using the report to plan for instruction, but none of them was described in depth. Examples included using the Learning Profile to develop lesson plans and creating instructional groupings when students working on different skills were being taught together.

IEP Planning. Participants described using score report contents primarily for two parts of IEP development: statements on the student's present levels of performance and annual goals. The tendency was to use the performance level narrative (1a) and mastery skill list (3c) nearly verbatim in statements of present levels of performance:

> *I'd take this whole thing and say use this. So say over the assessment is covering fifty skills, for ten Essential Elements Hunter mastered 37 skills during the year and overall his mastery fell on to* at target*. And then I would say specifically what he has mastered. And then, if he didn't show skills: however, Hunter was tested, did not show these skills or he struggled with these skills, and then we'd say what he struggled with.*

The Learning Profile, and specifically the next skills that had not been mastered, were one source of information participants reported using to develop IEP goals. However, the expectation in their school was that the Learning Profile be considered along with other assessments and school-developed checklists in order to identify goals for the student in reading, writing, and math. The contents of IEP goals spanned multiple EEs, and the objectives associated with the goals were based on teacher estimates of reasonable instructional targets:

> *We look at all of the elements that are being assessed. We say where they're starting . . . We would look at where they're starting, either where they were assessed at or like this year we talked about they were at the Initial [Precursor] level. Most of our students are. And we created some scales, but we would look at where we felt like they could achieve within a year, and we kind of made it into a percentage. So this is where they're starting. These are the things that we would like to see them get to this year and so create a percentage within that.*

Besides these two uses of score reports to guide IEP development, one teacher pointed to another possible use of the information for IEP teams. When reviewing a sample score report that showed a student whose overall performance was at the highest performance level, she questioned that student's placement and eligibility for an alternate assessment. Both educational setting and assessment eligibility would be determined by an IEP team.

IX.5.C. Baseline Test Administrator Survey Responses

Test administrators were asked two questions on the spring 2015 survey[40] that assessed their perceptions of the assessment contents. These items will be repeated annually for longitudinal data collection. Test administrators completed these items based on their student with the best experience with DLM assessments, and again based on their student with the worst experience. Teachers who only administered a DLM assessment to one student only responded once. Table 104 summarizes the responses across all students: best experience, worst experience, and only student. Test administrators generally responded that content reflected high expectations for their students, but did not always agree that content measured important academic skills. DLM assessments represent a departure from many of the states' previous alternate assessments in the breadth of academic skills assessed. Given the short history of general curriculum access for this population and the tendency to prioritize functional academic skills for instruction (Karvonen, Wakeman, Browder, Rogers, & Flowers, 2011) test administrators' responses may reflect an awareness that DLM assessments contain challenging content, but that they are divided on its importance in the educational programs of students with the most significant cognitive disabilities.

[40] Recruitment and sampling described earlier in this chapter.

Table 104. Test Administrator Perceptions of Student Experience with Testlets, Spring 2015

	Strongly Disagree		Disagree		Agree		Strongly Agree	
Statement	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
Content measures important academic skills	697	23.9	770	26.4	1248	42.8	2917	6.9
Content reflects high expectations for this student	362	12.4	447	15.4	1608	55.2	495	17.0

IX.6. CONCLUSION

This chapter presents additional studies as evidence to support the overall validity argument for the DLM Alternate Assessment System. The studies are organized into categories (content, response process, internal structure, relations to other variables, and consequences of testing) as defined by the *Standards for Educational and Psychological Testing* (AERA et al., 2014), the professional standards used to evaluate educational assessments.

The final chapter of this technical manual, Chapter XI, references evidence presented through the technical manual, including this chapter, and expands the discussion of the overall validity argument. The concluding chapter also provides areas for further inquiry and ongoing evaluation of the DLM Alternate Assessment System.

X. TRAINING AND PROFESSIONAL DEVELOPMENT

The Dynamic Learning Maps 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: **required test administration training** and **optional professional development for instruction**. First, required test administrator training ensures that test administrators have both the context and practical knowledge of the assessment system design, administration, and security practices to administer the test with fidelity. All required test administrator training was therefore aligned with the *Test Administration Manual* (Dynamic Learning Maps, 2014a). See Chapter IV for a thorough discussion of test administration. The purpose of the professional development component is to provide professional learning opportunities to support instructional practices for the target population of students who participate in the DLM assessments.

Chapter X describes the training that was offered in 2014-15 for state and local education agency staff, the required test administrator training, and the optional professional development. Participation rates and evaluation results from 2014-2015 instructional professional development are included in this chapter (see Table 105 and Table 106 at the end of the chapter).

X.1. TRAINING FOR STATE EDUCATION AGENCY STAFF

State education agency staff are integral to the implementation of the DLM assessment system. While there was no formal, comprehensive training program for this audience in 2014-15, the staff had opportunities to participate in training designed for local education agency staff and test administrators. Throughout the year, they also received instruction during regularly scheduled meetings and through written documentation from the 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.

X.2. TRAINING FOR LOCAL EDUCATION AGENCY STAFF

In 2014-15, there were three roles that supported implementation of the assessment 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 oversaw 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 Steward managed educator, student, and roster data.
- The Technical Liaison verified that the network and testing devices were prepared for test administration.

Webinars were held in fall 2014 for each of these roles. The purpose of each webinar was to introduce staff to their roles, responsibilities, and timelines with regard to the DLM assessments. The webinars were advertised in advance, and participants from all states were invited to attend. Each webinar was also recorded, and a link to the recording was available from the DLM website for those who could not attend the webinar when it was scheduled.

Webinars were also held prior to the opening of the spring 2015 assessment window. The audience included district and building staff who were responsible for overseeing test administration. The purposes of these webinars was to provide reminders about the assessment administration process and describe strategies for monitoring assessment administration.

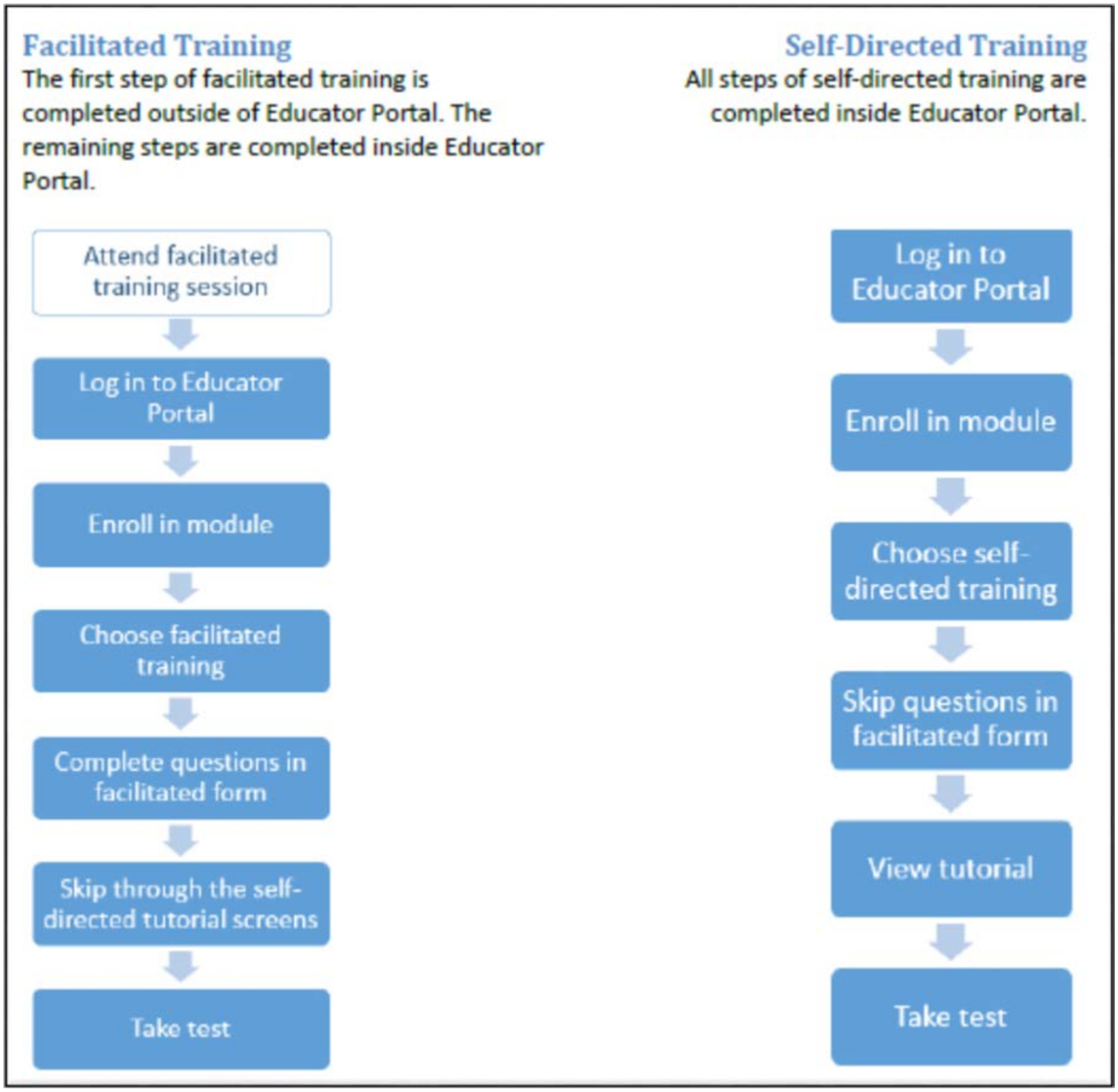
X.3. Required Training for Test Administrators

Training is required annually for educators who serve as test administrators and administer the DLM alternate assessments. In 2014-15, training was available in two formats: facilitated training (in-person training with quizzes in the Educator Portal) and self-directed (all content and quizzes within the Educator Portal).

In 2014-2015, training was required for all test administrators. The 2014-2015 required test administrator training system was built in the Educator Portal. All training materials were available through the online system managed by the DLM staff. The system provided a secure log-in. Users had access to a toll-free telephone and email helpdesk. Training materials were available as both PDF documents and videos. In-person training was also provided outside the online system using specially designed facilitated versions of the online training.

All test administrators had to successfully complete all modules before beginning testing; they weren't allowed access their students' log-in information for the student Kansas Interactive Testing Engine (KITE) platform until their training was successfully completed. Test administrators were required to complete seven modules and pass all seven post-tests with a score of 80% or higher. Test administrators were able to retake post-tests as many times as needed in order to pass all parts of the training.

Educators in each state had access to both facilitated and self-directed training options. Participants chose the correct version according to their state's guidelines. Figure 61 illustrates the differences between the two training formats. Modules were completed in the order of presentation, with a total training time for new test administrators estimated at approximately five hours, including videos and time to independently complete quizzes.



Facilitated Training

The first step of facilitated training is completed outside of Educator Portal. The remaining steps are completed inside Educator Portal.

Self-Directed Training

All steps of self-directed training are completed inside Educator Portal.

Attend facilitated training session
Log in to Educator Portal
Enroll in module
Choose facilitated training
Complete questions in facilitated form
Skip through the self-directed tutorial screens
Take test
Log in to Educator Portal
Enroll in module
Choose self-directed training
Skip questions in facilitated form
View tutorial
Take test

Figure 61. Required Training Processes Flows for Facilitated and Self-Directed Training.

X.3.A. Facilitated Training

The facilitated modules 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 determined its own policy guidance regarding who served as facilitators. Examples of individuals who served as facilitators included 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 other supports required to facilitate a meaningful, face-to-face training. Facilitators show the DLM-produced videos and implement learning activities as described in the facilitator guide. 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 scripts. Appendix G.1 includes the complete set of training materials for all seven required test administrator modules used in 2014-15.

In 2014-2015, the required test administrator training was offered through facilitated sessions in some states and local education agencies. Facilitators for these sessions prepared for the training by reviewing all videos and all sections of the *Test Administration Manual* (Dynamic Learning Maps, 2014a) addressed in the training. States also recommended that facilitators complete the training requirements themselves; facilitators who were also test administrators were required to pass the post-tests. Facilitators were asked to ensure that participants had Educator Portal accounts and access to them prior to the facilitated training session. Their responsibilities included setting up the training area with equipment, delivering the facilitated training modules, and directing users to return all equipment. Finally, facilitators directed test administrators to take each module quiz in the Educator Portal with support from the *Guide to DLM Required Test Administrator Training* (Dynamic Learning Maps, 2014b) for detail and access procedures. Facilitated training was flexibly structured so quizzes could be taken onsite during training sessions (e.g., in a computer lab) or independently after the training session was complete.

X.3.B. Self-Directed Training

The self-directed modules were designed to meet the needs of educators in rural and remote areas who were unable to attend facilitated sessions and those who otherwise needed access to on-demand training. 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 are learning. The videos are identical to those used in facilitated training. Each module ends with a quiz.

In 2014-15, the self-directed training was completed entirely within the DLM Educator Portal with support from the *Guide to DLM Required Test Administrator Training* (Dynamic Learning Maps, 2014b) for detail and access procedures, including the review of all module slides and procedures for completing all quizzes.

X.3.C. Training Content

X.3.C.i. Module 1: Overview of the Dynamic Learning Maps Alternate Assessment System

Module 1 of the test administrator training provided an overview of the DLM system components. Topics included illustration and discussion of the DLM maps, Claims and Conceptual Areas, Essential Elements, testlets, and linkage levels. Participants were expected to demonstrate an understanding of the DLM maps, including the academic nature of the knowledge, skills, and abilities described within them. They were also expected to develop a working definition of the Essential Elements and differentiate them from functional skills. Participants were also expected to define Claims and place them within the context of instructional practice.

Module 1 explained how the DLM testlets were developed. It also emphasized the fact that Target Level testlets are aligned directly to the Essential Element being tested, while explaining that testlets at other linkage levels are developed using the DLM map nodes that build up to, and extend from, the target node(s). In addition, participants were taught about the dynamic nature of the assessment, explaining that students could potentially see all five levels of testlets (Initial Precursor, Distal Precursor, Proximal Precursor, Target, and Successor) in their assessment, whether ELA or mathematics. They were introduced to mini-maps that specifically detail the nodes that are assessed at each linkage level.

X.3.C.ii. Module 2: Test Security in the Dynamic Learning Maps Alternate Assessment

In Module 2, participants were expected to know all the DLM security standards. These standards apply to anyone working with the DLM assessment. The standards are meant to ensure that assessment content is not compromised, and they include not reproducing or storing testlets, not sharing testlets via email, social media or file sharing, and not reproducing testlets by any means, except in clearly specified situations (e.g., braille forms of the testlets).

Participants agreed to uphold the DLM security expectations by signing an annual agreement document and committing to integrity. In addition, participants were instructed to follow their own state’s additional policies that govern test security.

X.3.C.iii. Module 3: Accessibility for All Students

Module 3 of the required training focused on accessibility. Participants were shown the characteristics of the DLM system that were designed to be optimally accessible to diverse learners, as well as the six steps for customizing supports for specific student needs, as described in detail in the DLM ACCESSIBILITY MANUAL.

The training emphasized how Universal Design for Learning was used to ensure that test content was optimally accessible. The technology platform used to deliver assessments, the KITE Client, was introduced, along with explanation of its accessibility features, including guidelines for selecting features for the Personal Needs and Preferences Profile (PNP).

Participants were expected to demonstrate understanding of test accommodations, their purpose, student eligibility, and appropriate practice. In addition, participants were shown how to complete the PNP and how the PNP and First Contact survey responses combined to develop a personal learning profile to guide administration decisions for each student.

Module 3 demonstrated how to actualize all accessibility features for an individual student, both within the KITE Client and through external supports, in conjunction with Testlet Information Pages (TIP).

Module 3 addressed flexibility in the ways that students access the items and materials, including what is considered appropriate flexibility (e.g., test administrator adapts the physical arrangement of the response options) and what is not (e.g., test administrator reduces the number of response options).

Finally, participants were taught how accessibility supports must be consistent with those that students receive in routine instruction and how those supports may extend beyond testing accommodations that are specifically mentioned in the child's IEP.

X.3.C.iv. Module 4: How the Assessment Works

Module 4 focused on participants' understanding and delivery of content through testlets within the KITE system. Topics included assessment content; types of assessments; design of the assessment, including testlet structure; item types; how to complete testlets; and all standard and allowable test administration practices. Finally, the module showed how students' responses led to test results used for accountability purposes.

Participants were expected to understand the two primary parts of all testlets: engagement activities and actual items. Participants learned about the design features of the engagement activities versus the tested items, as well as key aspects of test directions. In addition, they learned how testlets are administered using the KITE platform, clarifying the role of the test administrator and the role of the student in the computer-administered testlets. Participants learned about optional instructionally embedded assessments and how the blueprint sampled content during the spring window.

X.3.C.v. Module 5: Preparing for the Test

Module 5 prepared participants in their role as test administrators. They learned to check data, complete the First Contact survey, use the practice activities and release testlets, and plan and schedule assessment administration.

Participants reviewed the test administrators' role in completing data management requirements in the Educator Portal, supported by full instructions in the *Test Administration Manual* (Dynamic Learning Maps, 2014a). Participants reviewed the DLM assessment components, which are accessed through the Educator Portal (e.g., First Contact survey), and where student information is entered. Participants learned about students' required activities during operational testing as opposed to opportunities to practice through released testlets or practice activities available in KITE Client.

The training specifically addressed the First Contact survey, which is completed before testing begins. It uses test administrator responses to questions about student communication and academic skills to determine which linkage level is best to start students at the first time they encounter the DLM assessments. The First Contact survey is completed online, but test administrators also have access to all the questions in advance in an appendix to the *Test Administration Manual.* The First Contact survey includes questions regarding special education services and primary disability categorizations as well as sensory and motor capabilities, communication abilities, academic skill, attention and computer access.

X.3.C.vi. Module 6: Computer-Delivered Testlets

The sixth module provided participants with focused information on how the assessments are delivered via computer. Contents included the testlet structures used in the assessment system, the various item types used (e.g., single-select multiple choice, matching, sorting, drag and drop), how to navigate and complete testlets, and what to do on test day. Also included were details on the standard administration processes, allowable practices, and practices to be avoided.

X.3.C.vii. Module 7: Teacher-Administered Testlets

The final module focused on educator-administered testlets, including the specific structures used and the processes for completing testlets by administering them outside the KITE Client. The module also covered how the test administrator entered responses in to the KITE Client. The training emphasized the importance of educator directions provided within the testlet and specific directions to each content area (i.e., reading, mathematics, and writing). This module also included details on how to prepare for test day, including retrieval and use of Testlet Information Pages (TIPs), space arrangements, standard administration processes, allowable practices, and practices to be avoided.

X.3.D. Completion of All Modules

Each of the seven required training modules included a post-test. Participants were required to complete each post-test with at least 80% accuracy, and they had to score at least an 80% to be able to move on to the next module. Participants were allowed to retake the post-test as many times as necessary to achieve a passing score. They had access to the module contents at all times, and the recommendation was that they review module contents before attempting the post-test again. Whether participants completed the self-directed or facilitated version of the required training, post-tests were completed independently in the Educator Portal. Individuals with an appropriate role in Educator Portal were able to generate reports that summarized test administrators' completion of modules. Details of participants' progress were made available to state educational agencies.

X.4. Instructional Professional Development

The DLM professional development system was built to support educators in the efforts to teach English language arts and mathematics. The modules also teach educators about the DLM system. While the modules were originally intended for educators working directly with students with the most significant cognitive disabilities, demographic information suggests that pre-service educators, related service providers, parents, and others also accessed and completed the modules.

The professional development system was built in WordPress, an open-source website content management system. The professional development modules and instructional support materials are available through this site for anyone's use. In addition, there is a virtual community of practice that requires users to register and log-in with each use.

The instructional professional development system is accessed through a separate website at http://dlmpd.com. The system includes 50 modules available in both self-directed and facilitated formats. These modules address instruction in English language arts and mathematics and support educators in creating Individual Education Programs that are aligned with the DLM Essential Elements. This system also supports the communication needs of the students they teach and helps them understand the components of the DLM assessment system more completely. Appendix G.2 and Table 106 include a list of the 50 modules.

Certificates are provided upon course completion and can be emailed directly to facilitators. 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.

In addition to the 50 modules, the instructional professional development site provides instructional resources for educators. These resources include sample lesson plans, instructional vignettes, augmentative and alternative communication supports, and texts educators may choose to use in their day-to-day instruction. The number and range of resources is expanding. In addition, when educators register for the virtual practice community, they can upload and share instructional materials through an instructional materials exchange.

The final component of the instructional professional development system is an interactive system of groups, blogs, and discussion boards that allow educators across the DLM Consortium to pose questions and interact with one another.

X.4.A. Professional Development Participation and Evaluation

A total of 78,319 modules were completed in the self-directed format between the fall of 2012, when the first module was launched, and September 30, 2015 (Table 105). Data is not available regarding the number of educators who have completed the modules in their facilitated format, but it is known that several states (e.g., Iowa, Missouri, West Virginia) use the facilitated modules extensively.

Table 105. Number of Self-Directed Modules Completed by Educators in DLM States and Other Localities, through September 2015

State	Total Self-Directed Modules Completed
Missouri	20,456
Mississippi	13,804
Kansas	13,772
New Jersey	8,330
Colorado	3,128
Wisconsin	2,745
Utah	2,192
Illinois	1,930
North Carolina	1,620
Oklahoma	1,509
Vermont	1,105
Iowa	647
New Hampshire	599
Alaska	594
Pennsylvania	524
North Dakota	438
West Virginia	129
New York	80
Non-DLM states and other locations	18,521
Total	**78,319**

To evaluate educator perceptions of the utility and applicability of the modules, the DLM staff asked educators to respond to a series of evaluation questions upon completion of each self-directed module. Through September 2015, on average, educators completed the evaluation questions 75% of the time. Response rates ranged from 20% to 94%. The responses are summarized in Table 106.

Table 106. Results of Educator Responses to Instructional Professional Development Module Evaluation Questions

Module Name	Total # Modules Completed	Response Rate	The module addressed content that is important for professionals working with students with significant cognitive disabilities.	The module presented me with new ideas to improve my work with students with significant cognitive disabilities.	Completing this module was worth my time and effort.	I intend to apply what I learned in the module to my professional practice.
0: Who are Students with Significant Cognitive Disabilities?	10046	0.33	3.42	3.07	3.22	3.76
1: Common Core Overview	5614	0.30	3.12	2.87	3.04	3.66
2: Dynamic Learning Maps Essential Elements	8618	0.36	3.31	3.20	3.17	3.74
3: Universal Design for Learning	5116	0.34	3.31	3.23	3.23	3.75
4: Principles of Instruction in English Language Arts	4550	0.41	3.29	3.20	3.20	3.76
5: Standards of Mathematics Practice	7027	0.20	3.24	3.19	3.20	3.71
6: Counting and Cardinality	3274	0.43	3.34	3.28	3.28	3.76
7: IEPs Linked to the DLM Essential Elements	4119	0.37	3.28	3.20	3.21	3.73
8: Symbols	3241	0.26	3.36	3.30	3.31	3.74
9: Shared Reading	3964	0.47	3.39	3.32	3.28	3.80
10: DLM Claims and Conceptual Areas	2274	0.65	3.23	3.08	3.11	3.66
11: Speaking and Listening	2492	0.46	3.31	3.23	3.22	3.73
12: Writing: Text Types and Purposes	2494	0.60	3.22	3.15	3.10	3.68
13: Writing: Production and Distribution	1256	0.92	3.25	3.20	3.18	3.70

Module Name	Total # Modules Completed	Response Rate	The module addressed content that is important for professionals working with students with significant cognitive disabilities.	The module presented me with new ideas to improve my work with students with significant cognitive disabilities.	Completing this module was worth my time and effort.	I intend to apply what I learned in the module to my professional practice.
14: Writing: Research and Range of Writing	1502	0.70	3.22	3.17	3.15	3.70
15: The Power of Ten-Frames	927	0.92	3.23	3.22	3.18	3.65
16: Writing with Alternate Pencils	1080	0.92	3.34	3.28	3.25	3.65
17: DLM™ Core Vocabulary and Communication	1153	0.93	3.39	3.34	3.36	3.74
18: Unitizing	747	0.88	3.18	3.14	3.13	3.62
19: Forms of Number	828	0.87	3.14	3.09	3.08	3.58
20: Units and Operations	685	0.90	3.12	3.09	3.06	3.57
21: Place Value	716	0.88	3.13	3.10	3.07	3.53
22: Fraction Concepts and Models Part I	568	0.89	3.12	3.09	3.07	3.53
23: Fraction Concepts and Models Part II	472	0.89	3.12	3.09	3.08	3.55
24: Composing, Decomposing, and Comparing Numbers	554	0.84	3.15	3.14	3.11	3.55
25: Basic Geometric Shapes and Their Attributes	519	0.88	3.15	3.11	3.06	3.57
26: Writing Information and Explanation Texts	481	0.92	3.16	3.16	3.15	3.63
27: Calculating Accurately with Addition	294	0.91	3.09	3.08	3.01	3.52
28: Measuring and Comparing Lengths	235	0.91	3.07	3.05	2.98	3.48
29: Emergent Writing	595	0.93	3.31	3.27	3.28	3.71

Module Name	Total # Modules Completed	Response Rate	The module addressed content that is important for professionals working with students with significant cognitive disabilities.	The module presented me with new ideas to improve my work with students with significant cognitive disabilities.	Completing this module was worth my time and effort.	I intend to apply what I learned in the module to my professional practice.
30: Predictable Chart Writing	313	0.94	3.33	3.29	3.31	3.72
31: Calculating Accurately with Subtraction	182	0.91	3.13	3.12	3.08	3.54
32: Teaching Text Comprehension: Anchor-Read-Apply	230	0.91	3.31	3.26	3.27	3.69
33: Generating Purposes for Reading	226	0.84	3.22	3.22	3.22	3.65
34: Exponents and Probability	154	0.88	3.08	3.10	3.08	3.46
35: Beginning Communicators	390	0.93	3.41	3.27	3.34	3.77
36: Time and Money	208	0.92	3.25	3.21	3.20	3.67
37: DR-TA and Other Text Comprehension Approaches	165	0.86	3.28	3.24	3.26	3.66
38: Supporting Participation in Discussions	158	0.89	3.26	3.23	3.23	3.66
39: Algebraic Thinking	173	0.93	3.17	3.14	3.09	3.52
40: Composing and Decomposing Shapes and Areas	160	0.89	3.14	3.12	3.10	3.48
41: Writing: Getting Started with Writing Arguments	110	0.89	3.05	3.09	3.05	3.48
42: Calculating Accurately with Multiplication	85	0.86	3.23	3.06	2.99	3.44
43: Perimeter, Volume, and Mass	84	0.89	3.02	3.05	2.98	3.41
44: Writing: Getting Started in Narrative Writing	52	0.92	3.15	3.16	3.05	3.48
45: Patterns and Sequence	68	0.91	2.91	2.89	2.88	3.34

Module Name	Total # Modules Completed	Response Rate	The module addressed content that is important for professionals working with students with significant cognitive disabilities.	The module presented me with new ideas to improve my work with students with significant cognitive disabilities.	Completing this module was worth my time and effort.	I intend to apply what I learned in the module to my professional practice.
46: Functions and Rates	47	0.85	2.89	2.93	2.84	3.24
47: Calculating Accurately with Division	35	0.80	3.21	3.26	3.21	3.48
48: Organizing and Using Data to Answer Questions	12	0.58	3.38	3.38	3.33	3.50
49: Strategies and Formats for Presenting Ideas	26	0.85	3.23	3.30	3.31	3.56
50: Properties of Lines and Angles [41]						
Total	78319					
Average		0.75	3.21	3.17	3.15	3.61

[41] Module launched in late September 2015. No evaluation results were available as of 9/30/15.

Across modules, the average responses to survey items tended to be positive. For example, responses to the question, “the module addressed content that is important for professionals working with students with significant cognitive disabilities” averaged 3.21 on a scale from 1 (strongly disagree) to 4 (strongly agree). Respondents also perceived the modules as worth their time and tended to indicate an intent to apply what they had learned. Especially where response rates were high, these ratings provide evidence of overall quality and relevance of the modules.

XI. CONCLUSION AND DISCUSSION

The Dynamic Learning Maps Alternate Assessment System is based on the core belief that all students should have access to challenging, grade-level academic content. Therefore, the DLM assessments provide students with the most significant cognitive disabilities the opportunity to demonstrate what they know and can do.

The DLM System completed its first operational administration year in 2014-2015. This technical manual provides evidence to support the propositions and assumptions that undergird the assessment system as described at the onset of its design in the DLM Theory of Action (Chapter I, Figure 2). The contents of this manual address the information summarized in Table 107.

Table 107. Review of Technical Manual Contents.

Chapter(s)	Contents
I, II	Reviews the foundations of the assessment system, including the development of the theory of action to guide each subsequent step and the learning map, the DLM learning and cognition model.
III, IV, X	Provides procedural evidence of test content development and administration, including alignment to the learning maps and college and career readiness standards, accessibility features and procedures, security protocols, and test administrator training.
V	Describes the statistical model used to produce scores based on student responses.
VI	Provides a description of how cut points were developed to interpret results via performance levels.
VII, VIII	Describes results and analysis of the first operational administration’s data, evaluating how students performed on the assessment, the distributions of those scores, aggregated and disaggregated results, and analysis of the internal consistency of student responses.
IX	Provides additional studies focused on specific topics related to validity and in support of the score propositions and purposes.

This chapter reviews the evidence provided in this technical manual and places it within a validity framework in order to assess the program’s overall success at producing scores that mean what they are intended to mean. In addition, future research studies are discussed as part of ongoing and iterative processes of program responsiveness, validation, and evaluation.

XI.1. Validity Framework

The *Standards for Educational and Psychological Testing* (AERA et al., 2014) are the professional standards used broadly to evaluate educational assessments; the DLM Alternate Assessment System is no exception. The *Standards* define validity as “the degree to which evidence and theory support the interpretations of test scores for proposed uses of the test” (p. 11) and assert that validity is the “most fundamental consideration in developing tests and evaluating tests” (p. 11). Using the *Standards* as a baseline for the evaluation of the DLM assessments, this manual’s primary purpose is to provide evidence and theory to support the propositions laid out in the DLM Theory of Action (see Chapter I). The four propositions serve as an organizing framework for the summary and evaluation of validity evidence in this chapter. To this end, Chapter XI looks back at the previously presented evidence in support of the score purposes and their proposed interpretations and uses.

All aspects of the validity argument must be carefully evaluated (Lissitz, 2009; Sireci, 2009). The purpose of the assessment with its resultant scores is critical to the overall validity argument as it is indicative of the model from which the assessment was originally designed (Mislevy, 2009). It follows, then, that the evidence collected throughout the entire development process should point to a clear and persuasive link between the original assessment purpose and the uses and interpretations of the results. Clarity between what can be observed (e.g., student responses to assessment tasks) and what must be inferred (e.g., student ability in the content area) must inform the validity and interpretative arguments (Kane, 2006). In addition, the dimensions and organization of the overall validity argument matter, as they include not only the content sampled and procedural bases of the assessment, but also evidence for the underlying construct to be assessed, what may be included on the assessment that is irrelevant to the construct, and the relative importance of the consequences of the resulting scores (Messick, 1989; Linn, 2009).

Validation is the process of evaluating the evidence and theory presented in the overall validity argument. Using the *Standards* as our foundation, the DLM System began the validation process “with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use” (AERA et al., 2014, p. 11). These propositions[42] then informed the development of the theory of action (as described in Chapter I, Figure 2), which focused overall on combining high expectations for students with the most significant cognitive disabilities with appropriate educational supports for teachers, to result in improved academic experiences and outcomes for students.

XI.2. Propositions for Score Interpretation and Use

The DLM Consortium developed an argument-based approach to validity that established four propositions to support the intended uses and interpretations of DLM scores. These propositions are laid out within a context of precursors, assessment design assumptions, and

[42] The term “proposition” 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 III).

ultimate goals for the program within the theory of action (Chapter I, Figure 2). The propositions relate directly to the ultimate program goals and specific score purposes, providing the framework within which validity evidence can be judged. The four propositions are as follow:

1. Scores represent what students know and can do.
2. Achievement level descriptors provide useful information about student achievement.
3. Inferences regarding student achievement, progress, and growth can be drawn at the conceptual area level.
4. Assessment scores provide useful information to guide instructional decisions.

Summative scores from the DLM assessments are intended for use for several purposes:

1. Reporting achievement and growth within the taught content aligned to grade-level content standards to a variety of audiences including educators and parents
2. Inclusion in state accountability models to evaluate school and district performance
3. Planning instructional priorities and program improvements for the following school year

Appropriate interpretations and uses of DLM scores support the overall goals of the DLM Alternate Assessment System:

1. Students with the most significant cognitive disabilities are able to demonstrate what they know and can do.
2. Teachers make sound instructional decisions based on data.
3. Parents, teachers, and students have high expectations for students' academic achievement.
4. The trajectory of student growth in academic knowledge and skills improves.

Holding high expectations for students with the most significant cognitive disabilities and providing appropriate educational supports for teachers will lead to improved academic experiences and outcomes for students.

XI.3. SUMMARY AND EVALUATION OF VALIDITY EVIDENCE

To build the validity argument, the examination of the proposed score interpretations and purposes necessarily points back to evidence previously presented in this technical manual. This validation review was conducted by examining evidence associated with each proposition, organized by categories of evidence as presented in the *Standards* (AERA et al., 2014). These categories are (a) test content, (b) response processes, (c) internal structure, (d) other variables, and (e) consequences of testing.

Within each category, we describe related evidence. Although some evidence supports more than one proposition, for the sake of conciseness it is only described with one proposition. Table 108 in the Evaluation Summary section of this chapter summarizes the sources of validity evidence as organized by the propositions and each evidence category.

XI.3.A. PROPOSITION 1: SCORES REPRESENT WHAT STUDENTS KNOW AND CAN DO

XI.3.A.i. Evidence Based on Content

Evidence based on test content relates to the evidence "obtained from an analysis of the relationship between the content of the test and the construct it is intended to measure" (AERA et al., 2014, p. 14). The DLM Assessment System is intended to support claims about what students know and can do in English language arts and mathematics.

The interpretation and use of DLM scores depends on evidence of the relationships among the content components of the assessment system. Starting with the learning map models, assumptions related to test content focus on whether the DLM maps themselves address the content domains with fidelity. The Essential Elements, grade-level expectations for students with the most significant cognitive disabilities, must be adequately linked to the college and career readiness standards, in this case the Common Core State Standards. Coverage of content, as specified by test blueprints, provides evidence of representation of the target domain overall. Additionally, Essential Elements must be accurately linked to nodes within the DLM maps. Groups of ordered nodes related to the Essential Element – linkage levels – are identified for assessment. Thus, items within testlets are aligned to the Essential Elements via the map nodes at the associated linkage levels. Finally, teachers must have instructed the student on the content prior to assessment in order for students to have had the necessary opportunity to learn.

Content-related evidence to support this proposition is described primarily in terms of the goal of alignment. Alignment is "at the heart of the process" of content-oriented evidence of validation and involves evaluating the degree to which test content corresponds to student learning standards (AERA et al., 2014, p.15), which are the Essential Elements in the DLM system. Alignment was considered across the design, development, and operational stages. A second source of content-related evidence in the development phase was the use of procedures to ensure that items and testlets maximize construct-relevant and minimize construct-irrelevant features.

XI.3.A.i.a Design Phase

Chapter II describes procedural evidence that supports the representation of the content domains of English language arts and mathematics. Through an iterative process and with expert and educator feedback, teams developed highly dimensional representations of the content consisting of map nodes and pathways to connect them.

While the DLM maps represent the architecture of the content domain, Essential Elements convey the grade-level expectations for students with the most significant cognitive disabilities. As described in Chapter III, the Essential Elements were carefully developed to align to college and career readiness standards in each grade, representing high expectations for students so they would be prepared for college, career, and citizenship. Chapter III also explains how the Essential Elements were aligned with the DLM maps and grouped into claims and conceptual areas. The development of the test blueprint demonstrates how content was sampled to cover the content domain with coverage defined by the conceptual area.

XI.3.A.i.b Development Phase

Using a variant of Evidence-Centered Design (ECD), the consortium developed Essential Element Concept Maps (EECM) to support assessment development. As described in Chapter III, EECMs are graphic organizers for each Essential Element that define ELA and mathematics content specifications for assessment. They link the Essential Elements (content standards) to the test content itself, including descriptions of the nodes at each linkage level, key vocabulary, misconceptions and definitions, prerequisite and requisite skills, and accessibility requirements.

Testlet development procedures (Chapter III) followed guidance in the *Standards* (AERA et al., 2014). Item writers were recruited from multiple states in the consortium and were selected based on their qualifications in academic content areas and/or experience teaching students with the most significant cognitive disabilities. Item writers received comprehensive training and had opportunities for guided practice and feedback throughout the item writing session. Training focused on accessibility, universal design for learning, content development, and bias and sensitivity. The DLM testlets were designed to be accessible to all students in the target population, starting from the first delivered testlets. Item writers were taught to use DLM core vocabulary to minimize unnecessary barriers to student demonstrations of conceptual understanding that might be introduced by using excessively complex vocabulary in items. The vast majority of item writers evaluated the process and their products positively.

Testlets were reviewed (see Chapter III) for content, accessibility, instructional relevance, and bias and sensitivity at multiple points before field testing. Internal reviews for content and accessibility preceded external reviews by educators from across the consortium. The DLM test development staff considered feedback from all panelists when deciding whether to reject items or revise them before field testing. External reviews looked at item-level content criteria (alignment, depth of knowledge, quality and appropriateness, accuracy), accessibility (instructional relevance, clarity and appropriateness of images and graphics, minimizing barriers to students with specific needs), and bias/sensitivity (identifying items that require prior knowledge outside the bounds of the targeted content, ensuring fair representation of diversity, avoiding stereotypes and negative naming, removing language that affects a student's demonstration of their knowledge on the measurement target, and removing any language that is likely to cause strong emotional response). Across grades, subjects, and pools, the percentage of items or testlets rated as "accept" ranged from 72% to 91% in ELA and from 76% to 88% in math. The rate at which content was recommended for rejection was 5% or less across grades, pools, and rounds of review.

The final step of the development phase – field testing – provided additional content-related evidence (Chapter III). DLM staff used item flagging rules that allowed them to check for the reasonableness of the fungibility assumption that would later be applied in the diagnostic classification model used for scoring (Chapter V). In ELA and mathematics, a total of 515 items (12.2% of total) were flagged in Field Test 1 through Field Test 3, and 1,876 items (17.8%) were flagged during Phases A through C as needing review by content teams. The procedural evidence presented about the construction of the DLM maps and assessments provides strong evidence of alignment between the definition of the constructs as represented in the maps and the content of the testlets developed using principles of universal design for learning and evidence centered design.

XI.3.A.i.c Operational Phase

Chapter IX provides the results of an external alignment study. Overall, the external alignment study provided strong evidence of relationships among the content structures within the DLM assessment system: College and Career Ready standards to Essential Elements, Essential Elements to Target nodes, vertical progressions of nodes at linkage levels associated with each Essential Element, and item-node relationships. The study indicates that students with the most significant cognitive disabilities have access to challenging academic content at each grade level, with fidelity to content and performance centrality in the associated map nodes. Areas for improvement include re-evaluating which College and Career Ready standards are the best match to some Essential Elements that were evaluated as mismatched to the identified standard and reviewing items where panelist- and item writer-identified cognitive process dimension ratings differed. A full written response to the alignment study findings is provided in Appendix H.1, DLM Response to Alignment Study. Since the external alignment study was delimited to samples of testlets, additional evidence will also be needed to evaluate alignment of the assessment as it was administered (i.e., the student-level experience with a series of testlets).

XI.3.A.i.d Curriculum Alignment

Implicit in the intended uses of the DLM results is that the outcomes reflect content the student has had an opportunity to learn. Evidence that students have received instruction in the grade-level Essential Elements supports the use of results for accountability and school evaluation purposes. One form of procedural evidence is the coherent professional development system that supports instruction (see Chapter X). Modules support teachers in knowing how to teach related content within each conceptual area. Nearly 80,000 self-directed versions of the modules were successfully completed by the end of 2014-15 and post-training survey responses were positive about the importance of the content and likelihood of applying the information to their professional practice.

Preliminary evidence of students' opportunity to learn the assessed content came from spring 2015 surveys in which teachers estimated the number of testlets that had content that matched what the students experienced during instruction (see Chapter IX). Responses were distributed across the full continuum (i.e., 0 to 7 testlets). Respondents indicated that more than half of testlets (i.e., 4 or more) had matching content for 45% of students in ELA and 36% of students in math. While these figures may reflect differences between how content is instructed offline versus how it is assessed in online DLM assessments, responses provide weak evidence of curriculum alignment in this first year of operational assessment.

XI.3.A.ii. Evidence Based on Response Process

The interpretation and use of DLM scores depends in part on the validation of whether the cognitive processes that students are engaged in when taking the test match the claims made about the test construct. Evidence is needed to analyze the response processes of test takers in order to determine the fit between the test construct and how students actually experience test content (AERA et al., 2014). Both theoretical and empirical evidence is appropriate and should come from the individual test taker and external observation. Given the cognitive and communication challenges of students with the most significant cognitive disabilities, this category includes procedural evidence as well as empirical evidence that relies on direct observation, teacher feedback, and, to a lesser extent, student verbalization.

XI.3.A.ii.a Assessment Design and Development

Along with procedures and evidence described earlier regarding test content, several aspects of the assessment development process were intended to minimize response barriers and promote construct-relevant interactions with items. For example, as described in Chapter III, the item writing process began with assignment of an Essential Element and EECM, and featured training and practice activities that included discussion of how a student might demonstrate the knowledge, skills, or understanding in the nodes included on the EECM. Similarly, item writers were provided with guidance and feedback during the item writing process to promote the production of testlets accessible to the largest number of students possible. Strategies to maximize accessibility of the assessment content and avoid barriers to meaningful student interaction with items included using the DLM core vocabulary, avoiding terminology that could advantage or disadvantage particular students, and consideration of issues that could cause potential barriers for students at every step of the item writing process. Item writers and external reviewers were from diverse backgrounds and different states within the consortium. Having diverse perspectives represented by external reviewers minimized the chance that students would be disadvantaged due to the inclusion of unnecessary regional or cultural content in testlets. External review panelists evaluated items and testlets for accessibility of graphics, clear use of language that minimized the need for inference or prior knowledge, and instructional relevance for students. Additionally reviewers were asked to judge testlets to be reasonably free of barriers for students with limited working memory, communication disorders and/or limited implicit understanding of the intentions and emotions of others. The application of these criteria supported the development of content designed to allow all students to interact meaningfully with the assessments.

XI.3.A.ii.b Interaction with Testlet Content

To support assertions that knowledge and skills demonstrated on an assessment reflect students' true abilities, assessment items must "elicit cognitive processes associated with the underlying cognitive model so that observed item responses can lead to valid inferences about the construct under investigation" (Ketterlin-Geller, 2008, p. 10). As described in Chapter IX, cognitive labs provided evidence that test administrators and students interact as intended with the assessments and that there are not barriers to the intended response process due to construct-irrelevant testlet features or item response demands. Student labs identified response demands of various item types used in computer-administered testlets and evaluated the extent to which students had difficulty with those demands. Test administrator labs were used to evaluate clarity of written instructions in teacher-administered testlets and the degree to which teachers were able to correctly identify and record student behaviors when students used various response modes. While the use of test administrator labs is only in the first phase of data collection at the time of publication for this manual, preliminary evidence on interpretation of student behaviors indicates that the ease of determining student intent depends in part on the student's response mode. This method of data collection will continue, particularly as refinements are made to improve educator directions and supports for test administrators.

XI.3.A.ii.c Fidelity of Administration

The DLM assessments are intended to be administered with as much standardization as possible, and with the expectation that test administrators maintain fidelity to the important aspects of the administration process where flexibility is needed. This balance of standardization and flexibility is necessary given the heterogeneity of students with the most significant cognitive disabilities. General guidance is provided on these practices through multiple manuals and required test administrator training (see Chapters IV and X). Testlet Information Pages (TIPs; see Chapter IV) support teachers' readiness to deliver specific testlets to specific students with integrity. The majority of respondents to a spring 2015 survey indicated they had confidence in their ability to deliver computer-administered and teacher-administered testlets (Chapter IV). They also evaluated KITE Client as easy to use to administer testlets.

Test administration observations (Chapter IX) were conducted to further understand response processes for students. Observations were designed to understand whether students were able to interact with the system as intended and to respond to items irrespective of a sensory, mobility, health, communication, or behavioral constraint. The observations provided information on student interaction with testlet contents (e.g., images, figures, engagement activities) and manipulatives where applicable. They also provided evidence of the teacher's actions during administration. Test administrations were observed for the full range of students eligible for DLM assessments, across multiple states and multiple testing windows. Results provided evidence that test administrators accurately captured student responses. Across all test administration observations and student response modes, test administrators recorded responses reliably in 93.3% of teacher-administered testlets observed. In 98.6% of cases where test administrators entered responses on behalf of students in computer-administered testlets, the entered response matched the student's response. This evidence supports the assumption that test administrators entered student responses with accuracy.

In limited cases during the spring 2015 administration, constancy was compromised by an interruption in the adaptive delivery algorithm (see Chapter IV). The impact of these incidents on score interpretations and inferences was mitigated in most cases by having students revert to the last correctly assigned testlet and resume testing. To support appropriate uses of results for impacted students, the state was provided an incident file (Chapter VII) to assist them in making decisions about how to treat those students' scores within the context of their accountability systems.

XI.3.A.ii.d Accessibility

Accessibility must be evaluated to identify evidence that the delivery of items and testlets are accessible and appropriate for the full range of students with the most significant cognitive disabilities. Student and test administrator interaction with the KITE system must be evaluated to see if the system provides the necessary supports. Procedures for determining each student's personal needs and executing the correct system features to meet those needs must be in place.

Test administrators recorded accessibility supports in the student's Personal Needs and Preferences (PNP) profile. To support test administrators in making appropriate decisions about those supports, accessibility was addressed through manuals (Chapter IV), required test administrator training (Chapter X) and additional resources, such as access to released testlets with several simulated students (Chapter IV). Test administration observations revealed that students were able to respond to task using multiple response modes including verbal, gesture and eye-gaze. Evidence in support of accessibility was collected by having observers note difficulty with accessibility supports during observations of teacher-administered testlets. Of the 30 observations of teacher-administered testlets, observers noted difficulty in two cases (6.7%).

Surveys of teachers at the end of 2014-15 test administration provided feedback related to assumptions about accessibility during the assessment process. More than three-fourths of teachers indicated they knew how to use accessibility supports and allowable practices (Chapter IV). Evidence of the effectiveness of these supports was mixed. While 83% agreed that students had access to all needed supports, 76% indicated the student responded to the best of his or her ability, and 62% agreed that the student was able to respond regardless of health, behavior, or disability concerns (Chapter IX).Also, fewer than three-fourths of survey respondents indicated the accessibility features were similar to those used during instruction. This pattern suggests some students still encounter barriers during the assessment process. It is not known whether those barriers are due to gaps between students' accessibility needs and existing supports in the DLM assessment system, whether students were assessed outside of optimal times (e.g., during behavioral difficulties), or due to other issues. It is likely that some of the barriers were related to the challenges in transitioning to a new online assessment system and learning about the available supports in a timely manner.

Where accessibility gaps are identified due to limited compatibility between types of assistive devices and the KITE system, technology enhancements will be scheduled to improve accessibility. The DLM Consortium has already partnered with the Assistive Technology Industry Association to collect input from manufacturers on compatibility of their devices with the KITE Client, and this partnership is expected to continue. More research will be necessary to determine whether students have more opportunities to use those features during instruction in the future, or whether differences may remain because of variations in delivery mode (i.e., instruction delivered directly by the test administrator versus the DLM assessments administered online).

XI.3.A.iii. Evidence Based on Internal Structure

Analyses to support evaluation of evidence based on internal structure indicate the degree to which "relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA et al., 2014, p. 16). In this category of evidence, the DLM maps provide multi-dimensional representations of content in the academic domains. Reliability analyses describe the consistency of measurement at the linkage level, Essential Element, and overall content area. Additionally, given the heterogeneous nature of the student population and the various and interrelated subgroup categories (e.g., communication mode), differential item functioning (DIF) analyses examine whether particular items function differently for specific subgroups.

XI.3.A.iii.a Learning Map Models and Statistical Modeling

The architecture of the DLM assessment system is the learning map models, which are networks of sequenced learning targets. Evidence to support the validity argument begins with learning map model development as it describes learning and knowledge acquisition (Chapter II). Through the use of the most current research and theoretical evidence, and with input from educators and accessibility experts, the map development process addressed the assumption that nodes were sequenced in the correct order of acquisition and relationships between nodes were appropriate. Empirical evaluation of these assumptions will take place once sufficient data are collected for node-based modeling of the DLM maps. This process is anticipated to begin in 2017.

Several other sources of evidence have been collected related to the structure of the content. For example, evidence from the pilot test (Chapter III) indicates that when students were assigned testlets from multiple linkage levels, the percentage of correct responses generally decreased from testlet 1 (lowest linkage level) to testlet 3 (highest linkage level). Also, the external alignment study (Chapter IX) included an evaluation of the relationships between content at different linkage levels associated with an EE.

Consistent with the assessment system design, diagnostic classification models are used for statistical modeling. Chapter V provides evidence for the appropriateness of the statistical model, given the learning map basis and the scoring approach used in the DLM system. In addition, evidence provided in Chapter V illustrates how linkage levels can describe mastery at appropriate levels of specificity and are distinct from one another.

The other organizing structure in the DLM assessment system is the grouping of Essential Elements into conceptual areas nested within claims (Chapter I). Essential Elements are aligned with nodes located at appropriate intervals within the DLM maps to reflect within-grade and across-grade relationships (Chapter II). Blueprints were constructed to allow inferences to be made at the Conceptual Area level as described within the context of the DLM maps (Chapter III).

XI.3.A.iii.b Reliability

"[T]he general notion of reliability/precision is defined in terms of consistency over replications of the testing procedure" (AERA et al., 2014, p. 35). Evidence of reliability must show "appropriate evidence of reliability/precision" (AERA et al., 2014, p. 42). Because the DLM Alternate Assessment System uses non-traditional psychometric models (diagnostic classification models) to produce student score reports, evidence for the reliability of scores is based on methods that are commensurate with the models used to produce score reports.

Reliability evidence for the DLM assessments must address the assumption of internal consistency, including decision consistency and accuracy. For the DLM assessments, reliability is provided at three levels:[43] (a) the number of linkage levels mastered within a content area; (b) the number of linkage levels mastered within each EE; and (c) the mastery status of each of the 1,275 linkage levels across all EEs. Reliability estimates are provided for three overall metrics: correct classification rate, classification kappa, and correlation between true and estimated values.

As described in Chapter VIII, the reliability summaries for the number of linkage levels mastered within an EE presented reasonable levels of reliability (45% of EEs with Pearson correlations ≥ .70). However, 84.4% of classification accuracy values were ≥ .80. Similarly, the reliability summaries for mastery classification status of each linkage level showed reasonable levels of reliability (68.7% of linkage levels with tetrachoric correlations ≥ .80). However, roughly one-third of linkage level kappa values for the year-end model fell below 0.6. The low linkage level kappa and EE correlation values may be due to students taking fewer items per EE. Overall, reliability measures for the DLM assessment system address the *Standards* (AERA et al., 2014), using methods that were consistent with assumptions of the diagnostic classification model. The analyses yielded evidence to support the argument for internal consistency of the program. Results also pointed to the need for students to take adequate numbers of items per EE.

XI.3.A.iii.c Evaluation of Item-Level Bias

Differential item functioning (DIF) addresses the broad problem created when some test items are "asked in such a way that certain groups of examinees who are knowledgeable about the intended concepts are prevented from showing what they know" (Camilli & Shepard, 1994, p. 1). Studies that use DIF analyses can uncover internal inconsistency if particular items are functioning differently and systematically for identifiable subgroups of students (AERA et al., 2014). While DIF does not always indicate a weakness in the test items, it can help point to

[43] Evidence for reliability of results in the content area is presented with proposition #2.

construct-irrelevant variance or unexpected multidimensionality, thereby contributing to an overall arguments for validity and fairness.

As described in Chapter IX, both uniform and a combined model analysis of gender DIF yielded flags for between 1 and 17% of items by grade level and content area, with few flagged items having moderate to large effect sizes. The existence of DIF does not necessarily indicate a flaw in the assessment; rather, results serve to inform future steps in the development cycle. For example, items flagged for DIF will be inspected and could be revised or eliminated by content developers. The limited existence of DIF in the current analysis provides additional evidence strong internal structure.

XI.3.A.iv. Evidence Based on Relationships to Other Variables

To date, evidence on the relationship between student responses on the DLM assessments and other measures is limited to teacher evaluations of student academic knowledge and skills as measured by the First Contact survey and teacher evaluations of testlet difficulty. Teacher ratings of students' academic skills on the First Contact survey are translated into complexity bands in order to assign the testlet linkage level. The pilot administration (Chapter III) provided evidence that student complexity band was associated with linkage level difficulty. Students were assigned to different complexity bands and received three testlets at multiple linkage levels. The percentage of correct responses at the item level was lowest for students assigned to the foundational complexity band, as expected, and increased as the complexity band increased to band 3.

Additionally, preliminary evidence of the relationship between student responses and other variables comes from test administrator ratings of testlet difficulty for specific students. While testlet difficulty is not a direct measure of academic content, judgment about difficulty was based on the complexity of the academic content in the testlet. As described in Chapter IX, field test results from spring 2014 indicate that across subjects and complexity bands, roughly two-thirds of teachers reported the testlets assigned were at about the right level of difficulty for the student. Testlets were more likely perceived to be too difficult for students assigned to the foundational complexity band.

XI.3.A.v. Evidence for Consequences of Assessment

Consequential evidence may be limited in the first year of an operational assessment system as the system has not yet had an opportunity to have an effect. As described in Chapter IX, spring 2015 survey responses describe teachers' baseline perceptions of the academic content in the DLM assessments. Perhaps not surprisingly, just half of respondents indicated the assessment content measured important academic skills, while nearly three-fourths indicated the content reflected high expectations for the student. The DLM assessments represent a departure from many of the states' previous alternate assessments in the breadth of academic skills assessed. The Essential Elements reflect challenging learning targets for students, while the alternate academic achievement standards set high expectations for achievement; fewer students reached the At Target and Advanced performance levels (see Chapter VII) than on the states' previous alternate assessments. Teacher responses may reflect the awareness that the DLM assessments contain challenging content, but that they are divided on its importance.

XI.3.B. PROPOSITION 2: ACHIEVEMENT LEVEL DESCRIPTORS PROVIDE USEFUL INFORMATION ABOUT STUDENT ACHIEVEMENT

The DLM approach to standard setting relied on mastery profiles to anchor panelists' content-based judgments to arrive at performance level cut points based on multiple rounds of range finding and pinpointing. Cut points were set to distinguish four performance levels describing student achievement. Grade and content-specific performance level descriptors (PLDs) were not used during the standard setting workshop. Instead, they emerged based on the final cut points and were completed after standard setting in 2015.

XI.3.B.i. Evidence Based on Content

Cut points for the four performance levels were determined during the standard setting workshop as described in Chapter VI. Well-qualified panelists fully engaged in a process by which they made use of mastery profiles that summarized linkage level mastery by EE to specify cuts for the total number of linkage levels a student must master to be classified in a performance level. Panelists also relied on content-based evidence when classifying profiles to performance levels, including node description booklets, example items and testlets, and assessment blueprints.

Following specification of cut points for the four performance levels, grade and content-specific performance level descriptors were created. Beginning at the standard setting workshop, and continuing with DLM staff content team development, the specific content being assessed at each linkage level was used to guide the development of grade and content-specific performance level descriptors.

Standard setting panelists began the process by drafting lists of skills and understandings that they determined were characteristic of specific performance levels, after establishing cut points. These skills were used as a starting point for the DLM content teams as they developed language for grade and content-specific descriptions for each performance level. Content teams reviewed the EEs, EECMs, and linkage level descriptors on the profiles to determine skills and understandings assessed at the grade level. Using multiple sources of information, all anchored in the EEs and the structure of the DLM maps, the content teams evaluated the placement of skills into each of the four performance levels. These sources of evidence provide support for the claim that achievement level descriptors provide useful information about student achievement, describing grade-level content expectations.

XI.3.B.ii. Evidence Based on Internal Structure

As presented in Chapter VIII, content-area (performance level) reliability indicates consistency of measurement for the content area as a whole. These statistics are analogous to total score reliability in assessments that use classical or IRT-based models. Reliability evidence was demonstrated by the correlation between true and estimated number of linkage levels mastered, which ranged from .837 to .950 These values indicate that measurement is generally consistent and reveal low measurement error in the total number of linkage levels a student is determined to have mastered, which translates to greater accuracy in assigning students to performance levels. As such, the descriptions of knowledge, skills, and ability typical of students in each performance level has a high likelihood of describing individual students classified to the particular performance level, increasing their utility for meaningful interpretative use by educators and parents.

XI.3.B.iii. Evidence for Consequences of Assessment

In order to establish sound score interpretations and delimit score use, score reports must be useful and provide relevant information for teachers to inform instructional choices and goal setting. Teachers must use results to plan subsequent instruction, and scores can only be interpreted and used for purposes called out in the theory of action as part of the validity argument.

Assessment results (Chapter VII) were provided to all DLM member states to be reported to parents and to educators at state and local education agencies. Individual reports were provided to teachers and parents. State users received a general research file, which included the student's overall performance level. Individual student score reports also included performance level and a summary of skills the student mastered, resulting in the assignment of the performance level. In addition, aggregated reports were provided to state and local education agencies summarizing student achievement by performance level (Chapter VII). Score reports for the 2015-2016 academic year will include the grade and content-specific performance level descriptors in place of the bulleted list of skills mastered by conceptual area.

Evidence of intended use of performance level information in score reports is summarized in the research to inform DLM score reports (Chapter IX). Teachers indicated they used the overall performance level when discussing the student's achievement with parents or guardians, but referred to other parts of the score report when planning for instruction. Future research will include usability studies to determine how educators use the overall performance level and the grade/content performance level descriptors, which describe what students in a performance level typically know and can do to inform instructional choices and goal setting.

XI.3.C. PROPOSITION 3: INFERENCES REGARDING STUDENT ACHIEVEMENT, PROGRESS, AND GROWTH CAN BE DRAWN AT THE CONCEPTUAL AREA LEVEL

Within each content area, four broad claims were developed, and then subdivided into nine conceptual areas (Chapter I). Conceptual areas are comprised of multiple, conceptually related content standards (Essential Elements) and nodes that support and extend beyond them.

Individual student score reports (Chapter VII) support interpretation and score use by providing information about student achievement at the conceptual area level. The individual student score report is comprised of two parts: the Performance Profile and the Learning Profile. The Performance Profile, a summary report of individual student results, includes bar graphs indicating the percent of skills mastered within each conceptual area, as well as a bulleted list of the specific skills mastered in each conceptual area. The Learning Profile, a more fine-grained summary of student mastery of specific knowledge, skills and understandings, includes linkage level mastery reported within each Essential Element and conceptual area (Chapter VII).

XI.3.C.i. Evidence Based on Content

Conceptual areas organize groups of EEs within claims to support understandings of how students make progress in the content of the claim (see Chapter III). The DLM test blueprints, which specify the pool of available EEs and requirements for coverage, impose constraints on EE choices to ensure student results reflect performance adequately across several conceptual areas. Specifying blueprint coverage requirements at the conceptual area level ensures representation and supports inferences at this level.

XI.3.C.ii. Evidence Based on Internal Structure

The learning map conceptual areas are the theoretical framework for reporting scores; item interrelationships within and across conceptual areas are imperative to understanding the internal structure (see Chapter VIII). Reliability evidence for 2014-2015 was calculated at the overall content area and at the Essential Element level, but not at the conceptual area level (which is the level between overall content and Essential Element). The reliability summaries for the number of linkage levels mastered within an EE showed moderate levels of reliability (45% of EEs with Pearson correlations ≥ .70). However, 84.8% of classification accuracy values were >= .80. The reliability statistics at the Essential Element level provide indirect evidence that the conceptual areas are sufficiently reliable for reporting and to support inferences about student achievement. Future reliability studies may be conducted to obtain reliability evidence for each conceptual area to inform future test development efforts and better support making inferences regarding student achievement at the conceptual area.

XI.3.C.iii. Evidence for Consequences of Assessment

Validity evidence is necessary to support the assumption that teachers use score reports to inform instructional choices and goal setting, and that score reports are useful and provide relevant information for teachers. Preliminary evidence from score report usability studies described in Chapter IX indicate that teachers refer to the performance profile results regarding conceptual areas when explaining reports to parents and when identifying patterns of strength and areas for improvement. Future studies will include usability studies to gain information as to how educators use score report information at the conceptual area level to guide instruction.

XI.3.D. PROPOSITION 4: ASSESSMENT SCORES PROVIDE USEFUL INFORMATION TO GUIDE INSTRUCTIONAL DECISIONS

This proposition is especially intended to support the intended use of results to plan instructional priorities and program improvements (use #3). Guiding instructional decisions may be conceptualized as individual student level decisions (i.e., those that teachers might make after receiving a student score report from the previous year) or school/program decisions (e.g., decisions about strategic priorities or curricular changes based on aggregated information). In 2014-15, evidence came from the design of score reports and interpretive materials, and studies on score report design and interpretation. To support this proposition, there must be evidence that scores are interpreted and used only for their intended purposes, and that teachers can use score reports to inform instructional choices and goal setting. While consequential evidence presented for earlier propositions also supports proposition 4, evidence for this proposition specifically addresses interpretation and use of report contents.

XI.3.D.i. Evidence for Consequences of Assessment

As described in Chapter VII, various guiding documents and supporting resources were created to help key stakeholders interpret assessment results as intended. The *Parent Interpretive Guide* provided a sample Individual Student Report to explain how the assessment measures student performance on alternate achievement standards for students with the most significant cognitive disabilities. Explanatory letter templates were developed to be used by teachers and state superintendents to introduce the student reports. These letters provide context for the reports including what the DLM assessment is, when it was administered, and what results tell about student performance. A teacher interpretive guide was provided for all those who would discuss results with parents or other stakeholders. The *Scoring and Reporting Guide for Administrators* was designed for principals and district administrators. It covered each type of report provided for the DLM assessments, presented suggestions for how to interpret each report, and suggested uses for the information.

As described in Chapter IX, research that informed the development of score reports included qualitative data collection and analysis to understand (1) parents' needs for information in score reports, (2) how stakeholders read and interpret score reports, and (3) how teachers would use assessment results to plan for individual and group instruction. Prototype score reports were developed based on parent perceptions of the challenges with previous alternate assessment score reports. Prototypes were reviewed and refined after multiple rounds of input from parents, educators, and parent advocates. The summative reports contain Performance Profiles and Learning Profiles.

There is preliminary evidence from stakeholder focus groups, teacher interviews, and paired discovery activities (see DLM Score Report Design and Use section in Chapter IX) that stakeholders can read the reports accurately and find them useful. In teacher interviews, the Learning Profile portion of the individual score report was most useful for the purpose of planning instruction, including re-teaching skills. Participants described using score report contents primarily for two parts of IEP development: statements on the student's present levels of performance and annual goals. Teachers also tended to use the performance level narrative and mastery skill list nearly verbatim in statements of present levels of performance.

Considering the newness of the DLM assessment system and the length and complexity of information in the individual student score reports, this line of score report research offers strong evidence in support of the proposition that scores provide information that can be used for instructional decision-making. Follow-up studies are planned on teacher decision-making and how score report interpretation translates into actual instructional change, within and across years. Evidence is still needed on score report interpretation by other stakeholder groups, including parents from diverse backgrounds and school administrators, and on the interpretation and use of aggregated reports for decision-making at the school and program levels. To date, this research has been limited to stakeholder interpretation of score reports, without the use of interpretive resources. Future research will also evaluate the extent to which these resources support appropriate interpretations and uses.

XI.3.E. Evaluation Summary

The accumulated evidence available by the end of 2014-15 provides preliminary support for the validity argument, particularly at a level that would be expected by the end of the first operational year of an assessment system designed from scratch on a compressed timeline. Each proposition is addressed by evidence in one or more of the categories of validity evidence, as summarized in Table 108. While many sources of evidence support multiple propositions, Table 108 lists the primary associations. For example, proposition 4 is indirectly supported by content-related evidence described for propositions 1 through 3. Table 109 shows the titles and sections for the chapters cited in Table 108.

Table 108. Dynamic Learning Maps Alternate Assessment System Propositions and Sources of Related Evidence for 2014-15.

	Sources of Evidence*				
Proposition	**Test Content**	**Response Processes**	**Internal Structure**	**Relations with Other Variables**	**Consequences of Testing**
Scores represent what students know and can do.	3, 4, 5, 6, 7, 8, 10, 23, 28	7, 11, 12, 20, 24, 29	1, 2, 9, 13, 22, 23, 25	9, 26	18, 27
Achievement level descriptors provide useful information about student achievement.	1, 6, 14, 15, 16, 17		22		19, 20, 27
Inferences regarding student achievement, progress, and growth can be drawn at the conceptual area level.	1, 5, 19		22		27
Assessment scores provide useful information to guide instructional decisions.					21, 27

*Note: * See Table 109 for a list of evidence sources. Only direct sources of evidence are listed. Some propositions are also supported indirectly by evidence presented for other propositions.*

Table 109. Evidence Sources Cited in Previous Table.

Evidence #	Chapter	Section
1	I	System Components
2	II	All
3	II	Development Process
4	III	Development of the Essential Elements
5	III	Test Blueprints
6	III	Essential Element Concept Maps for Test Development
7	III	Item Writing
8	III	External Reviews
9	III	Pilot Administration
10	III	Field Testing
11	IV	Test Administration Resources and Materials
12	IV	Implementation Evidence from 2014-15 Test Administration
13	V	All
14	VI	Standard Setting Approach
15	VI	Panelists
16	VI	Meeting Procedures
17	VI	Grade Level/Content Performance Level Descriptors
18	VII	Student Performance
19	VII	Score Reports
20	VII	Data Files
21	VII	Score Report Interpretation Resources
22	VIII	Reliability Evidence
23	IX	Evidence Based on Test Content
24	IX	Evidence Based on Response Process
25	IX	Evidence Based on Internal Structure
26	IX	Evidence Based on Relations to Other Variables
27	IX	Evidence Based on Consequences of Testing

Evidence #	Chapter	Section
28	X	Instructional Professional Development
29	X	Required Training for Test Administrators

The overall evaluation of the extent to which each proposition is supported by the evidence collected by 2014-15 is summarized in Table 110.

Table 110. Evaluation of Evidence for Each Proposition.

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

XI.4. CONTINUOUS IMPROVEMENT

XI.4.A. OPERATIONAL ASSESSMENT

As noted previously in this manual, 2014-2015 was the first year the DLM Alternate Assessment System was operational. While the 2014-2015 assessments were carried out in a manner that supports the validity of the proposed uses of the DLM information for the intended purposes, the Dynamic Learning Maps™ Alternate Assessment Consortium is committed to continual improvement of assessments, teacher and student experiences, and technological delivery of the assessment system. Through formal research and evaluation as well as informal feedback, some improvements have already been implemented for 2015-2016. This section describes examples of those improvements in test development, administration, scoring, and reporting; KITE system functionality; and design changes to support enhancements in psychometric modeling.

Improvements to test development procedures for 2015-2016 and planned improvements for future years focus on ensuring accurate, high quality assessment content. The guidelines and procedures for item writing are reviewed annually using multiple sources of information from the field and research findings and data collected throughout the school year. For example, internal reviews of operational content from 2014-2015 resulted in a number of technical corrections that improved experiences for teachers and students without changing the construct being assessed in any specific item. These types of corrections have also led to refined quality control processes. Information included on the TIPs used in 2015-2016 was also revised based on input from the field. Changes focused on increased usability, logical ordering and specific instructions for educators on how materials are to be used in teacher-administered testlets that require them. A description of assessment content improvements for 2015-2016 is provided in Appendix H.2, DLM Improvements for 2015-2016 – Memo to States.

Improvements to the 2015-16 test administration procedures focused on ensuring accessibility, accurate delivery of testlet assignments and a high-quality assessment experience for teachers and students. Improvements to synthetic read-alouds were made with a significant number of testlets receiving updated audio files to support student use of text-to-speech and alternate text for images. These updates made synthetic audio more consistent across testlets and improved the quality of read-alouds. Other improvements to accessibility features based on feedback from the field included enhancements to the quality of color contrast. The accessibility manual was updated to include improved explanations of supports and the use of accessibility features. Case examples of student with complex needs were included to assist educators with decision-making for students who require a combination of supports and other allowable practices.

Significant improvements were also made to the 2015-2016 required training for test administrators. Project staff and an ad-hoc committee of state partners reviewed the content of the required training. As a result, the training content was streamlined, and differentiated versions were created for new and returning DLM test administrators. Module quizzes were also improved and a new learning platform was selected, allowing better course design and management features for training modules.

Automated assessment delivery was improved for 2015-2016 by incorporating more rigorous checks of enrollment in the staging environment of the KITE system. This allowed project staff to use simulated data to identify problems that could lead to misadministration. A set of technology enhancements to the KITE system allowed more stringent data controls to be established for 2015-2016 to prevent misadministration due to student data changes during adaptive delivery. (Student data changes were the most frequent cause of delivery errors in 2014-2015).

Improvements were also made in 2015-2016 scoring and reporting. An internal audit of scoring and reporting procedures led to changes in 2015-2016 including revised quality control processes for data files and automated data checks for score files and score reports. Due to concerns about the number of linkage levels with moderate reliability statistics, the Learning Profile portion of the individual score report, which contained linkage level mastery information, was removed from the individual student score report beginning in 2015-2016.

Several score reporting improvements were intended to support use of the results. Beginning in 2015-2016 score reports will be available to states and districts through Educator Portal, providing easier local use of assessment results. Grade and content-specific PLDs were published to help states support the attainment of higher expectations for students with the most significant cognitive disabilities. Improvements were also made to resources available to support interpretation of score reports.

XI.4.B. Future Research

The continuous improvement process also leads to future directions for research to inform and improve the DLM Alternate Assessment System in 2016-2017 and beyond. Some areas for investigation have been described earlier in this chapter and throughout the manual.

As mentioned in Chapter II, as additional data is collected through operational assessment and field testing of testlets, work on empirical analyses of the DLM maps can begin. Based on changes to the 2015-2016 field test design, the DLM Alternate Assessment Consortium expects to have sufficient data by 2017-2018 to support improved modeling, perhaps providing a shift to node-based scoring rather than linkage level scoring. With this additional data several aspects of the structure of the learning maps can be evaluated, including the uniqueness of the hypothesized nodes, directionality of connections, and quality of model fit in a node-based DCM analysis. Work related to validation of the DLM maps is expected to continue to be a part of the ongoing effort to improve the accuracy and representativeness of learning map models that underlie the assessment system.

The next few years will also bring opportunities to more closely evaluate technical information from 2014-15. For example, reliability estimates will be calculated at the conceptual area level and indices will be more closely examined for variations across linkage levels and EEs. Correct classification rates for subgroups of students will provide additional insight on consistency of measurement for students across the performance continuum. These analyses, in conjunction with feedback from the field and data from the system on time taken for administration, can inform a discussion with states about whether future blueprint revisions should balance differently the demands of testing time, breadth of content coverage and the number of items that should be assess each skill. Adaptation data described in Chapter IV will be further mined for evidence regarding how often and for which EEs students regularly adapt in the system, which may inform future difficulty thresholds used for system adaptation.

Other research is also anticipated in the near future, as sample sizes increase across the second and subsequent years of operational delivery. For example, DIF analyses, which were limited in 2014-2015, may be replicated across additional items and with different focal and reference groups after the 2015-2016 administration. Studies on the comparability of results for students who use various combinations of accessibility supports are also dependent upon the availability of larger data sets. This line of research is expected to begin in 2017.

In the near future we also anticipate working with states to collect additional, state-level validity evidence. For example, states may collect data (e.g., online progress monitoring) that would be appropriate for use to evaluate the relationship of student responses on DLM assessments to other variables. Since states are responsible for making policy decisions and setting expectations regarding the use of assessment data, they are also well-positioned to provide additional procedural evidence on uses of DLM results for various purposes.

Longitudinal data collection is ongoing as part of the regular operations of the assessment system. An annually administered survey of educators will provide a source of data from which to investigate changes over time in some of the key assumptions of the validity argument. Additionally the survey will provide a means of investigating the long term effects of the assessment system for students and educators. Project staff are planning more intensive studies to collect evidence related to consequences of the assessment system. An example of an investigation now being planned focuses on how educators use reports of results to inform instructional choices and goal setting for students. Additional studies will focus on the extent to which overall system goals are met and negative consequences are avoided.

All future studies will be guided by advice from the DLM Technical Advisory Committee and the state partners, using processes established over the life of the consortium.