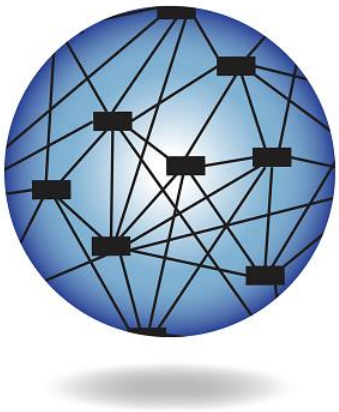
THE CENTER FOR EDUCATIONAL TESTING & EVALUATION
The University of Kansas



DYNAMIC® LEARNING MAPS

2015–2016 Technical Manual Update

Year-End Model

July 2017

Dynamic Learning Maps® Consortium. (2017, June). *2015–2016 Technical Manual Update – Year-End*. Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation.

Acknowledgments

The publication of this technical manual update builds upon documentation presented in the *2014–2015 TECHNICAL MANUAL* and represents further contributions to a body of work in the service of supporting a meaningful assessment system designed to serve students with the most significant cognitive disabilities. Hundreds of people have contributed to this undertaking. We acknowledge them all for their contributions.

Many contributors made the writing of this technical manual update possible. We are especially grateful for the contributions of 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*
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 update are listed below with gratitude.

Amy Clark, Ph.D., *Psychometrician Senior*
Meagan Karvonen, Ph.D., *Director*
Russell Swinburne Romine, Ph.D., *Associate Director for Test Development and Production*

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.

Sue Bechard
Brianna Beitling
Karen Erickson
Claire Greer
Lisa Harkrader
Neal Kingston
Allison Lawrence
Lee Ann Mills
Michael Muenks
Brooke Nash
Michelle Shipman
Jonathan Templin
Susan K. Thomas
William Jacob Thompson
Lisa Weeks

Additional thanks are given to Annie Davidson for her contributions to the original technical manual.

Table of Contents

List of Tables

List of Figures

I. INTRODUCTION

During the 2015–2016 academic year, the Dynamic Learning Maps® (DLM®) Alternate Assessment System offered assessments of student achievement in mathematics, English Language Arts (ELA), and science for students with the most significant cognitive disabilities in grades 3–8 and high school. Because the 2015–2016 academic year was the first year science was administered operationally, a separate technical manual was prepared for science (see Dynamic Learning Maps [DLM] Consortium, 2017).

The purpose of the system is to improve academic experiences and outcomes for students with the most significant cognitive disabilities by setting high, 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]

A complete technical manual was created for the first year of operational administration, 2014–2015. This technical manual provides updates for the 2015–2016 administration; therefore only sections with updated information are included in this manual. For a complete description of the DLM assessment system, refer to the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016).

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

I.1. BACKGROUND

In 2015–2016, DLM assessments were administered to students in 16 states: Alaska,[2] Colorado, Illinois, Iowa, Kansas, Mississippi, Missouri, New Hampshire, New Jersey, New York, North Dakota, Oklahoma, Utah, Vermont, West Virginia, and Wisconsin.

Additional state partners who did not administer operational assessments in ELA and mathematics in 2015–2016 include North Carolina and Pennsylvania.

In 2015–2016, the Center for Educational Testing and Evaluation at the University of Kansas continued to partner with the Center for Literacy and Disability Studies at the University of North Carolina at Chapel Hill 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).

I.2. ASSESSMENTS

Assessment blueprints consist of Essential Elements (EE) prioritized for assessment by the DLM Consortium. To achieve blueprint coverage, each student is administered a series of testlets. Each testlet is delivered through the 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 own 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 reassessed 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

[2]Alaska administered assessments but stopped all statewide testing mid-window and did not receive summative results, so Alaska's results are not included in any of the data presented in later chapters.

embedded and spring windows are separate. In 2015–2016, the states participating in the integrated model included Iowa, Kansas, Missouri, North Dakota, Utah, and Vermont.

- **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. In 2015–2016, the states participating in the year-end model included Alaska, Colorado, Illinois, Mississippi, New Hampshire, New Jersey, New York, Oklahoma, West Virginia, and Wisconsin, and two Bureau of Indian Education schools, Miccosukee and Choctaw.

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 an overview of the assessment and administration for the 2015–2016 academic year and a summary of contents of the remaining chapters. While subsequent chapters describe the essential components of the assessment system separately, several key topics are addressed throughout this manual, including accessibility and validity.

Chapter II was not updated for 2015–2016. See the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016) for a description of the process by which the DLM maps were developed.

Chapter III outlines procedural evidence related to test content and response-process propositions.[3] Chapter III includes summaries of external reviews for content, bias, and

[3]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 of this manual).

accessibility. The final portion of the chapter describes the operational and field-test content available for 2015–2016.

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 provides updated evidence for spring routing in the system, as well as teacher survey results collected during 2015–2016.

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.

DLM Consortium, 2016) for a description of 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.

Chapter VII reports the 2015–2016 operational results, including student participation data. The chapter details the percentage of students at each performance level (impact); subgroup performance by gender, race, ethnicity, and English language learner status; and the percentage 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 performance level, content area, conceptual area, EE, linkage level, and conditional linkage level.

Chapter IX describes additional validation evidence not covered in previous chapters. The chapter details how the internal structure of the assessment was evaluated through differential items. In addition, it presents updated teacher survey results specific to the validity argument.

Chapter X describes the training and professional development that was offered across the DLM Consortium, including the 2015–2016 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 2015–2016 instructional professional development are included.

Chapter XI synthesizes the evidence provided in the previous chapters. It also provides future directions to support operations and research for DLM assessments.

II. MAP DEVELOPMENT

Learning map models are a unique key feature of the Dynamic Learning Maps® (DLM®) Alternate Assessment System and drive the development of all other components. For a description of the process used to develop the map models, including the detailed work necessary to establish and flesh out the DLM maps in light of the Common Core State Standards and the needs of the student population, see Chapter II of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016).

III. ITEM AND TEST DEVELOPMENT

Chapter III of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016) describes general item and test development procedures. This chapter provides an overview of updates to item and test development for the 2015–2016 academic year. The first portion of the chapter provides a summary of item and testlet information, followed by the 2015–2016 external reviews of items and testlets for content, bias, and accessibility. The next portion of the chapter describes the operational assessments for 2015–2016, followed by a section describing field tests administered in 2015–2016.

For a complete description of item and test development for Dynamic Learning Maps® (DLM®) assessments, including information on the use of evidence-centered design and Universal Design for Learning in the creation of concept maps to guide test development; external review of content; and information on the pool of items available for the pilot, field tests, and 2014–2015 administration, see the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016).

III.1. ITEMS AND TESTLETS

This section describes information pertaining to items and testlets administered as part of the DLM assessment system, including a summary of item-writer characteristics, English language arts (ELA) blueprint coverage, ELA reading passage development, information on ELA writing testlets, and the selection of accessible graphics for testlets. With the exception of the description of item-writer characteristics during the 2015–2016 test development cycle, the remainder of this section provides expanded information about item and testlet development practices in effect beginning in 2014–2015. This expanded information was included in the 2016–2017 update at stakeholder request. For a complete summary of item and testlet development procedures that began in 2014–2015 and were implemented in 2015–2016, see Chapter III of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016).

III.1.A. ITEM WRITER CHARACTERISTICS

Development of DLM items and testlets began in the summer of 2013. Additional items and testlets were developed during 2014. During these years, most item writing occurred during summer events in which content and special education specialists worked on-site in Lawrence, Kansas, to develop DLM assessments. For the 2015–2016 year, most item writers came from the previous item-writing events. The exception was four internal staff members: three graduate research assistants and one full-time staff member who received training and wrote testlets.

An item writer survey was used to collect demographic information about the teachers and other professionals hired to write DLM testlets. In total, 25 item writers contributed to testlets for the 2015–2016 year, including 15 for mathematics and 10 for ELA. The median and range of

number of years of teaching experience in four areas is shown in Table 1 for the ELA and mathematics item writers.

Table 1. *Item Writers' Years of Teaching Experience*

Area	ELA		Mathematics	
	Median	Range	Median	Range
Pre-K–12	7	0–27	15	0–37
ELA	9	0–27	18	1–34
Mathematics	9	9	16	1–35
Special Education	3	1–17	17	0–37

Item writers were also asked to indicate which grade(s) they had experience teaching. There were five ELA item writers with experience at the elementary level (grades 3–5), six with experience in middle school (grades 6–8), and four with experience in high school. Similarly, there were five mathematics item writers with experience at the elementary level, (grades 3–5), five with experience in middle school (grades 6–8), and four with experience in high school.

All 25 item writers held at least a bachelor's degree. The distribution and types of degrees held by item writers are shown in Table 2 and Table 3.

Table 2. *Level of Degree*

Degree	ELA Item Writers		Mathematics Item Writers	
	n	%	*n*	%
Bachelor's	10	100	15	100
Master's	4	40	9	60
Other	1	10	1	7

Table 3. *Degree Type for All Item Writers*

Degree	ELA Item Writers	Mathematics Item Writers
Bachelor's Degree		
Education	3	7
Content Specific	4	2
Special Education	1	1
Other	2	4
Master's Degree		
Education	2	4
Content Specific	1	0
Special Education	0	2
Other	2	3

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 lowest levels of experience occurred in the disability categories of deaf/hard of hearing and traumatic brain injury categories. All disability categories reported on the survey are listed in Table 4.

Table 4. *Item Writer Experience by Content Area*

Disability Category	ELA Item Writers		Mathematics Item Writers	
	n	%	*n*	%
Blind/Low Vision	0	0	5	33
Deaf/Hard of Hearing	1	10	3	20
Emotional Disability	5	50	8	53
Mild Cognitive Disability	5	50	11	73
Multiple Disabilities	3	30	4	27
Orthopedic Impairment	3	30	4	27
Other Health Impairment	4	40	6	40
Severe Cognitive Disability	2	20	3	20
Specific Learning Disability	6	60	9	60
Speech Impairment	3	30	4	27
Traumatic Brain Injury	1	10	2	13
None of the above	2	20	3	20

Of the item writers, 20% had experience administering an Alternate Assessment of Alternate Achievement Standards (AA-AAS) prior to their work on the DLM project, 24% reported working with students eligible for AA-AAS at the time of the survey, and 24% reported holding a National Board certification.

III.1.B. Blueprint Coverage

DLM Essential Elements (EEs) in ELA use the same strands found in the Common Core State Standards: Reading Literature (RL), Reading Information (RI), Language (L), Writing (W), and Speaking and Listening (SL). All grades include EEs in the RL, RI, L, and W strands. RL and RI EEs are assessed in reading testlets. Writing EEs are assessed in writing testlets. Language EEs are sometimes assessed in reading testlets, when the content of the EE lends itself to assessment in the context of reading, and sometimes in writing testlets, when the EE is better measured as part of a writing task. SL standards were not included on the ELA test blueprint. State partners indicated that general assessments delivered within their states at the time the blueprint was approved did not include SL standards on their test blueprints. Since DLM assessments were designed to be an alternate to the general assessments administered in DLM states, SL standards were left to be taught and assessed at the local level.

For a complete description of DLM test blueprints, including a complete description of mathematics blueprints, see Chapter III section I.1.B Test Blueprints of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016).

III.1.C. English Language Arts Reading Passage Development

Passages for DLM assessments include stories and informational texts. When administered to students via the Kansas Interactive Testing Engine (KITE®), stories and informational texts used in ELA reading testlets are presented in a page-by-page or screen-by-screen format, with one to three sentences per screen and an accompanying photographic illustration. Students can navigate through the passage at their own pace by using the **NEXT** button in the user interface and go back as desired within the passage by using the **BACK** button. Photographs were selected to illustrate texts but were intended solely to support the texts and provide an engaging assessment experience, not to be a replacement for the words in the text. During administration, students first read a story or informational text in its entirety. Then students read the text a second time. In the second reading, items are presented embedded within the story or informational text and/or at its conclusion.

DLM stories and informational texts were developed to use clear language and reduce the need for prior knowledge. To allow students to access the content, texts were written and reviewed internally to reduce linguistic structural barriers that may have occurred as a result of complex grammatical structures or syntax. DLM stories and texts are short, between 50 and 250 words. They include high-frequency, easily decodable words, such as those found on the research-supported *DLM Core Vocabulary List*, which includes words that are commonly used for expressive communication in social and academic contexts. Simple sentences were favored and pronoun use was reduced. Consistency in sentence structure within a story or informational text was favored.

Guidelines were developed to support passage writers in producing accessible texts for use in DLM assessments. Four criteria were applied to all DLM passages, and an additional four guidelines were developed and applied only to informational texts. These guidelines are listed in Table 5. Passage writers received training on the use of the guidelines, which were subsequently used by external reviewers when evaluating reading passages. For a complete summary of external review of ELA passages, see *Results from External Review During the 2014–2015 Academic Year* (Clark, Swinburne Romine, Bell, & Karvonen, 2016).

Table 5. *Text Development Content Guidelines*

	Criterion	Guidelines
All DLM Texts	1. Accessible Text Language	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.
	2. Accessible Text Content	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 conform to the specifications for where it will be used in assessments.
	3. Instructional Relevance	The text is instructionally relevant to students for whom it was written. It is grade-level appropriate and engaging.
Informational Texts	4. Fair Construct	The text represents the topic accurately without requiring prior knowledge.
	5. Diversity	Where applicable, there is a fair representation of diversity in race, ethnicity, gender, disability, and family composition.
	6. People Positive	The text uses appropriate labels for groups of people. People first language is used for individuals with disabilities. Populations are not depicted in a stereotypical manner.
	7. Fair Language	The language in the text neither prevents nor advantages any regional or cultural group from demonstrating what they know about the targeted content.

III.1.D. English Language Arts Writing Testlets

In 2014–2015 and 2015–2016, every grade level had an Emergent and Conventional writing testlet available, each of which measures several EEs. Writing testlets include EEs in the Writing strand, and in some grades, EEs in the Language strand. Emergent writing testlets measure the Initial Precursor and Distal Precursor linkage levels, while conventional writing testlets measure the Proximal Precursor, Target, and Successor linkage levels. Because writing testlets measure multiple EEs and linkage levels, the structure of writing testlets differs from that of other testlets.

All writing testlets are teacher administered. The testlet 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. Item types are either multiple-choice single select **[single-select multiple choice]** or multiple-choice multi select **[multi-select multiple choice]**. 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 they complete items in the testlet. Figure 1 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, both "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 1). The inclusion of multiple, correct response 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 first name. As such, each response option is associated with a different EE and linkage level.

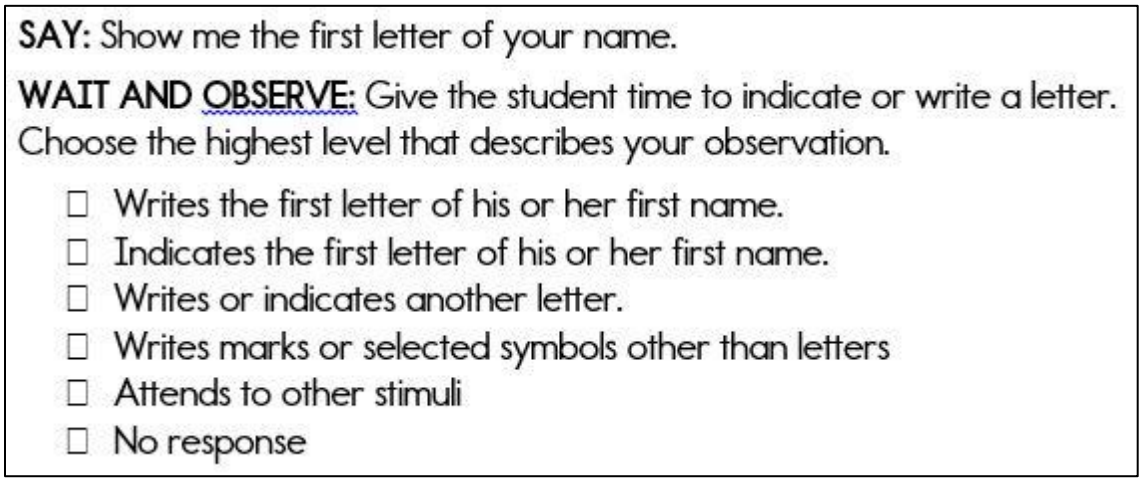


Figure 1. Example of ELA Emergent writing item focused on process.

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 2 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. A complete description of writing testlets can be found in Chapter III of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016).

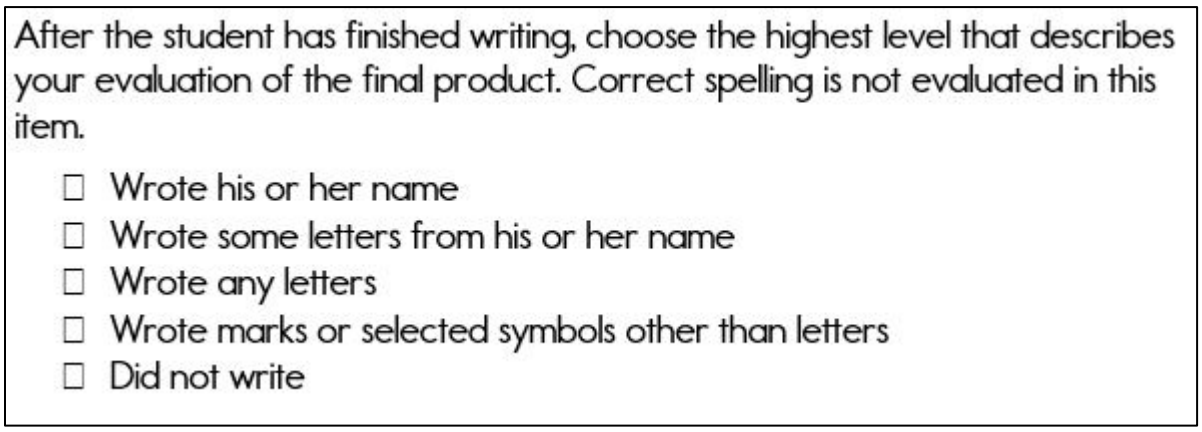


Figure 2. Example of ELA Conventional writing item focused on product.

Because writing items measure multiple EEs and linkage levels, writing items are scored at the option level rather than item level. This means that rather than having a single correct answer and several distractors for the item, each answer option is treated as a separate true or false item that is scored individually as evidence for the specific EE and linkage level it measures. For writing items that are multiple-choice single select **[single-select multiple choice],** the answer options often subsume other answer options. This means that selection of one response may inherently mean other answer options are also scored as correct. In the example provided in Figure 1, a selection of the first answer option, writes the first letter of his or her name, would result in the other answer options, such as "indicates the first letter of his or her first name," also being scored as correct.

The scoring process for DLM writing testlets follows. Data are extracted from the database that houses all DLM data. For writing items, the response-option identifiers are treated as item identifiers so that each response option can be scored as correct or incorrect for the EE and linkage level it measures. Additionally, response-option dependencies are built in, based on scoring directions provided by the ELA test-development team, to score as correct response options that are subsumed under other correct response options. Once the data structure has been transformed and response-option dependencies are accounted for, the writing data are combined with all other data to be included in the calibration process. For more information on calibration, see Chapter V of this manual.

III.1.E. Selection of Accessible Graphics for Testlets

Graphics for mathematics and ELA reading testlets and photographs used to illustrate ELA reading testlets were selected using guidelines developed with input from state partners to ensure accessibility for students. Graphics in mathematics testlets use colored line drawings; they are designed to employ high contrast and provide clear, simple, graphic representations of content. They are used only when required to assess the construct and for engagement

activities. Graphic designers and item writers received training to avoid the creation of items that rely on students' perception of color. Image quality and accessibility were reviewed as a part of the external-review process for items and testlets.

ELA reading assessments use photographs to support the presentation of a book format in reading text. Because independent interaction with text and linguistic comprehension depend on students representing the meaning of words and the concepts that they represent, the illustrations are of less importance than the words in a text. Photographs used in DLM assessments are intended to support the text, not to link so closely to the text that a student could infer the story based only on the images. Photographic images in ELA reading testlets were selected from free, publicly available, Creative Commons-licensed images available on an Internet photo-sharing site. After initial internal guidelines were used for pilot testing in 2013, revised guidelines for the use of images in ELA reading testlets were developed with input from partner states in 2014 and applied to all ELA reading testlets. The guidelines addressed both accessibility considerations including image clarity, contrast, and consistency within texts, as well as the exclusion of biased or sensitive material. Text writers and staff used the guidelines to select images for use in ELA reading testlets. The guidelines for images were incorporated into external review processes to ensure accessibility and to avoid biased or sensitive content. A complete list of the guidelines is included in Appendix A.

III.2. 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 content be accepted, revised, or rejected. Feedback from external reviewers was used to make final decisions about assessment items before they were field-tested.

Overall, the process and review criteria for external review in 2015–2016 remained the same as in 2014–2015. Minor changes were made, including using fewer reviewers who completed more assignments and increasing the amount paid to reviewers per review.

III.2.A. Review Recruitment, Assignments, and Training

In 2015–2016, a volunteer survey was used to recruit external review panelists. Volunteers for the external review process completed the Qualtrics survey to capture demographic information and information about volunteers' education and experience. These data were then used to identify panel types (content, bias and sensitivity, and accessibility) for which the volunteer would be eligible. A total of 19 people from year-end model states completed the required training, and 14 of those were placed on external review panels.

Of the 14 reviewers placed on panels, nine completed reviews. Each reviewer was assigned to one of the three panel types. Five ELA reviewers were on panels: zero on accessibility panels, three on content panels, and two on bias and sensitivity panels. Four mathematics reviewers

were on panels: two on accessibility panels, one on a content panel, and one on a bias and sensitivity panel. In addition, three power reviewers and two hourly reviewers reviewed all three panel types as needed for each content area.

Year-end model panelists primarily reviewed testlets, comprising three to eight tasks and measuring multiple EEs. However, when needed, year-end reviewers also reviewed testlets designed for instructionally embedded assessments, which are available in all states regardless of the assessment model.

The professional roles reported by the 2015–2016 reviewers are shown in Table 6. Reviewers who reported Other roles included specialized teachers and individuals identifying multiple categories.

Table 6. *Professional Roles of External Reviewers*

Role	ELA		Mathematics	
	n	%	*n*	%
Classroom Teacher	3	60.0	2	50.0
District Staff	0	0.0	2	50.0
Other	2	40.0	0	0.0

Reviewers had varying experience teaching students with the most significant cognitive disabilities. ELA reviewers had a median of 5 years of experience with a minimum of 1 year and a maximum of 9 years. Mathematics reviewers had a median of 13 years of experience teaching students with the most significant cognitive disabilities, with a minimum of 3 years and a maximum of 30 years.

All ELA and mathematics reviewers were female, non-Hispanic/Latino, and Caucasian. Table 7 reports the population density of schools in which reviewers taught or held a position. Within the survey, *rural* was defined as a population living outside settlements of 1,000 or fewer inhabitants, *suburban* was defined as an outlying residential area of a city of 2,000–49,000 or more inhabitants, and *urban* was defined as a city of 50,000 inhabitants or more.

Table 7. *Population Density for Schools of External Reviewers*

Population Density	ELA		Mathematics	
	n	%	*n*	%
Rural	1	20.0	2	50.0
Suburban	2	40.0	1	25.0

Population Density	ELA		Mathematics	
	n	%	*n*	%
Urban	2	40.0	1	25.0

Review assignments were given throughout the year. Reviewers were notified by email each time they were assigned collections of testlets. Each review assignment required 1.5 to 2 hours to complete. In most cases, reviewers had between 10 days and 2 weeks to finish an assignment.

III.2.B. Results of Reviews

Most of the content reviewed externally during the 2015–2016 academic year was included in the spring testing window. On a limited basis, reviewers examined content for the upcoming 2016–2017 school year. For ELA, the percentages of items or testlets rated as *accept* across grades, pools, and rounds of review ranged from 85% to 92%. The rate at which content was recommended for rejection ranged from 1% to 3% across grades, pools, and rounds of review. For mathematics, the percentages of items or testlets rated as *accept* ranged from 87% to 94%. The rate at which content was recommended for rejection ranged from less than 1% to 1%. A summary of the content team decisions and outcomes is provided here. A more detailed report and outcomes from external reviews are included in the external review technical report for 2015–2016 (Clark, Beitling, Bell, & Karvonen, 2016).

III.2.B.i. Content Team Decisions

Because multiple reviewers examined each item and testlet, external review ratings were compiled across panel types, following the same process used for 2014–2015. For each item and testlet, DLM content teams reviewed and summarized the recommendations provided by the external reviewers. 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, or (e) more information needed.

DLM 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 97% 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 329 minor revisions to items and 225 minor revisions to testlets. The mathematics content team retained 69% of items and testlets sent out for external review. As with ELA, most item and testlet revisions were minor. The mathematics team made 310 minor revisions to items and 171 minor revisions to testlets. Additional detail on review outcomes is included in the 2015–2016 external review technical report (Clark et al., 2016).

III.3. Operational Assessment Items for 2015–2016

Operational assessments were administered during the spring testing window. A total of 848,596 operational test sessions were administered; one test session is one testlet taken by one student. Only test sessions that were complete or in progress at the close of the testing window counted toward the total test sessions.

Testlets were made available for operational testing in 2015–2016 based on the 2014–2015 operational pool and the promotion of testlets field-tested during 2014–2015 to the operational pool following their review. Table 8 and Figure 9 summarize the total number of operational testlets by content area for 2015–2016. A total of 754 operational testlets were available across grades and content areas. This total also included 236 EE/linkage-level combinations (108 mathematics, 128 ELA) for which more than one testlet was available because both a braille and general versions were available.

Table 8. *2015–2016 ELA Operational Testlets*

Grade	*n*
3	43
4	43
5	41
6	34
7	33
8	29
9	41
10	19
11	25
English 2	21
English 3	20
Grand Total	**349**

Table 9. *2015–2016 Mathematics Operational Testlets* (N = *405)*

Grade	*n*
3	29
4	34
5	29
6	29
7	29
8	29
9	42
10	43
11	46
Algebra I	30
Algebra II	29
Geometry	36
Grand Total	**405**

Similar to 2014-2015, *p*-values were calculated for all operational items to summarize information about item difficulty.

Figure 3 and Figure 4 include the *p* values for each operational item for ELA and mathematics. To prevent items with small sample size from potentially skewing the results, the student sample-size cutoff for inclusion in the *p* values plots was 20. In general, ELA items were easier than mathematics items, as evidenced by more items falling in the higher bin (*p*-value) ranges. Writing items were omitted from this plot because scoring occurred at the option level rather than item level.

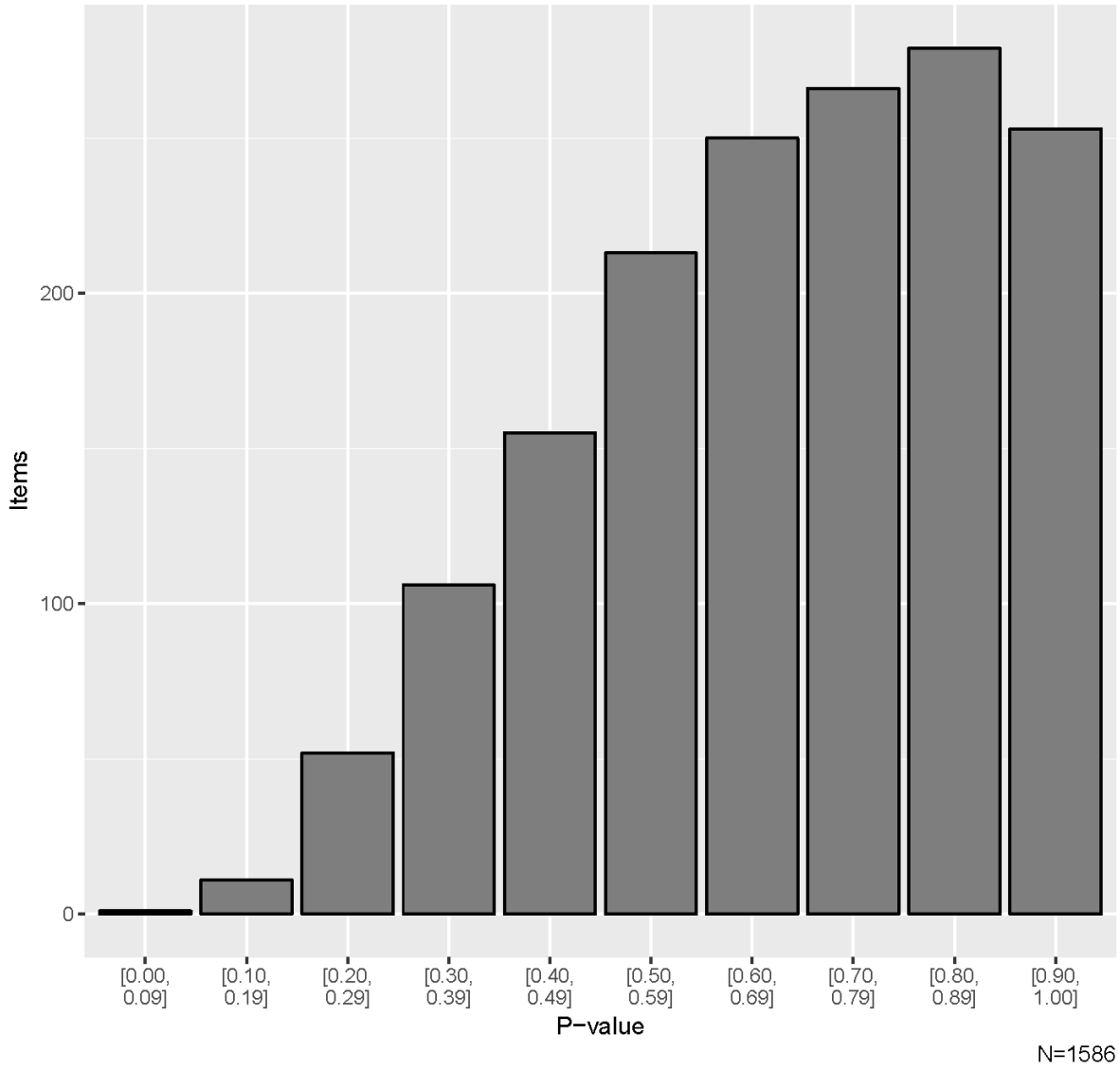


Figure 3. P-values for ELA 2015-2016 operational items.

Note: Writing items and items with a sample size less than 20 were omitted.

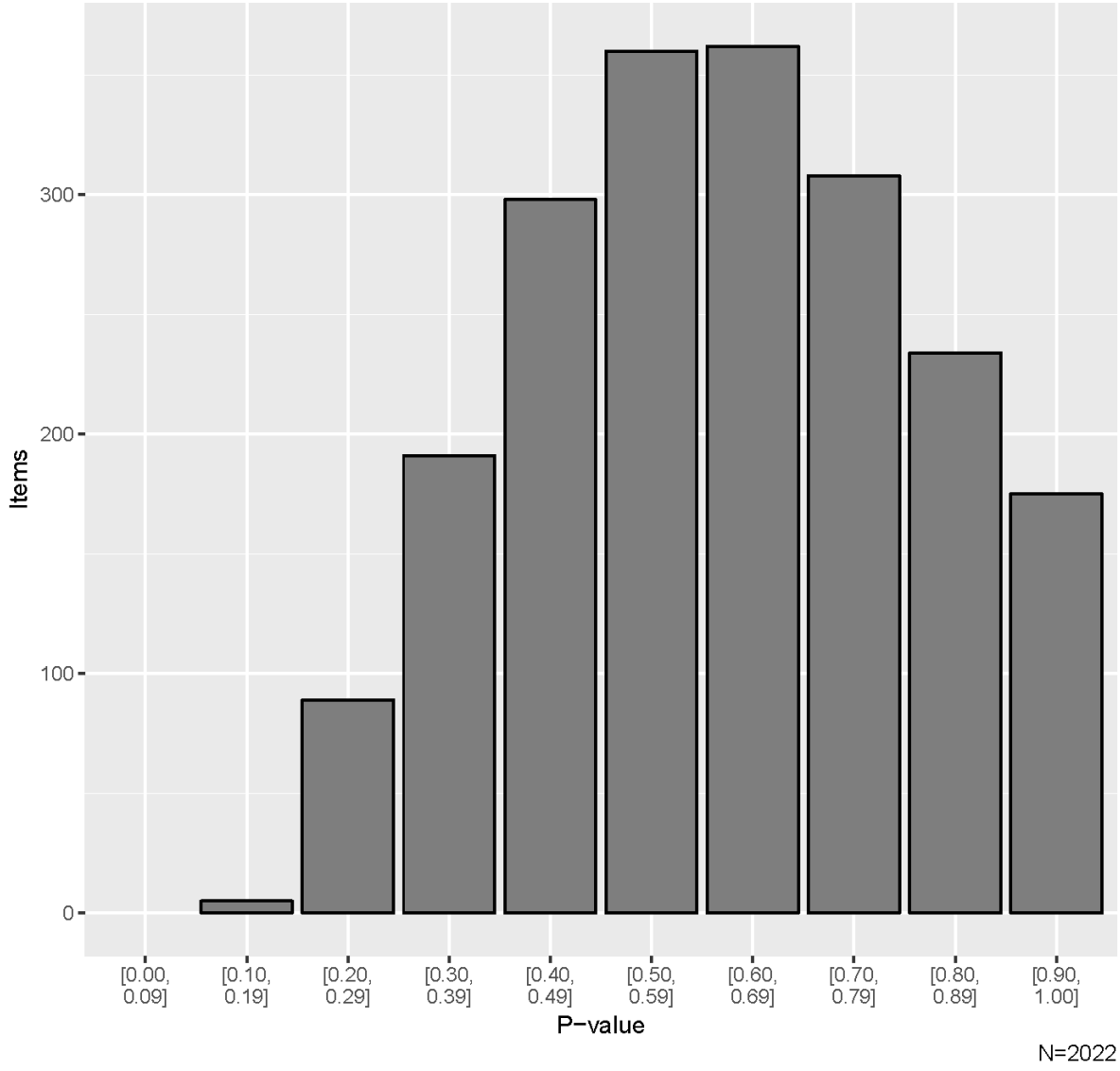


Figure 4. P-values for mathematics 2015-2016 operational items.

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

Standardized difference values were also calculated for all operational items with a student sample size of at least 20 to compare the *p* value for the item to all other items measuring the same EE and linkage-level combination. The standardized difference values provide one source of evidence of internal consistency. Figure 5 and Figure 6 summarize the standardized difference values for operational items. Most items fell within two standard deviations of the mean for the EE and linkage level. As additional data are collected and decisions are made

regarding item pool replenishment, item standardized difference values will be considered in accompaniment with item misfit analyses to determine items and testlets that are recommended for retirement.

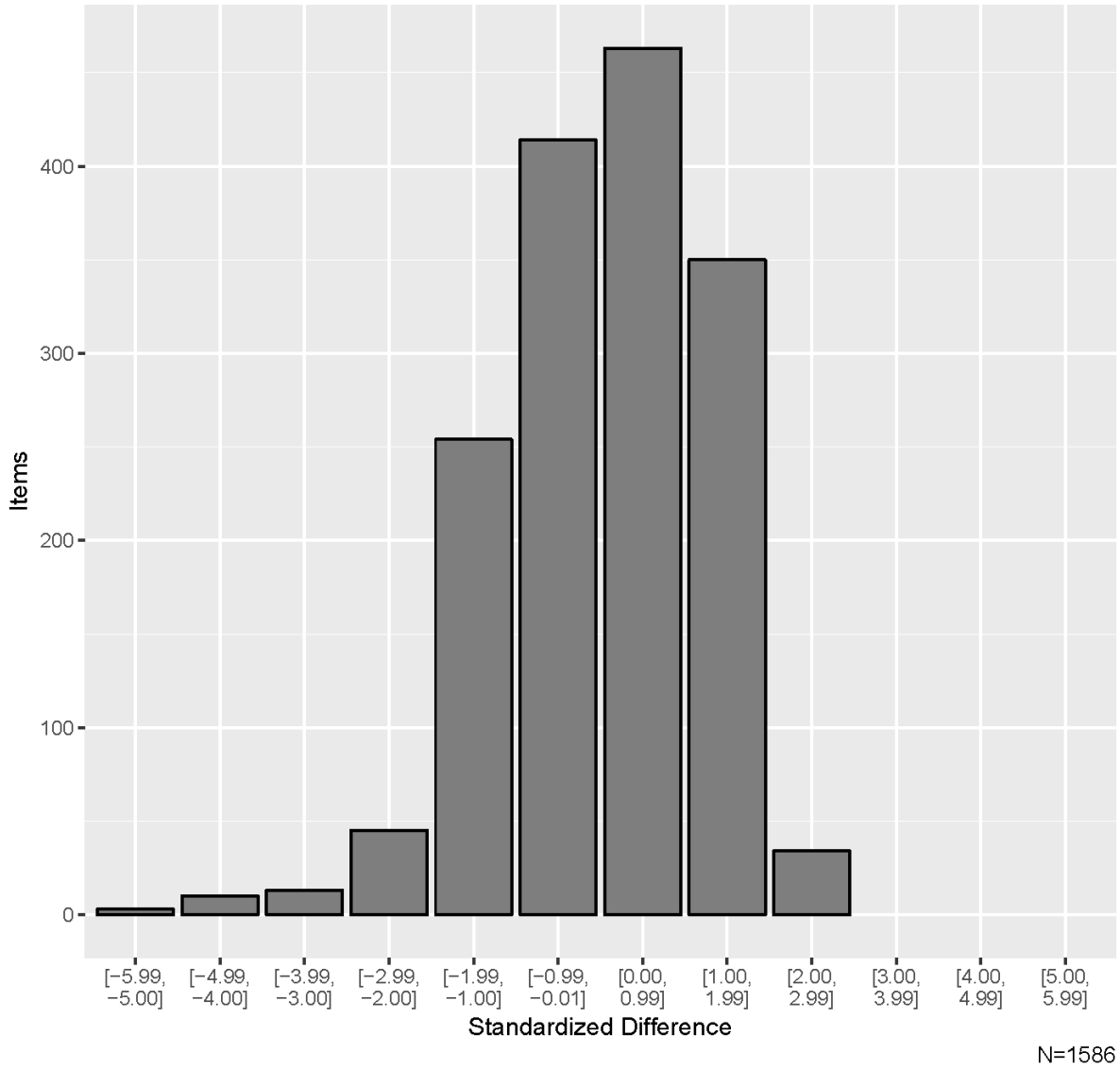


Figure 5. Standardized difference *z* scores for ELA 2015-2016 operational items.

Note: Writing items and items with a sample size less than 20 were omitted.

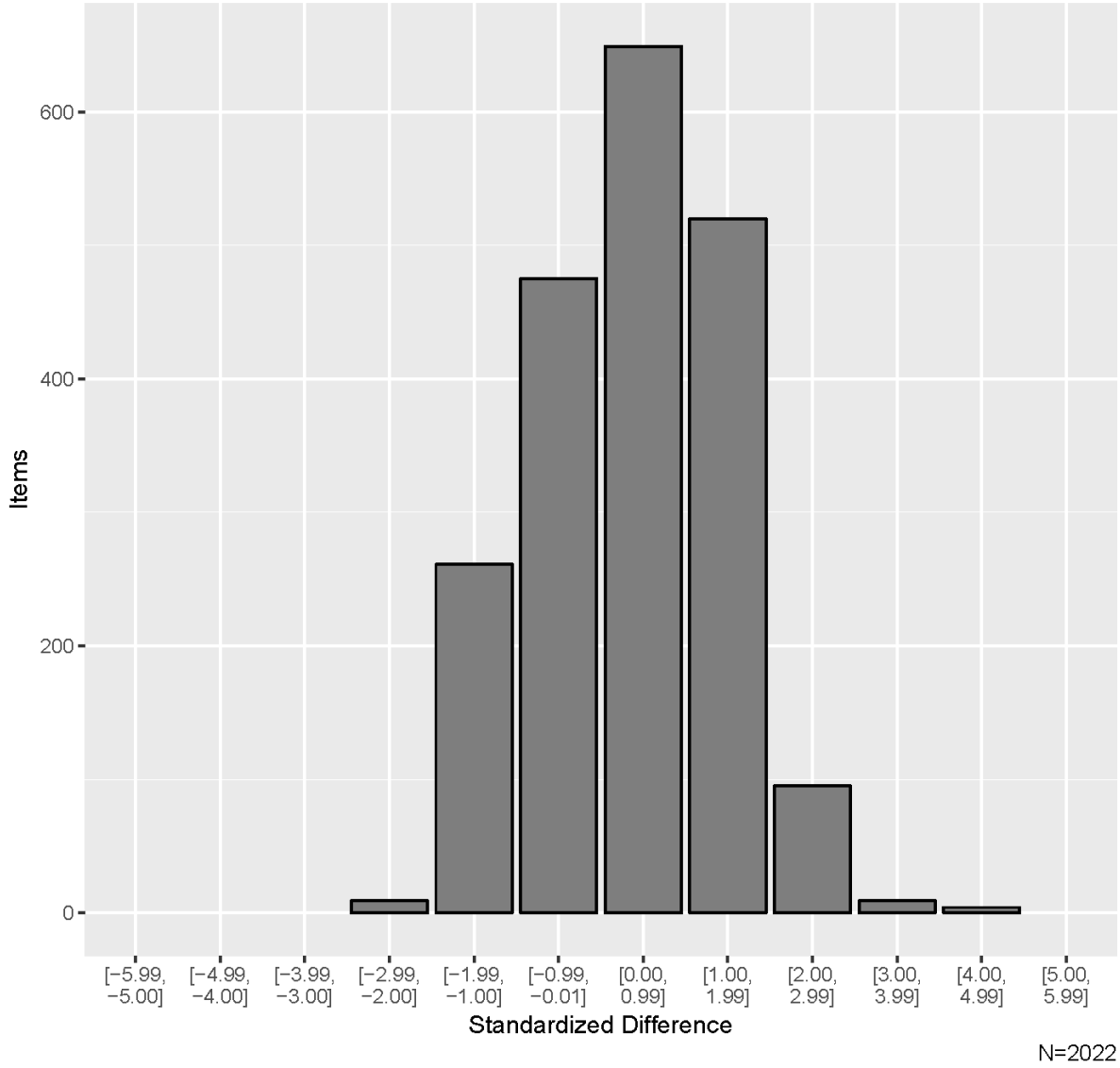


Figure 6. Standardized difference *z* scores for mathematics 2015-2016 operational items.

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

III.4. FIELD TESTING

During the 2015–2016 academic year, DLM field tests were administered to evaluate item quality for EEs assessed at each grade level for ELA and mathematics. Field testing is conducted to deepen operational pools so that multiple testlets are available in spring windows. By deepening the operational pools, testlets can also be evaluated for retirement in instances where

other testlets perform better. Additionally, as testlet exposure rates are examined in subsequent years, deeper pools will allow retirement of testlets that have exceeded exposure specifications. A complete summary of prior 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) and in Chapter III of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016).

III.4.A. Description of Field Tests

Collection of field test data during the spring window was first implemented in the 2015–2016 academic year. During the spring administration, all students received up to four field-test testlets for each content area upon completion of all operational testlets. While the test name did not indicate the testlet was a field-test testlet, teachers could likely surmise the difference due to multiple field-test testlets populating for each content area following completion of the final testlet.

The spring field-test administration was designed to collect data for each participating student at more than one linkage level for an EE to support future modeling development. (See Chapter V of this manual for more information.) As such, the field-test testlets were assigned at one linkage level below the last linkage level at which the student was assessed. Due to the process of assigning the testlet one linkage level lower than the last testlet, no Successor-level testlets were field-tested during the spring window.

Testlets were made available for spring field testing in 2015–2016 based on the availability of field-test content for each section of the assessment. Table 10 and Table 11 summarize the total number of field-test testlets by content area and grade level for 2015–2016. A total of 250 field-test testlets were available across grades and content areas.

Table 10. *2015–2016 ELA Field-Test Testlets (N = 81)*

Grade	*n*
3	11
4	11
5	10
6	7
7	7
8	5
9	8
10	13
11	9
English 2	0
English 3	0
Grand Total	**81**

Table 11. *2015–2016 Mathematics Field-Test Testlets (N = 169)*

Grade	*n*
3	9
4	13
5	13
6	14
7	14
8	13
9	19
10	22
11	23
Algebra 1	9
Algebra 2	9
Geometry	11
Grand Total	**169**

Participation in spring field testing was not required in any state, but teachers were encouraged to administer all available testlets to their students. Participation rates for ELA and mathematics in 2015–2016 are shown in Table 12 below. High participation rates allowed for all testlets to meet sample-size requirements (responses from at least 20 students) and thus undergo statistical and content review prior to moving to the operational pool.

Table 12. *2015–2016 Participation Rates in Spring Field Testing by Content Area*

Content Area	*n*	%
ELA	31,319	44.8
Mathematics	30,435	43.3

III.4.B. Field-Test Results

Data collected during each field test are 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.4.B.i. Item Flagging

Criteria used for item flagging during previous field-test events were retained for 2015–2016. Content teams flagged items for review if they met any of the following statistical criteria:

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

Items that had a sample size of at least 20 cases were reviewed. Items with a sample size of less than 20 were slated for retest in a subsequent field-test window to collect additional data prior to making item-quality decisions.

Figure 7 and Figure 8 summarize the *p* values for items field-tested during the 2015–2016 spring window. Most items fell above the 35% threshold for flagging. Items below the threshold were reviewed by test-development teams for each content area.

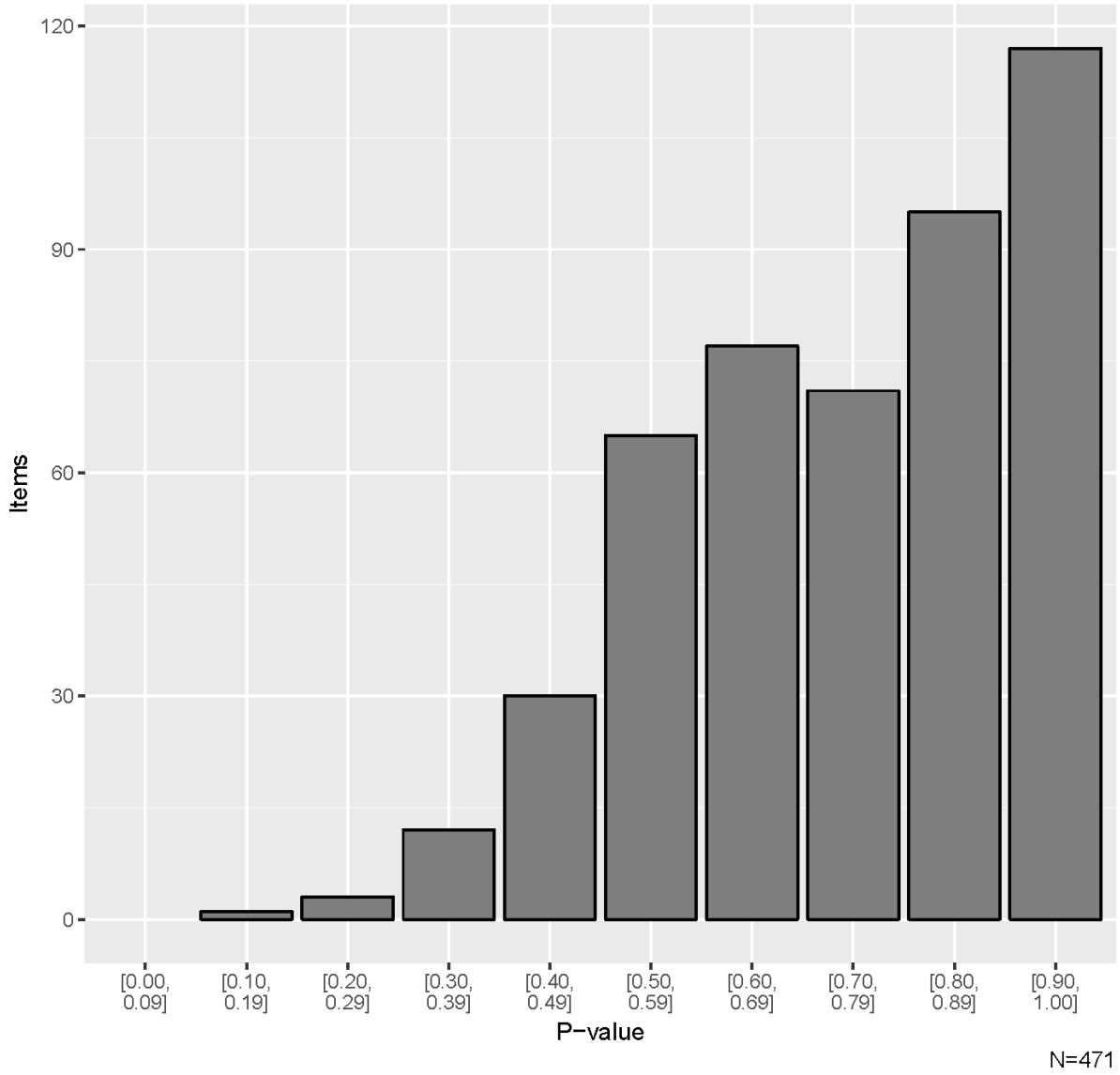


Figure 7. P-values for 2015-2016 ELA items field-tested during spring window.

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

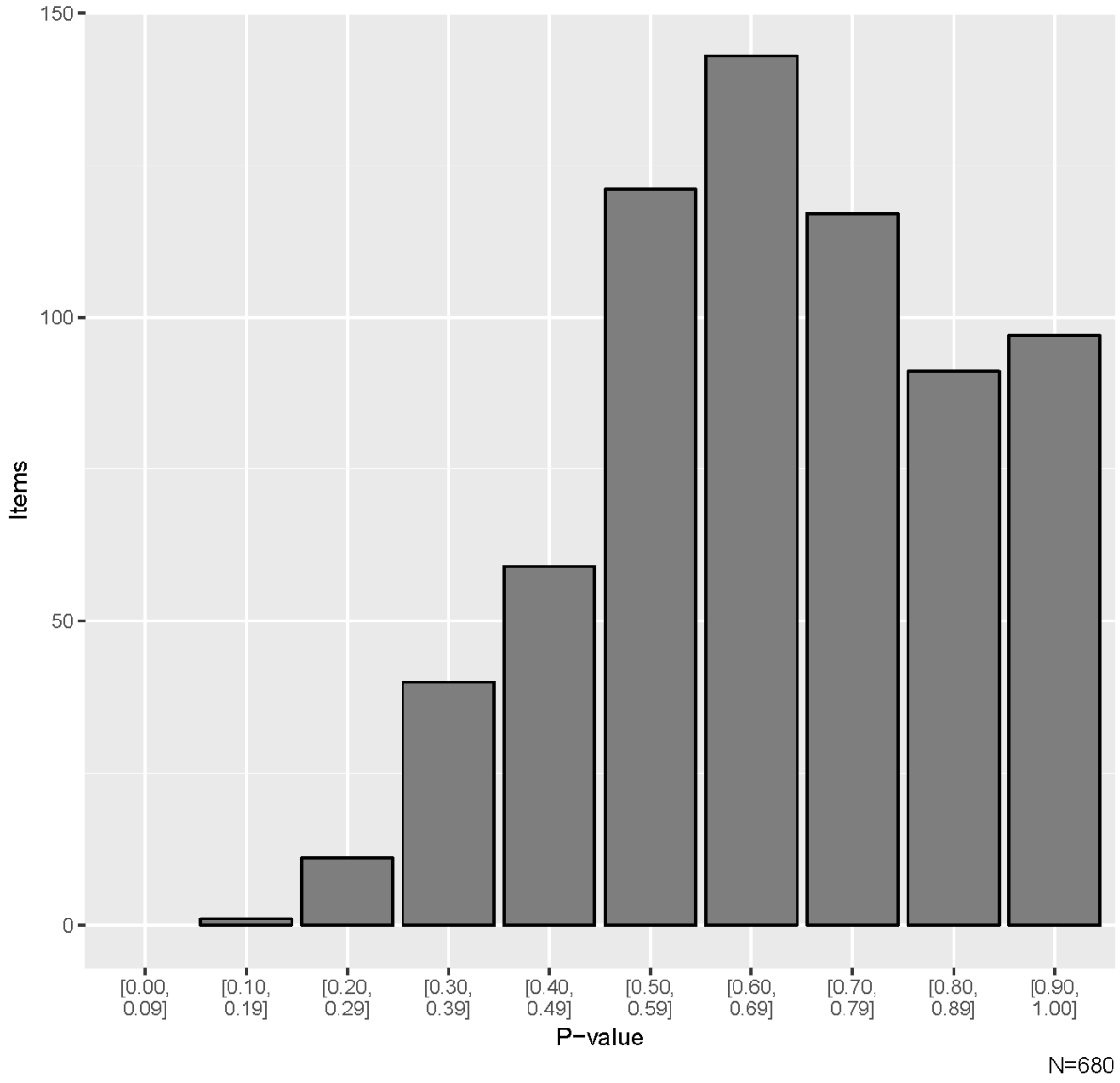


Figure 8. *P*-values for 2015-2016 mathematics items field-tested during spring window.

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

Figure 9 and Figure 10 summarize the standardized difference values for items field-tested during the 2015–2016 spring window. Most items fell within two standard deviations of the mean for the EE and linkage level. Items beyond the threshold were reviewed by test-development teams for each content area.

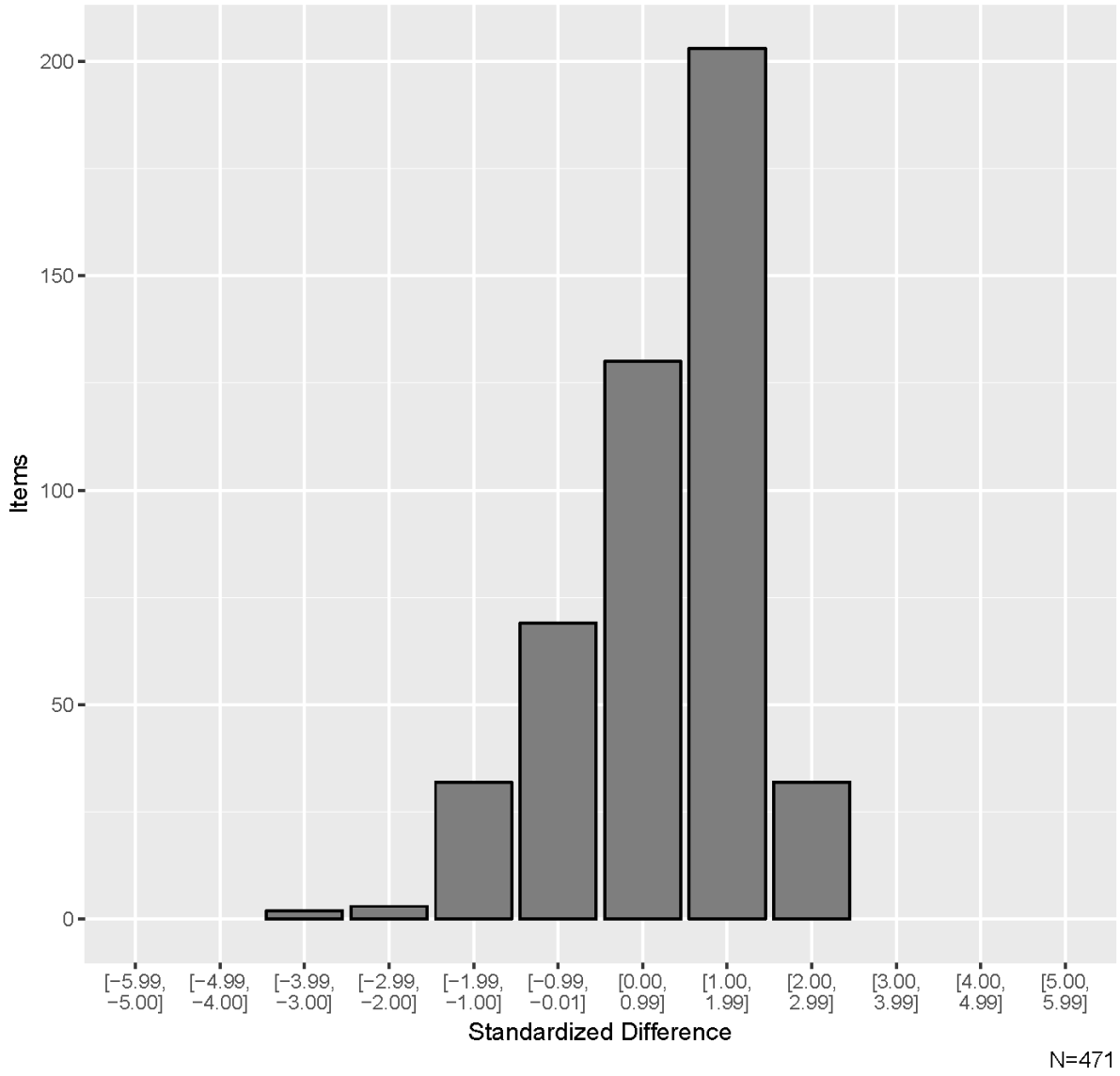


Figure 9. Standardized difference *z* scores for 2015-2016 ELA items field-tested during spring window.

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

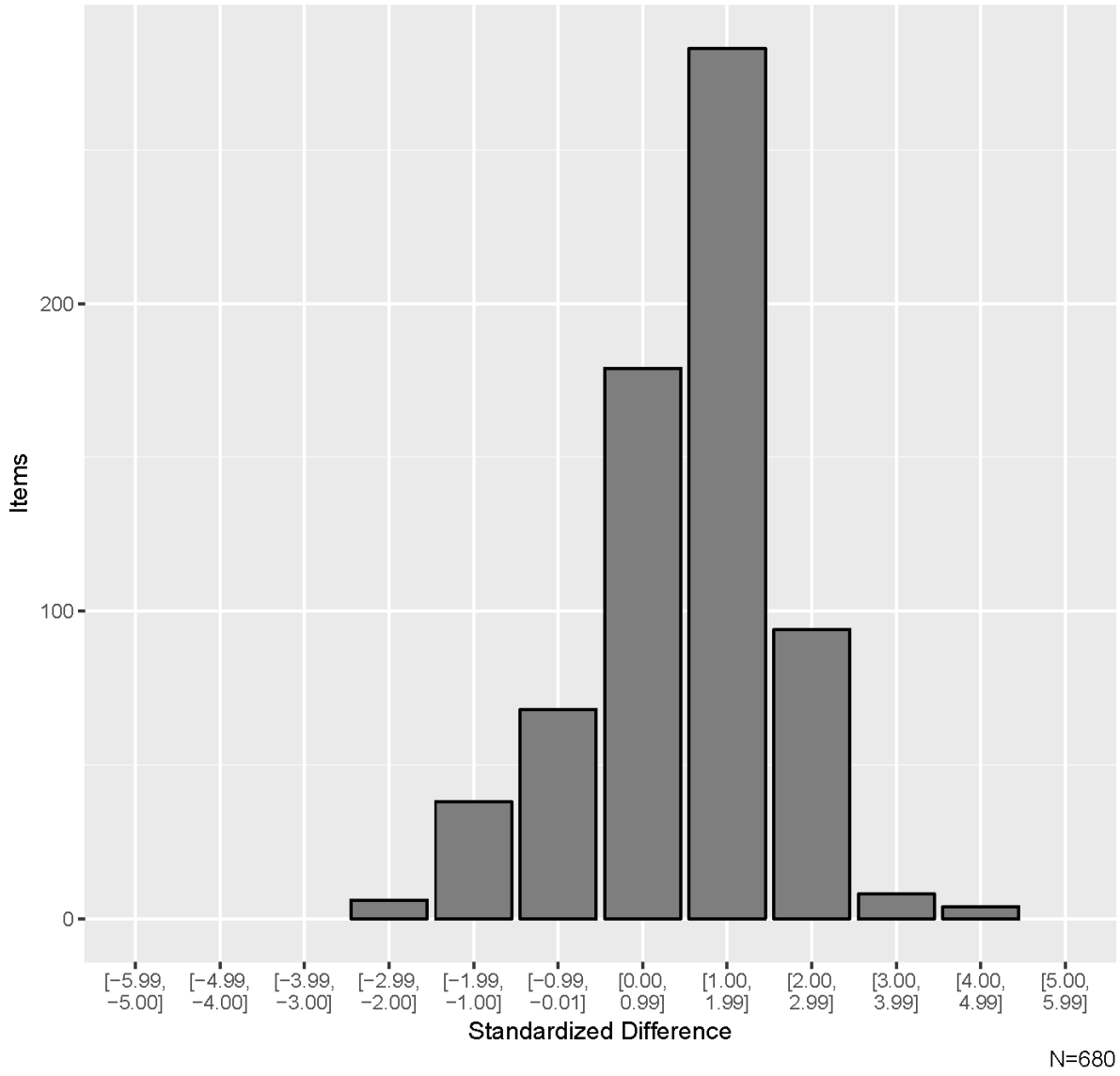


Figure 10. Standardized difference z scores for 2015-2016 mathematics items field-tested during spring window.

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

III.4.B.ii. Item Data Review Decisions

Using the same procedures from prior field-test windows, test-development teams for each content area made four types of item-level decisions as they reviewed field-test items flagged for either a *p* value or standardized difference value beyond the threshold.

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

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

Common reasons for flagging an item for modification included items that were 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, reviewers looked at all items rated as 3 or 4 within the testlet to help determine whether the testlet would be retained or rejected. Here, the test-development 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. The entire testlet was rejected if the test-development team determined the flagged items could not be adequately revised.

III.4.B.iii. Results of Item Analysis and Content-Team Review

A total of 33 ELA items and 157 mathematics items were flagged due to their *p* values and/or standardized difference values. Test-development teams reviewed all flagged items and their context within the testlet to identify possible reasons for the flag and to determine whether an edit was likely to resolve the issue.

Table 13 and Table 14 provides the content team's counts for acceptance, revision, and rejection by content area for all field-test flagged items, for ELA and mathematics respectively. In ELA, three items and their associated testlets were rejected, compared to 11 items in mathematics. Items were rejected when test-development team review indicated (a) the item had more than one correct response option; (b) the text used for ELA testlets was outdated; or (c) the images, materials, or item format for mathematics testlets were outdated.

Table 13. *ELA Content Team Response to Item Flags for Each Grade*

Grade	Flagged Item Count	Accept		Revise		Reject	
		n	%	*n*	%	*n*	%
3	3	3	100.0	0	0.0	0	0.0
4	3	3	100.0	0	0.0	0	0.0
5	1	0	0.0	0	0.0	1	100.0
6	4	4	100.0	0	0.0	0	0.0
7	2	2	100.0	0	0.0	0	0.0
8	3	3	100.0	0	0.0	0	0.0
9	4	4	87.5	0	0.0	0	0.0
10	12	10	83.3	0	0.0	2	16.7
11	1	1	100.0	0	0.0	0	0.0
English 2	0	0	0.0	0	0.0	0	0.0
English 3	0	0	0.0	0	0.0	0	0.0

Table 14. *Mathematics Content Team Response to Item Flags for Each Grade*

Grade	Flagged Item Count	Accept		Revise		Reject	
		n	%	*n*	%	*n*	%
3	3	3	100.0	0	0.0	0	0.0
4	7	7	100.0	0	0.0	0	0.0
5	11	9	81.8	0	0.0	2	18.2
6	13	12	92.3	0	0.0	1	7.7
7	24	20	83.3	0	0.0	4	16.7
8	8	5	62.5	3	37.5	0	0.0
9	24	24	100.0	0	0.0	0	0.0
10	18	17	94.4	1	5.6	0	0.0
11	29	27	93.1	2	6.9	0	0.0

Grade	Flagged Item Count	Accept		Revise		Reject	
		n	%	*n*	%	*n*	%
Algebra I	20	16	80.0	0	0.0	4	20.0
Algebra II	0	0	0.0	0	0.0	0	0.0
Geometry	0	0	0.0	0	0.0	0	0.0

Decisions to recommend testlets for retirement occur on an annual basis following the completion of the operational testing year. In instances where multiple testlets are available for an EE and linkage-level combination, test-development teams may recommend the retirement of testlets that perform poorly compared to others measuring the same EE and linkage level. The retirement process will begin following the 2016–2017 academic year.

IV. TEST ADMINISTRATION

Chapter IV of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016) describes general test administration and monitoring procedures. This chapter describes procedures and data collected in 2015–2016, including a summary of adaptive routing, administration errors, Personal Needs and Preferences (PNP) profile selections, and teacher survey responses regarding user experience and accessibility.

Overall, administration features remained consistent with the prior year's implementation, including the availability of optional instructionally embedded testlets, spring administration of testlets, the use of adaptive delivery during the spring window, and the availability of accessibility supports. No changes were made to the assessment blueprints or testlet construction during the 2015–2016 administration year.

For a complete description of test administration for Dynamic Learning Maps® (DLM®) assessments, including information on administration time, available resources and materials, and monitoring assessment administration, see the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016).

IV.1. Overview of Key Administration Features

This section describes updates to the key, overarching features of DLM test administration for 2015–2016. For a complete description of key administration features, including information on assessment delivery, the Kansas Interactive Testing Engine (KITE®), and linkage-level selection, see Chapter IV of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016) and the *Test Administration Manual 2015–2016* (DLM Consortium, 2015).

IV.1.A. Test Windows

During the consortium-wide spring testing window, which occurred between March 16 and June 10, 2016, all students were assessed on each Essential Element (EE) on the blueprint. Each state set its own testing window within the larger consortium spring window.

IV.1.B. Special Circumstance Codes

In 2015–2016, state partners were given the option to allow entry of special circumstance codes in Educator Portal, the administrative application for staff and educators to manage student data, complete required test administrator training, retrieve resources needed for each assigned testlet, and retrieve reports. For states implementing the use of special circumstance codes, state partners defined the list of allowable codes, including correspondence of the Common Education Data Standards codes to state-specific codes and definitions.

Special circumstance codes were available for entry in the event that a student could not participate in a testlet that generates a performance level used for federal and state

accountability. Special circumstance codes could be entered in Educator Portal to provide an explanation for why a student was not tested.

The special circumstance fields were located in Educator Portal on the same screen where the Testlet Information Page (TIP) was accessed and included descriptive terms such as *medical waiver* or *parental refusal*. Only educators with the role of district assessment coordinator, building test coordinator, or state assessment administrator had permission to choose the code. DLM staff recommended that the special circumstance code not be entered until late in the state's spring testing window to allow adequate time for testing to occur but before the window closed. Codes needed to be entered once per content area associated with the first testlet delivered or as needed when test administration could no longer occur due to a special circumstance. Data files delivered to state partners summarizing special circumstance codes are described in Chapter VII.

IV.2. Implementation Evidence

This section describes evidence collected for 2015–2016 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.

IV.2.A. Adaptive Delivery

During the spring 2016 test administration, the English language arts (ELA) and mathematics assessments were adaptive between testlets, following the same routing rules applied in 2014–2015. That is, the linkage level associated with the next testlet a student received was based on the student's performance on the most recently administered testlet, 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 the student responded correctly to at least 80% of the items measuring the previously tested EE. If the previous testlet was at the highest linkage level (i.e., Successor), the student remained at that level.
- The system adapted down one linkage level if the student responded correctly to less than 35% of the items measuring the previously tested EE. If the previous testlet was at the lowest linkage level (i.e., Initial Precursor), the student remained at that level.
- Testlets remained at the same linkage level if the student responded correctly to between 35% and 80% of the items on the previously tested EE.
- When a testlet contained items aligned to more than one EE,[4] a percentage of items answered correctly was calculated for each group of items measuring the same EE. The

[4]This rule applied only to testlets in the year-end and End-of-Instruction models.

minimum of these values was then used to determine the next linkage level, based on the above thresholds.

The linkage level of the first testlet assigned to a student was based on First Contact survey responses. Table 15 shows the correspondence between the First Contact complexity bands and first assigned linkage levels.

Table 15. *Correspondence of Complexity Bands and Linkage Levels*

First Contact Complexity Band	Linkage Level
Foundational	Initial Precursor
1	Distal Precursor
2	Proximal Precursor
3	Target

For a complete description of adaptive delivery procedures, see Chapter IV of the *2014-15 Technical Manual – Year-End Model* (DLM Consortium, 2016).

Following the spring 2016 administration, analyses were conducted to determine 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 to second testlet administered for students within a grade, content area, and complexity band. The aggregated results can be seen in Table 16 and Table 17.

Overall, results were similar to those found in 2014–2015. For the majority of students across all grades who were assigned to the Foundational complexity band by the First Contact survey, testlets did not adapt to a higher linkage level after the first assigned testlet. Consistent patterns were not as apparent for students who were assigned to Complexity Band 1 and Complexity Band 2. Generally, there was a more even split between students assigned at Band 1 whose testlets did not adapt a linkage level and 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 across grade and content area. That is, for some combinations of grade and content area, 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 Complexity Band 3, the linkage level of the assessment between the first and second testlets either did not adapt or adapted up a level.

The 2015–2016 results build on earlier findings from the pilot study and 2014–2015 operational assessment administration (see Chapter III and Chapter IV of the *2014-2015 Technical Manual – Year-End Model*, respectively) and suggest that the First Contact survey complexity-band assignment was an effective tool for assigning students content at appropriate linkage levels. Results also indicated that linkage levels of students assigned to Complexity Band 2 are more variable with respect to the direction in which students 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 16. *Adaptation of Linkage Levels Between First and Second English Language Arts Testlets* (n = *70,214)*

Grade	Foundational		Band 1			Band 2			Band 3		
	Adapted Up (%)	Did Not Adapt (%)	Adapted Up (%)	Did Not Adapt (%)	Adapted Down (%)	Adapted Up (%)	Did Not Adapt (%)	Adapted Down (%)	Adapted Up (%)	Did Not Adapt (%)	Adapted Down (%)
3	19.4	80.6	32.4	39.6	28.0	75.4	12.4	12.2	92.9	3.1	4.1
4	31.4	68.6	16.3	44.5	39.1	37.3	38.8	23.9	53.2	45.3	1.5
5	22.3	77.7	23.0	31.2	45.8	60.1	28.1	11.7	78.9	18.8	2.3
6	17.6	82.4	22.2	10.5	67.3	41.4	22.5	36.1	29.7	33.2	37.1
7	18.9	81.1	17.1	30.3	52.6	28.9	36.4	34.8	28.7	28.5	42.8
8	36.5	63.5	26.6	44.5	28.9	48.0	41.0	11.1	81.1	14.7	4.2
9	15.9	84.1	18.5	10.9	70.6	32.0	15.4	52.6	43.1	10.8	46.1
10	17.1	82.9	10.1	32.4	57.5	16.6	57.3	26.0	49.0	41.0	10.0
11	15.6	84.4	3.7	25.3	71.0	24.2	43.2	32.6	36.2	47.2	16.6
English 2	29.1	70.9	33.8	38.0	28.3	18.2	65.0	16.8	59.6	34.4	6.0
English 3	61.0	39.0	45.5	25.5	29.1	66.4	28.7	4.9	57.3	30.5	12.2

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

Table 17. *Adaptation of Linkage Levels Between the First and Second Mathematics Testlet* (n = *70,525)*

Grade	Foundational		Band 1			Band 2			Band 3		
	Adapted Up (%)	Did Not Adapt (%)	Adapted Up (%)	Did Not Adapt (%)	Adapted Down (%)	Adapted Up (%)	Did Not Adapt (%)	Adapted Down (%)	Adapted Up (%)	Did Not Adapt (%)	Adapted Down (%)
3	6.8	93.2	6.2	31.2	62.6	15.5	27.2	57.3	8.9	56.9	34.2
4	12.9	87.1	49.3	13.7	37.0	60.5	17.6	21.9	47.1	24.0	28.9
5	23.8	76.2	10.2	16.3	73.6	15.0	8.9	76.1	54.5	7.1	38.5
6	13.9	86.1	12.9	25.2	61.9	16.4	33.7	49.8	28.3	38.6	33.1
7	9.7	90.3	7.7	17.1	75.1	31.1	35.3	33.6	38.2	9.7	52.1
8	19.4	80.6	13.8	6.2	80.0	3.0	39.1	57.9	17.4	25.8	56.7
9	21.6	78.4	8.2	30.7	61.1	9.0	50.2	40.8	19.5	49.6	30.9
10	14.7	85.3	0.5	34.8	64.7	2.1	20.3	77.5	20.4	46.1	33.5
11	15.7	84.3	1.3	47.3	51.4	1.5	24.0	74.5	9.9	49.5	40.6
Algebra 1	36.0	64.0	46.4	26.5	27.1	75.4	13.3	11.3	61.2	18.9	19.9
Algebra 2	55.6	44.4	58.3	41.7	0.0	44.0	12.0	44.0	0.0	0.0	100.0
Geometry	70.0	30.0	88.1	9.0	3.0	78.1	21.9	0.0	59.3	31.5	9.3

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

IV.2.B. Administration Errors

Monitoring of testlet assignment during the 2015–2016 operational assessment windows uncovered several incidents that affected test assignment to students. These incidents included routing errors, in which students may have received a testlet for the incorrect linkage level, and scoring errors, which, because routing thresholds are based on the percentage correct in a testlet, may have indirectly affected routing to subsequent testlets. Scoring errors were corrected prior to calculation of summative results.

Table 18 provides a summary of the number of students affected by each of the incidents, as delivered to states in the Incident File (see Chapter VII of this manual for more information). The most frequent error was a potential misrouting caused by an incorrectly specified EE for an item. For Incident Codes 1 and 2 (i.e., misrouting due to local caching server use and missing responses not scored as incorrect, respectively), states were provided with lists of students affected and given the option to revert each student's assessment back to the end of the last correctly completed testlet (i.e., the point at which routing failed) and have the students complete the remaining testlets as intended. Remaining issues (e.g., incorrectly scored items) were corrected in the system upon discovery of the incident. Overall, the administration incidents affected between less than 0.01% and 9.64% of students.

Table 18. *Number of Students Affected by Each 2016 Incident, Year-End Model (*n = *75,086)*

Incident Code	Incident Description	*n*	%
1	Potential misrouting due to use of the local caching server.	16	0.02
2	Potential misrouting due to missing responses not scoring as incorrect.	1,112	1.48
3	Potential misrouting due to an item with an incorrect key.	519	0.69
4	Potential misrouting due to an item with multiple correct keys.	400	0.53
5	Potential misrouting due to mis-specified Essential Element.	7,237	9.64
7	BVI test form administered to non-BVI student.	16	0.02
8	Non-BVI test form administered to a BVI student.	1	<0.01
9	Potential misrouting due to simultaneous testing on multiple devices.	1	<0.01

Note. BVI = Blind and visually impaired.

Additional details about the eight incidents are described in Table 19.

Table 19. *Incident Summary for 2015–2016 Operational Testing, Year-End Model*

#	Issue	Type	Summary
1	Potential misrouting due to use of the local caching server	Technology: Administration	Use of a local caching server prevented transmission of item responses in real time. Thus, when a student testing on a local caching server submits responses, a percentage correct could not be calculated. In the system, the percentage correct would default to 0, causing the student to always adapt down, regardless of performance on the testlet.
2	Potential misrouting due to missing responses not scoring as incorrect	Technology: Scoring	Items left blank on the assessment are scored as incorrect. However, when calculating percentage correct for adaptation, missing responses were omitted, rather than scored as 0. Thus, the calculated percentage correct did not always have the correct denominator, leading to incorrect adaptations.
3	Potential misrouting due to an item with an incorrect key	Assessment: Content	One End-of-Instruction item was marked in the system with an incorrect key, causing students who provided a correct response to be scored as incorrect and vice versa. To solve this problem, DLM psychometric staff developed and QC'd a manual scoring script to ensure scoring was accurate for score reporting for all students responding to these items prior to the fix. However, the system score was used to determine routing to a subsequent testlet during the operational window.
4	Potential misrouting due to an item with multiple correct keys	Assessment: Content	For two items, distractor response options were correct but not keyed, causing student responses to be mistakenly marked incorrect. To solve this problem, DLM psychometric staff developed and QC'd a manual scoring script to ensure scoring was accurate for score reporting for all students responding to these items prior to the fix. However, the system score was used to

#	Issue	Type	Summary
			determine routing to a subsequent testlet during the operational window.
5	Potential misrouting due to mis-specified Essential Element	Assessment: Content	Five items were not assigned to the correct EE. On Year-End model testlets, the percentage correct is calculated for each EE assessed, and adaptations are determined by the lowest of those percentage-correct values. Thus, incorrectly specified EEs could result in incorrect percentage-correct values and incorrect adaptations. To solve this problem, DLM psychometric staff developed and QC'd a manual scoring script to ensure the EEs were correct for all students responding to these items prior to the fix. However, the system EE was used to determine routing to a subsequent testlet during the operational window.
6	BVI test form administered to non-BVI student	Technology: Enrollment	Testlets with accessibility supports intended for blind or visually impaired students were assigned to students who did not require those supports.
7	Non-BVI test form administered to a BVI student	Technology: Enrollment	Students requiring accessibility supports for blindness or visual impairment were assigned testlets that did not include those supports, when testlets with those supports were available in the system.
8	Potential misrouting due to simultaneous testing on multiple devices	Technology: Scoring	Students began testing on one device and then switch to another device without closing the session on the original device. Following completion of the testlet on the second device, the next testlet would be assigned. When students returned to the original device and closed their testing session, response data from the first testlet were erased because the system noted the test session for testlet was ended prior to completion. Because the next testlet had already been assigned, testing was allowed to continue despite the earlier testlet's incomplete status.

Note. BVI = Blind and visually impaired.

As in 2014–2015, the Incident File was delivered to state partners with the General Research File (GRF; see Chapter VII for more information), providing a list of all students affected by each

issue. States could use the Incident File and their own accountability policies and practices to determine possible invalidation of student records. All issues were corrected for subsequent administration. Testlet assignment will continue to be monitored in 2016–2017 to track any potential incidents and report them to state partners.

IV.2.C. User Experience with Assessment Administration and KITE System

User experience with the 2015–2016 assessments was evaluated through a spring 2016 survey disseminated to teachers who had administered a DLM assessment during the 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 Studies). For responses to the 2014–2015 version of the survey, see Chapter IV and Chapter IX of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016).

A total of 2,320 teachers from states participating in the DLM assessment responded to the survey (estimated response rate of 11.5%). Most respondents reported having assessed a relatively small number of students during the testing window; 61.7% reported assessing four or fewer students. The self-reported numbers of students assessed per teacher for the year-end assessment model are summarized in Table 20.

Table 20. *Self-Reported Number of Students Assessed* (n = *2,320)*

Reported Number of Students Assessed	*n*	%
1	549	23.7
2	403	17.4
3	273	11.8
4	205	8.8
5	222	9.6
6	207	8.9
7	115	5.0
8	88	3.8
9	81	3.5
10	49	2.1
11	28	1.2
12	41	1.8
13	15	0.6
14	2	0.1
≥15	42	1.8

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

IV.2.C.i. Educator Experience

Respondents were asked to reflect on their experiences with the assessments and their comfort level and knowledge in administering them. Most questions required respondents to use a 4-point scale: *Strongly Disagree*, *Disagree*, *Agree*, or *Strongly Agree*. Responses are summarized in Table 21. The first two questions (regarding comfort level with the administration of computer-administered and teacher-administered testlets) were displayed only if respondents previously stated that they had administered the corresponding testlet type.

Table 21. *Teacher Response Regarding Test Administration*

Statement	SD		D		A		SA		A+SA	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
Confidence in ability to deliver computer-administered testlets.	45	3.2	86	6.0	638	44.7	657	46.1	1,295	90.8
Confidence in ability to deliver teacher-administered testlets.	28	2.5	80	7.2	549	49.1	460	41.2	1,009	90.3
Test administrator training prepared respondent for responsibilities of test administrator.	219	10.7	393	19.1	1,135	55.3	306	14.9	1,441	70.2
Respondent knew how to use accessibility features, allowable supports, and options for flexibility.	96	4.7	263	12.8	1,371	66.7	325	15.8	1,696	82.5
Testlet Information Pages helped respondent to deliver the testlets.	192	9.4	447	21.8	1,136	55.4	277	13.5	1,413	68.9

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

Teachers responded that they were confident in administering either kind of testlet; 90.8% responded Agree or Strongly Agree for computer-administered testlets and 90.3% responded 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; 70.2% responded Agree or Strongly Agree. Most teachers also said that they knew how to use accessibility supports, allowable supports, and options for flexibility (82.5%) and that the TIPs helped them to deliver the testlets (68.9%).

IV.2.C.ii. KITE System

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

KITE Client is used for the administration of DLM testlets. Teachers were asked to consider their experiences with KITE Client and to evaluate the ease of each step using a 5-point scale: *Very Hard*, *Somewhat Hard*, *Neither Hard Nor Easy*, *Somewhat Easy*, or *Very Easy*. Table 22 summarizes teacher responses.

Table 22. *Ease of Using KITE Client*

Statement	VH		SH		N		SE		VE		SE+VE	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Enter the site	49	2.5	154	7.7	366	18.4	640	32.1	784	39.3	1,424	71.4
Navigate within a testlet	41	2.1	112	5.6	330	16.6	654	32.8	854	42.9	1,508	75.7
Submit a completed testlet	33	1.7	69	3.5	298	15.0	568	28.5	1,022	51.4	1,590	79.9
Administer testlets on various devices	78	4.0	143	7.3	612	31.1	559	28.4	573	29.2	1,132	57.6

Note. VH = very hard; SH = somewhat hard; N = neither hard nor easy; SE = somewhat easy; VE = very easy; SE+VE = somewhat easy and very easy.

Respondents found it to be either Somewhat Easy or Very Easy to enter the site (71.4%), navigate within a testlet (75.7%), submit a completed testlet (79.9%), and administer testlets on various devices (57.6%). Open-ended survey responses indicated issues with display on some devices, including scrolling and glitches in the display of testlet response options. These issues were forwarded to the technology team for evaluation ahead of the 2016–2017 administration.

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, using the same scale used with KITE Client; the data are summarized in Table 23. Overall, respondents' feedback was mixed: Fewer teachers than expected found it somewhat easy or very easy to navigate the site (43.6%), to enter PNP and First Contact information (57.3%), to manage student data (43.5%), and to manage their own accounts (48.5%).

Table 23. *Ease of Using Educator Portal*

Statement	VH		SH		N		SE		VE		SE+VE	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Navigate the site	144	7.0	452	22.0	52	27.4	593	28.9	302	14.7	895	43.6
Enter PNP and First Contact information	87	4.2	286	14.0	503	24.5	783	38.2	391	19.1	1,174	57.3
Manage student data	170	8.3	448	21.8	543	26.4	609	29.7	283	13.8	892	43.5
Manage your account	116	5.6	336	16.4	605	29.5	687	33.4	310	15.1	997	48.5

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

Open-ended survey responses indicated Educator Portal was not very user-friendly. Respondents indicated they had challenges navigating due to unclear terminology and labeling in the system, and they noted that information was spread across multiple screens, requiring substantial forward and backward clicking by the user. Suggestions for improvement included adding direct links on the main page, alternative organization mechanisms for TIPs, and providing a to-date summary of test administration by student. This feedback informed technology development plans for the 2016–2017 academic year.

Open-ended survey feedback indicated teachers thought there were too many dropdown boxes to select EEs when creating instructional plans; teachers thought it would be useful to be able to set up instructional plans for multiple students at once. Additionally, teachers would prefer that all information fit on one screen, rather than scrolling or navigating to create instructional plans. This feedback was incorporated into changes made to the Instructional Tools Interface for the 2016–2017 academic year.

Finally, respondents were asked to rate their overall experience with KITE Client and Educator Portal on a 4-point scale: *Poor*, *Fair*, *Good*, and *Excellent*. Results are summarized in Table 24. The majority of respondents reported a positive experience with KITE Client. A total of 65.0% of respondents rated their experience as Good or Excellent, while 53.7% rated their overall experience with Educator Portal to be Good or Excellent.

Table 24. *Overall Experience with KITE Client and Educator Portal*

Interface	Poor		Fair		Good		Excellent	
	n	%	*n*	%	*n*	%	*n*	%
KITE Client	202	10.2	490	24.7	873	44.0	417	21.0
Educator Portal	272	13.6	650	32.6	832	41.7	239	12.0

Overall feedback from teachers indicated that KITE Client was easy to navigate and user-friendly. Additionally, teachers provided useful feedback for improvements to Educator Portal that will be considered for subsequent technology development to improve user experience for 2016–2017 and beyond.

IV.2.C.iii. Accessibility

Accessibility supports provided in 2015–2016 were the same as those available in 2014–2015. Accessibility guidance provided by the DLM system distinguishes between accessibility 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 25 shows selection rates for three categories of PNP supports, sorted by rate of use within each category. For a complete description of the available accessibility supports, see Chapter IV of the *2014-2015 Technical Manual – Year-End Model* (DLM Consortium, 2016). Generally, the percentage of students for whom supports were selected in 2015–2016 was similar to that observed in 2014–2015.

Table 25. *Personal Needs and Preferences Profile (PNP) Supports Selected for Students* (N = *66,211)*

Supports	*n*	%
Supports activated by PNP		
Read aloud (TTS)	641	1.0
Magnification	5,175	7.8
Color contrast	3,503	5.3
Overlay color	3,539	5.4
Invert color choice	2,770	4.2
Supports requiring additional tools/materials		
Individualized manipulatives	25,613	38.7
Calculator	16,262	24.6
Single-switch system	3,880	5.9
Alternate form – visual impairment	1,414	2.1
Two-switch system	822	1.2
Uncontracted braille	126	0.2
Supports provided outside the system		
Human read aloud	57,035	86.1
Test administration enters responses for students	30,311	45.8
Partner-assisted scanning	4,981	7.5
Sign interpretation	1,121	1.7
Language translation	1,243	1.9

Table 26 describes teacher responses to survey items about the accessibility supports used during administration. Teachers were asked to respond to three items using a 4-point Likert-type scale (*Strongly Disagree*, *Disagree*, *Agree*, or *Strongly Agree*). The majority of teachers agreed that students were able to effectively use accessibility supports (74.7%), that accessibility supports were similar to ones the student used for instruction (70.3%), and that allowable options for flexibility were necessary to meet students' needs when administering the assessment (65.9%). These data support the conclusions that the accessibility supports of the DLM alternate assessment were effectively used by students, emulated accessibility supports used during instruction, and met student needs for test administration. Additional data will be collected during the spring 2017 survey to determine whether results improve over time.

Table 26. *Teacher Report of Student Accessibility Experience*

Statement	SD		D		A		SA		A+SA	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
Student was able to effectively use accessibility features.	386	12.3	410	13.0	1,242	39.5	1,105	35.2	2,347	74.7
Accessibility features were similar to ones student uses for instruction.	288	9.3	635	20.4	1,862	59.9	325	10.5	2,187	70.4
Allowable options for flexibility were needed when administering test to meet student needs.	332	9.0	929	25.2	1,736	47.0	696	18.8	2,432	65.8

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

IV.3. CONCLUSION

During the 2015–2016 academic year, the DLM system was available for optional instructionally embedded use and during the operational spring window. Entry of special circumstance codes was available in Educator Portal to indicate why students may not have completed testing in a content area. Implementation evidence was collected in the forms of testlet adaptation analyses, a summary of students affected by incidents during operational testing, and teacher survey responses regarding test administration and accessibility. Results indicated that teachers felt confident administering testlets in the system, found KITE Client easy to use, but thought Educator Portal posed more challenges. This feedback resulted in changes for the 2016–2017 year to make the site easier to use, including improvements to data management features and streamlining of the process for creating instructional plans.

V. MODELING

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

V.1. PSYCHOMETRIC BACKGROUND

Learning map models, which are the networks of sequenced learning targets, are at the core of the DLM assessments in English language arts (ELA) and mathematics. In general, a learning map model is a collection of skills to be mastered that are linked 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. Stated in the vocabulary of traditional psychometric methods, a learning map model defines a large set of discrete latent variables indicating students' learning status on key skills and concepts relevant to a large content domain, as well as a series of pathways indicating which topics (represented by latent variables) are prerequisites for learning other topics.

Because of the underlying map structure and the goal to provide more fine-grained information beyond a single raw or scale score value when reporting student results, the assessment system provides a profile of skill mastery to summarize student performance. This profile is created using a form of diagnostic classification modeling, which draws upon research in cognition and learning theory to provide feedback about student performance. Diagnostic classification models (DCMs) are confirmatory, latent class models that characterize the relationship of observed responses to a set of categorical latent variables (e.g., Rupp & Templin, 2008; Rupp, Templin, & Henson, 2010). DCMs are also known as cognitive diagnosis models (e.g., Leighton & Gierl, 2007) or multiple classification latent class models (Maris, 1999) and are mathematically equivalent to Bayesian networks (e.g., Almond, Mislevy, Steinberg, Yan, & Williamson, 2015; Mislevy & Gitomer, 1995; Pearl, 1988). This is the main difference from more traditional psychometric models (e.g., item response theory), which model a single, continuous latent variable. DCMs provide information about student mastery on multiple latent variables or skills of interest.

DCMs have primarily been used in educational measurement settings in which detailed information about test-takers' skills is of interest, as in assessing mathematics (e.g., Bradshaw, Izsák, Templin, & Jacobson, 2014), reading (e.g., Templin & Bradshaw, 2014), and science (e.g., Templin & Henson, 2008). To provide detailed profiles of student mastery of the skills, or attributes, measured by the assessment, DCMs require the specification of an item-by-attribute Q-matrix, indicating the attributes measured by each item. In general, for a given item, i, the Q-

matrix vector is represented as $q_i = [q_{i1}, q_{i2}, \ldots, q_{iA}]$. Similar to a factor pattern matrix in a confirmatory factor model, Q-matrix indicators are binary; either the item measures an attribute ($q_{ia} = 1$) or it does not ($q_{ia} = 0$).

For each item, there is a set of conditional item-response probabilities that corresponds to a student's possible mastery patterns. When an item measures a single binary attribute, only two statuses are possible for any examinee: a master of the attribute or a nonmaster of the attribute.

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

V.2. Essential Elements and Linkage Levels

Because the primary goal of the DLM assessments is to measure what students with the most significant cognitive disabilities know and can do, alternate grade-level expectations called Essential Elements (EEs) were created to provide students in the population access to the general education grade-level academic content. See Chapter II of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016) for a complete description. Each EE has an associated set of linkage levels that are ordered by increasing complexity. There are five linkage levels for each EE: Initial Precursor, Distal Precursor, Proximal Precursor, Target, and Successor.

V.3. Overview of DLM Modeling Approach

Many statistical models are available for estimating the probability of mastery for attributes in a DCM. The statistical model used to determine the probability of mastery for each linkage level for DLM assessments is latent class analysis, which provides a general statistical framework for obtaining probabilities of class membership for each measured attribute (Macready & Dayton, 1977). Student mastery statuses for each linkage level are obtained from an Expectation–Maximization procedure that contributes to an overall profile of mastery.

V.3.A. DLM Model Specification

Due to the administration design, where overlapping data from students taking testlets at multiple linkage levels within an EE were unavailable, simultaneous calibration of all linkage levels within an EE was not possible. Instead, each linkage level was calibrated separately for each EE using separate latent class analyses. Additionally, because items were developed to a precise cognitive specification, all master and nonmaster probability parameters for items measuring a linkage level were assumed to be equal. That is, all items were assumed to be fungible, or exchangeable, within a linkage level. As such, each class (i.e., master or nonmaster) has a single probability of responding correctly to all items measuring the linkage level, as depicted in Table 27. Similarly, for each item measuring the linkage level, a student has the same probability of providing a correct response. Chapter III details item-review procedures

intended to support the fungibility assumption. Chapter X discusses future studies intended to continue evaluating the fungibility assumption.

Table 27. *Fungible Item Parameters for Items Measuring a Single Linkage Level*

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

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

The DLM scoring model for the 2015–2016 administration follows. Each linkage level within each EE was considered the latent variable to be measured (i.e., the attribute). Using latent class analysis, a probability of mastery on a scale of 0 to 1 was calculated for each linkage level within each EE. Students were then classified into one of two classes for each linkage level of each EE: master or nonmaster. As described in Chapter VI of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016), a posterior probability of at least .8 was required for mastery classification.

All items in a linkage level were assumed to measure that linkage level, meaning the Q-matrix for the linkage level was a column of ones. As such, each item measured one latent variable, resulting in two parameters per item: (a) the probability of answering the item correctly for examinees who have not mastered the linkage level (i.e., the reference group) and (b) the probability of answering the item correctly for examinees who have mastered the linkage level. According to the assumption of item fungibility, a single set of probabilities was estimated for all items within a linkage level. Finally, a structural parameter was also estimated, which was the proportion of masters for the linkage level (the analogous map parameter). In total, three parameters per linkage level are specified in the DLM scoring model: a fungible probability for nonmasters, a fungible probability for masters, and the proportion of masters. An explanation of the full model is provided below.

V.3.B. Model Calibration

Across all grades and content areas, there were 242 EEs, each with five linkage levels, resulting in a total of $242 \times 5 = 1{,}210$ separate calibration models. Each separate calibration included all items available for the EE and linkage level. Each model was estimated using marginal maximum likelihood using a program that was developed in the R Project for Statistical Computing (R Core Team, 2013).

Latent class analysis was used to obtain the posterior probabilities of mastery, or the likelihood a student mastered the skill being measured. As such, it did not provide scale score values, but rather a probability, on a scale of 0 to 1, representing the certainty of skill mastery. Values closer to 0 or 1 represent greater certainty of nonmastery or mastery, respectively, whereas values closer to .5 represent maximum uncertainty.

A latent class analysis was conducted for each linkage level for each EE. The calibration of the model and final scoring procedure used an Expectation–Maximization algorithm. If the probability of a correct response on item i for a person in class j is defined as π_{ij}, the likelihood of a given response pattern for an individual h over J classes and I items is defined as:

$$f(\mathbf{X}_h) = \sum_{j=1}^{J} \eta_j \prod_{i=1}^{I} \pi_{ij}^{x_i} \left(1 - \pi_{ij}\right)^{1 - x_i}$$

This likelihood (or the log-likelihood, if the log is taken) can be maximized by an Expectation–Maximization algorithm using three estimating equations. The Expectation step estimates the posterior probability for each student. It is expressed with the following formula (using notation consistent with Bartholomew, Knott, & Moustaki, 2011),

$$h(j|\boldsymbol{X}_h) = \frac{\eta_j \prod_{i=1}^{I} \pi_{ij}^{x_{ih}} \left(1 - \pi_{ij}\right)^{1 - x_{ih}}}{f(\boldsymbol{X}_h)},$$

where $h(j \mid \boldsymbol{X}_h)$ represents the posterior probability of a person's class membership given their responses. The numerator is the person's probability of item responses for a given class, $\prod_{i=1}^{I} \pi_{ij}^{x_{ih}} \left(1 - \pi_{ij}\right)^{1 - x_{ih}}$, times the probability of membership in that given class, η_j. The denominator ($\boldsymbol{X}_h$) is the probability of that person's item responses, or the full likelihood, defined above.

The Maximization step estimates the model parameters, including the item parameter, π_{ij}, for each item i and class j, and the proportion of people in a given class, η_j.

The item parameter was estimated using the following formula,

$$\pi_{ij} = \frac{\sum_{h=1}^{N} x_{ih} h(j|\boldsymbol{X}_h)}{N \eta_j},$$

where $h(j|\boldsymbol{X}_h)$ represents the posterior probability of a person's class membership given their responses, which was estimated during the Expectation step. The numerator is the sum of the item responses across all respondents, x_{ih}, weighted by the posterior probability of each respondent being in that class. The denominator is the number of respondents, N, times the proportion of people estimated to be in the class, j. Thus, the item parameters can be thought of

as item p values, conditional on group membership. Because the assessment system assumed a fungible item model, all items measuring a linkage level had the same parameter for each class.

The parameter η_j was estimated using the following formula,

$$\eta_j = \frac{\sum_{h=1}^{N} h(j|\boldsymbol{X}_h)}{N},$$

where $h(j|\boldsymbol{X}_h)$ represents the posterior probability of a person's class membership given their responses, which was estimated during the Expectation step. The numerator is the sum of the class membership probabilities across all respondents, and the denominator, N, is the number of respondents.

Model calibration in 2016 occurred in June and incorporated operational item responses from the 2015–2016 testing window. The model was calibrated using the Expectation–Maximization algorithm until the convergence criteria, change in log-likelihood to < 0.00001, was met. During the calibration process, initial values of 0.9 and 0.1 for the item parameters were provided for each class, master and nonmaster respectively, to prevent their definitions from switching during estimation. The initial value of η was set to 0.5 for each class.

The final calibrated model parameters from the Maximization step described above were used to run the Expectation step a final time, using all operational item responses obtained during the spring window. This process resulted in the final student posterior probabilities for each linkage level, which were used for scoring.

V.4. DLM Scoring: Mastery Status Assignment

Following calibration, results for each linkage level were combined to determine the highest linkage level mastered for each EE. Although the connections between linkage levels were not modeled empirically, they were used in the scoring procedures. In particular, if the latent class analysis determined a student was determined 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 and order the linkage levels that guided item and testlet development. Chapter III of *2014-2015 Technical Manual – Year-End Model* (DLM Consortium, 2016) provides evidence from the pilot test that supports the ordering of linkage levels. Additional validation studies for this scoring rule are currently underway.

In addition to the calculated posterior probability of mastery, students were able to demonstrate mastery of each EE in two additional ways: (a) correctly answering 80% of all items administered at the linkage level or (b) via the *two-down* scoring rule. The two-down scoring rule was implemented to guard against excessively penalizing students assessed at the highest linkage levels for incorrect responses. For example, students who tested at the Successor level but did not demonstrate mastery were assigned mastery status of two linkage levels lower (Proximal Precursor) to prevent them from being penalized for testing at the highest level and not demonstrating mastery. Students who did not demonstrate mastery at the Initial Precursor or Distal Precursor levels were considered nonmasters of all linkage levels within the EE

because the two-down rule was inapplicable. This scoring method was discussed and determined to be a reasonable approach by the DLM TAC during a conference call on July 21, 2015.

To evaluate the degree to which each mastery assignment rule contributed to students' linkage-level mastery status, the percentage of mastery statuses obtained by each scoring rule was calculated, as shown in Figure 11. Posterior probability was given first priority; if mastery was not demonstrated by meeting the posterior probability threshold, the next two scoring rules were imposed. Between 60% and 80% of mastered linkage levels were derived from the posterior probability obtained from the modeling procedure. The remaining percentage of linkage levels was assigned mastery status by the minimum mastery, or two-down rule, or the percentage-correct rule. These results indicate that the percentage-correct rule likely had strong overlap (but was second in priority) with the posterior probabilities, in that correct responses to all items measuring the linkage level were likely necessary to achieve a posterior probability above the .80 threshold. The percentage-correct rule does, however, provide mastery status in those instances where providing correct responses to all items still resulted in a posterior probability below the mastery threshold.

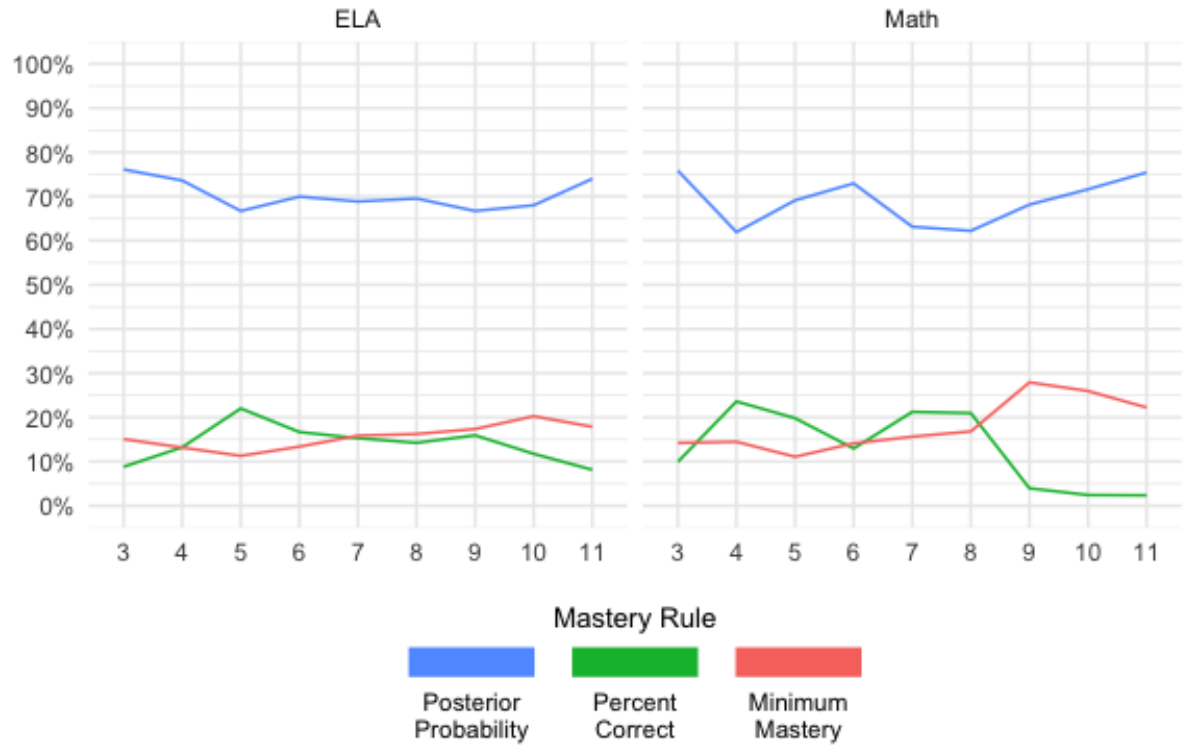


Figure 11. Linkage-level mastery assignment by mastery rule for each content area and grade.

V.5. Conclusion

In summary, the DLM modeling approach makes use of well-established research in the areas of Bayesian inference networks and diagnostic classification modeling to determine student

mastery of skills measured by the assessment. Latent class analyses are conducted for each linkage level of each EE to determine the probability of student mastery. Items within the linkage level are assumed to be fungible, with equivalent item-probability parameters for each class, due to the conceptual approach used to construct DLM testlets. For each linkage level, a mastery threshold of .8 is applied, whereby students with a posterior probability greater than or equal to the cut are deemed masters and students with a posterior probability below the cut are deemed nonmasters. In addition to posterior probabilities of mastery obtained from the model and to ensure students are not overly penalized by the modeling approach, two additional scoring procedures are implemented: percentage correct at the linkage level and the two-down scoring rule. An analysis of the scoring rules indicates most students demonstrate mastery of the linkage level based on posterior probability values obtained from the modeling results.

VI. STANDARD SETTING

The standard-setting process for the Dynamic Learning Maps® (DLM®) Alternate System in English language arts and mathematics derived cut points for placing students into four performance levels from results from the 2014–2015 DLM alternate assessments. For a description of the process, including the development of policy performance level descriptors, the four-day standard-setting meeting, follow-up evaluation of impact data and cut points, and specification of grade- and content-specific performance level descriptors, see Chapter VI of *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016).

VII. ASSESSMENT RESULTS

Chapter VII of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016) describes assessment results for the 2014–2015 academic year, including student participation and performance summaries and an overview of data files and score reports delivered to state partners. This chapter presents 2015–2016 student participation data; final results in terms of the percentage of students at each performance level; and subgroup performance by gender, race, ethnicity, and English language learner (ELL) status for the 2015–2016 administration year. This chapter also reports the distribution of students by the highest linkage level (LL) mastered during 2015–2016. Finally, this chapter describes updates made to Individual Student Score Reports, data files, and quality control procedures during the 2015–2016 operational year. For a complete description of and interpretive guides, see Chapter VII of the *2014-2015 Technical Manual – Year-End Model* (DLM Consortium, 2016).

VII.1. STUDENT PARTICIPATION

The spring 2016 assessments were administered to a total of 71,003 students in nine states and two Bureau of Indian Education schools. Counts of students tested in each state are displayed in Table *28*. The assessment sessions were administered by 16,578 educators in 10,245 schools and 2,914 school districts.

Table 28. *Student Participation by State* (N = *71,003)*

State	Students
Choctaw	20
Colorado	5,224
Illinois	11,191
Miccosukee Indian School	5
Mississippi	5,044
New Hampshire	805
New Jersey	10,422
New York	22,018
Oklahoma	6,657
West Virginia	2,504
Wisconsin	7,113

Table 29 summarizes the number of students tested in each grade during spring 2016. In grades 3 through 8, over 8,900 students participated in each grade. In high school, the largest number of students participated in grade 9, 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 29. *Student Participation by Grade* (N = *71,003)*

Grade	Students
3	8,943
4	9,185
5	9,262
6	9,604
7	9,746
8	9,601
9	5,907
10	2,841
11	5,432
12	482

Table 30 summarizes the demographic characteristics of students who participated in the spring 2016 administration. The majority of participants were male (67%) and white (59%). Only 6% of students were eligible or monitored for ELL services.

Table 30. *Demographic Characteristics of Participants*

Subgroup	*n*	%
Gender		
Female	23,394	32.95
Male	47,588	67.02
Missing	21	0.03
Race		
White	41,662	58.68
African American	16,344	23.02
Asian	3,205	4.51
American Indian	2,700	3.80
Alaska Native	43	0.06
Two or more races	6,838	9.63
Native Hawaiian or Pacific Islander	133	0.19
Missing	78	0.11
Hispanic Ethnicity		
No	55,705	78.45
Yes	15,050	21.20
Missing	248	0.35
English Language Learner (ELL) Participation		
Not ELL eligible or monitored	66,572	93.76
ELL eligible or monitored	4,431	6.24

In addition to the spring administration, instructionally embedded assessments are also made available for teachers to administer to students during the year. Results from these assessments do not contribute to final summative scoring but can be used to guide instructional decision-making. Table 31 summarizes the number of students participating in instructionally embedded testing by state. A total of 407 students took at least one instructionally embedded testlet during the 2015–2016 academic year.

Table 31. *Participation in Instructionally Embedded Testing by State*

State	*N*
Alaska	4
Colorado	8
Illinois	69
New Hampshire	1
New Jersey	4
New York	3
Oklahoma	281
West Virginia	36
Wisconsin	1

Table 32 and Table 33 summarize the number of instructionally embedded test sessions taken in English language arts (ELA) and mathematics, respectively. Across all states, students took a total of 2,107 ELA testlets and 2,398 mathematics testlets.

Table 32. *Instructionally Embedded English Language Arts Test Sessions by Grade (*N = *2,107)*

Grade	Total Test Sessions (*n*)
3	248
4	199
5	279
6	229
7	348
8	276
9	3
10	304
11	218
12	3

Table 33. *Instructionally Embedded Mathematics Test Sessions by Grade (*N = *2,398)*

Grade	Total Test Sessions (*n*)
3	282
4	233
5	368
6	229
7	359
8	328
9	30
10	382
11	168
12	19

VII.2. STUDENT PERFORMANCE

Student performance on Dynamic Learning Maps® (DLM®) assessments is interpreted using cut points, determined during standard setting (see Chapter VI in DLM Consortium, 2016), 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).

For the 2015–2016 administration, student performance was reported using the same four performance levels approved by the DLM Consortium for the 2014–2015 year.

- The student demonstrates Emerging understanding of and ability to apply content knowledge and skills represented by the EEs.
- The student's understanding of and ability to apply targeted content knowledge and skills represented by the EEs is Approaching the Target.
- The student's understanding of and ability to apply content knowledge and skills represented by the EEs is At Target.
- The student demonstrates Advanced understanding of and ability to apply targeted content knowledge and skills represented by the EEs.

VII.2.A. OVERALL PERFORMANCE

2015–2016 administration for ELA and mathematics. For ELA grades 3 through 11, the percentage of students who demonstrated performance at the At Target or Advanced level ranged from 24% to 37%; in English 2 approximately 20% of students were at the At Target or Advanced level; and in English 3 55% of students were in these categories. In mathematics grades 3 through 11, the percentage of students meeting or exceeding Target expectations ranged from approximately 7% to 30%, tending to decrease in the higher grades; 27% of students were At Target or Advanced in Algebra 1; and 28% and 64% of students were in these categories in Algebra 2 and Geometry, respectively.

Table 34. *Percentage of Students by Content Area, Grade, and Performance Level*

Grade	Performance Level				
	Emerging (%)	Approaching (%)	Target (%)	Advanced (%)	Target/Advanced (%)
ELA					
3 (*n* = 8,933)	59.0	16.5	22.1	2.4	24.6
4 (*n* = 9,166)	49.3	19.9	25.9	4.9	30.8
5 (*n* = 9,241)	47.8	19.4	26.9	5.9	32.8
6 (*n* = 9,585)	47.2	23.2	18.9	10.6	29.5
7 (*n* = 9,725)	37.0	27.9	26.0	9.2	35.1
8 (*n* = 9,577)	38.5	24.7	25.1	11.7	36.8
9 (*n* = 5,088)	35.4	28.9	27.2	8.5	35.7
10 (*n* = 1,667)	29.1	34.0	32.2	4.7	37.0
11 (*n* = 4,536)	38.3	31.1	26.6	4.0	30.6
English 2 (*n* = 2,278)	41.9	38.5	13.7	5.9	19.6
English 3 (*n* = 319)	22.9	21.9	41.7	13.5	55.2
Mathematics					
3 (*n* = 8,899)	58.0	15.7	17.8	8.4	26.3
4 (*n* = 9,161)	51.7	17.7	20.8	9.8	30.6
5 (*n* = 9,228)	56.9	19.9	12.7	10.5	23.2
6 (*n* = 9,563)	53.5	25.7	11.4	9.4	20.8
7 (*n* = 9,722)	65.2	22.8	7.7	4.3	11.9
8 (*n* = 9,567)	49.3	33.5	13.4	3.8	17.2
9 (*n* = 5,080)	44.6	34.7	16.9	3.7	20.7
10 (*n* = 1,668)	47.2	40.4	11.6	0.8	12.4
11 (*n* = 4,524)	65.3	27.9	6.6	0.2	6.8
Algebra 1 (*n* = 2,722)	60.0	13.0	12.0	15.0	27.0
Algebra 2 (*n* = 51)	49.0	23.5	15.7	11.8	27.5
Geometry (*n* = 250)	22.4	14.0	36.0	27.6	63.6

VII.2.B. SUBGROUP PERFORMANCE

Performance-level results for subgroups, including groups based on gender, race, ethnicity, and ELL status, were computed.

The distribution of students across performance levels was examined using demographic subgroups. Table 35 and Table 36 summarize the disaggregated frequency distributions for ELA and mathematics, respectively, collapsed across all assessed grade levels. Although each state has its 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. Rows labeled *Missing* indicate the student's demographic data were not entered into the system. The columns labeled *Not Assessed* reflect a small number of students who had records in the system and were tested in one but not both subjects. Overall, fewer demographic data were missing in 2015–2016 than in the previous year.

Table 35. *Students at Each ELA Performance Level by Demographic Subgroup* (N = *71,003)*

Subgroup	Performance Level									
	Emerging		Approaching		Target		Advanced		Not Assessed*	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
Gender										
Female	10,058	43.0	5,482	23.4	5,433	23.2	1,580	6.8	841	3.6
Male	20,344	42.8	10,751	22.6	11,490	24.1	3,451	7.3	1,552	3.3
Missing	8	38.1	8	38.1	2	9.5	3	14.3	0	0.0
Race										
White	17,638	42.3	9,578	23.0	10,170	34.4	3,087	7.4	1,189	2.9
African American	6,519	39.9	3,863	23.6	3,875	23.7	1,106	6.8	981	6.0
Asian	1,773	55.3	635	19.8	617	19.3	154	4.8	26	0.8
American Indian	973	36.0	596	22.1	755	28.0	235	8.7	141	5.2
Alaska Native	20	46.5	6	14.0	13	30.2	1	2.3	3	7.0
Two or more races	3,389	49.6	1,514	22.1	1,452	21.2	433	6.3	50	0.7
Native Hawaiian or Pacific Islander	68	51.1	32	24.1	23	17.3	8	6.0	2	1.5
Missing	30	38.5	17	21.8	20	25.6	10	12.8	1	1.3
Hispanic Ethnicity										
No	23,598	42.4	12,754	22.9	13,206	23.7	3,949	7.1	2,198	3.9
Yes	6,694	44.5	3,432	22.8	3,673	24.4	1,062	7.1	189	1.3
Missing	118	47.6	55	22.2	46	18.5	23	9.3	6	2.4
English Language Learner (ELL) Participation										
Not ELL eligible or monitored	28,571	42.9	15,119	22.7	15,811	23.8	4,766	7.2	2,305	3.5
ELL eligible or monitored	1,839	41.5	1,122	25.3	1,114	25.1	268	6.0	88	2.0

Note. *Students were not assessed on any English language arts Essential Elements.

Table 36. *Students at Each Mathematics Performance Level by Demographic Subgroup* (N = *71,003)*

Subgroup	Performance Level									
	Emerging		Approaching		Target		Advanced		Not Assessed*	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
Gender										
Female	13,133	56.1	5,699	24.4	2,797	12.0	1,296	5.5	469	2.0
Male	25,466	53.5	11,037	23.2	6,625	13.9	3,530	7.4	930	2.0
Missing	10	47.6	6	28.6	2	9.5	3	14.3	0	0.0
Race										
White	22,522	54.1	10,059	24.1	5,535	13.3	2,729	6.6	817	2.0
African American	8,672	53.1	3,887	23.8	2,237	13.7	1,168	7.1	380	2.3
Asian	2,011	62.7	581	18.1	377	11.8	205	6.4	31	1.0
American Indian	1,183	43.8	660	24.4	441	16.3	300	11.1	116	4.3
Alaska Native	21	48.8	9	20.9	4	9.3	5	11.6	4	9.3
Two or more races	4,080	59.7	1,501	22.0	811	11.9	401	5.9	45	0.7
Native Hawaiian or Pacific Islander	82	61.7	26	19.5	10	7.5	11	8.3	4	3.0
Missing	38	48.7	19	24.4	9	11.5	10	12.8	2	2.6
Hispanic Ethnicity										
No	30,407	54.6	13,150	23.6	7,176	12.9	3,725	6.7	1,247	2.2
Yes	8,064	53.6	3,542	23.5	2,222	14.8	1,077	7.2	145	1.0
Missing	138	55.6	50	20.2	26	10.5	27	10.9	7	2.8
English Language Learner (ELL) Participation										
Not ELL eligible or monitored	36,404	54.7	15,601	23.4	8,746	13.1	4,469	6.7	1,352	2.0
ELL eligible or monitored	2,205	49.8	1,141	25.8	678	15.3	360	8.1	47	1.1

Note. *Students were not assessed on any mathematics Essential Elements.

VII.2.C. LINKAGE-LEVEL MASTERY

As described earlier in the chapter, overall performance in each content area is based on the number of linkage levels mastered across all EEs. Based on the scoring method, for each EE, the highest linkage level the student mastered can be identified. This means that a student can be classified as a master of 0, 1 (Initial Precursor), 2 (Initial Precursor and Distal Precursor), 3 (Initial Precursor, Distal Precursor, and Proximal Precursor), 4 (Initial Precursor, Distal Precursor, Proximal Precursor, and Target), or 5 (Initial Precursor, Distal Precursor, Proximal Precursor, Target, and Successor) linkage levels. This section summarizes the distribution of students by highest linkage level mastered across all EEs in the grade/course and content area. For each EE, a student can demonstrate mastery of any of the five linkage levels. If the student did not master any of the linkage levels, the student's score report indicated no evidence of mastery, mastery of the Initial Precursor level, mastery of the Distal Precursor level, mastery of the Proximal-Precursor level, mastery of the Target level, and mastery of the Successor level (as the highest level of mastery) was summed across all EEs and divided by the total number of students assessed to obtain the proportion of students who mastered each linkage level.

Table 37 and Table 38 report the percentage of students who mastered each linkage level as the highest linkage level across all EEs for ELA and mathematics, respectively. For example, across all third grade ELA EEs, 19% of the time the highest level students mastered was the Initial Precursor level. For ELA, the average percent of students who mastered as high as the Target or Successor linkage level across all EEs ranged from approximately 21% in grade 3 to 35% in English 3. For mathematics, the average percent of students who mastered the Target or Successor linkage level across all EEs ranged from approximately 5% in Algebra 1 to 38% in Algebra 2.

Table 37. *Percentage of Students Demonstrating Highest Level Mastered Across ELA EEs, by Grade/Course*

Grade/Course	Linkage Level					
	No Evidence (%)	IP (%)	DP (%)	PP (%)	T (%)	S (%)
3 (*n* = 8,933)	23.6	18.8	19.9	16.9	12.4	8.4
4 (*n* = 9,166)	23.8	15.5	15.0	17.8	13.4	14.5
5 (*n* = 9,241)	20.3	17.1	17.9	18.7	12.0	14.0
6 (*n* = 9,585)	23.3	20.1	18.8	16.7	11.2	10.0
7 (*n* = 9,725)	20.3	20.3	18.3	15.1	11.6	14.3
8 (*n* = 9,577)	23.4	18.7	15.6	14.0	13.8	14.5
9 (*n* = 5,088)	23.3	19.5	13.4	20.7	13.8	9.2
10 (*n* = 1,667)	18.9	20.5	13.6	19.3	14.6	13.1
11 (*n* = 4,536)	26.8	19.9	17.3	15.0	12.4	8.6
English 2 (*n* = 2,278)	16.7	25.5	15.4	19.2	15.0	8.3
English 3 (*n* = 319)	22.4	12.9	6.8	22.7	19.0	16.1

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

Table 38. *Percentage of Students Demonstrating Highest Level Mastered Across Mathematics EEs, by Grade/Course*

Grade/Course	Linkage Level					
	No Evidence (%)	IP (%)	DP (%)	PP (%)	T (%)	S (%)
3 (*n* = 8,899)	37.2	28.3	14.4	10.6	5.9	3.5
4 (*n* = 9,161)	31.7	26.0	17.0	14.6	6.1	4.7
5 (*n* = 9,228)	35.4	30.4	15.0	9.8	5.8	3.6
6 (*n* = 9,563)	35.4	24.4	15.1	14.0	6.8	4.2
7 (*n* = 9,722)	33.4	33.9	14.7	9.7	5.4	2.8
8 (*n* = 9,567)	33.4	23.1	16.5	15.4	8.7	2.9
9 (*n* = 5,080)	18.9	20.5	13.6	19.3	14.6	13.1
10 (*n* = 1,668)	28.1	23.8	19.6	15.9	6.8	5.7
11 (*n* = 4,524)	32.7	28.5	19.7	9.7	5.2	4.1
Algebra 1 (*n* = 2,722)	43.8	31.8	15.2	4.3	3.6	1.4
Geometry (*n* = 250)	34.3	22.5	18.5	15.1	5.7	3.8
Algebra 2 (*n* = 51)	6.3	11.5	20.8	24.0	12.5	25.0

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

VII.3. Data Files

Three data files, made available to DLM state partners, summarized results from the 2015–2016 year. Similar to 2014–2015, the General Research File (GRF) contained student results, including each student's highest LL mastered for each EE and final performance level for the subject for all students who completed any testlets. The Incident File listed students who were affected by one of the known problems with testlet assignments during the spring 2016 window (see Chapter IV) using the same structure as 2014–2015. Finally, a new data file, called the Special Circumstances File, was delivered for 2015–2016 and provided information about which students and EEs were affected by extenuating circumstances (e.g., chronic absences), as defined by each state.

During 2016, the GRF structure remained largely the same as in 2015. However, one change made to the GRF in 2015–2016 was the addition of an Invalidation Code column, along with the inclusion of a two-week review window following delivery of the GRF. During the two-week review window, states were given the opportunity to review the GRF and make changes to records. State partners were able to make changes to demographic data in the GRF to ensure

accuracy of data in score reports. State partners were not able to make changes to any of the EE or final performance-level values. During the review window, state partners also had the opportunity to use the supplemental files to determine if an entire student record should be invalidated. The addition of the invalidation code column was to allow states the ability to invalidate students that should not be included for their specific reporting or accountability purposes. These decisions were made by each state based on their own state policies and procedures. When a state invalidated a record and resubmitted the GRF to DLM staff, the student did not receive an Individual Student Score Report and was excluded from aggregated reports.

The Date Time Supplemental File was not provided in 2016 because of enhancements made to the test delivery system. In 2014–2015, only a single, consortium-wide delivery window was available for spring assessments. While states defined state-specific administration windows, students were able to test outside those windows within the larger consortium spring window. System enhancements in 2016 prevented students from testing outside their state-specific spring window, thereby eliminating the need for the Date Time Supplemental File to identify students who tested outside their state-specific window.

VII.4. 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 Student Score Reports were provided to educators and parents/guardians. Several aggregated reports were provided to state and local education agencies, including reports for the classroom, school, district, and state. No changes were made to the aggregated reports or their delivery during 2016. Changes to the Individual Student Score Reports are summarized below. For a complete description of score reports, including aggregated reports, see Chapter VII of the *2014-2015 Technical Manual – Year-End Model* (DLM Consortium, 2016).

VII.4.A. Individual Student Score Reports

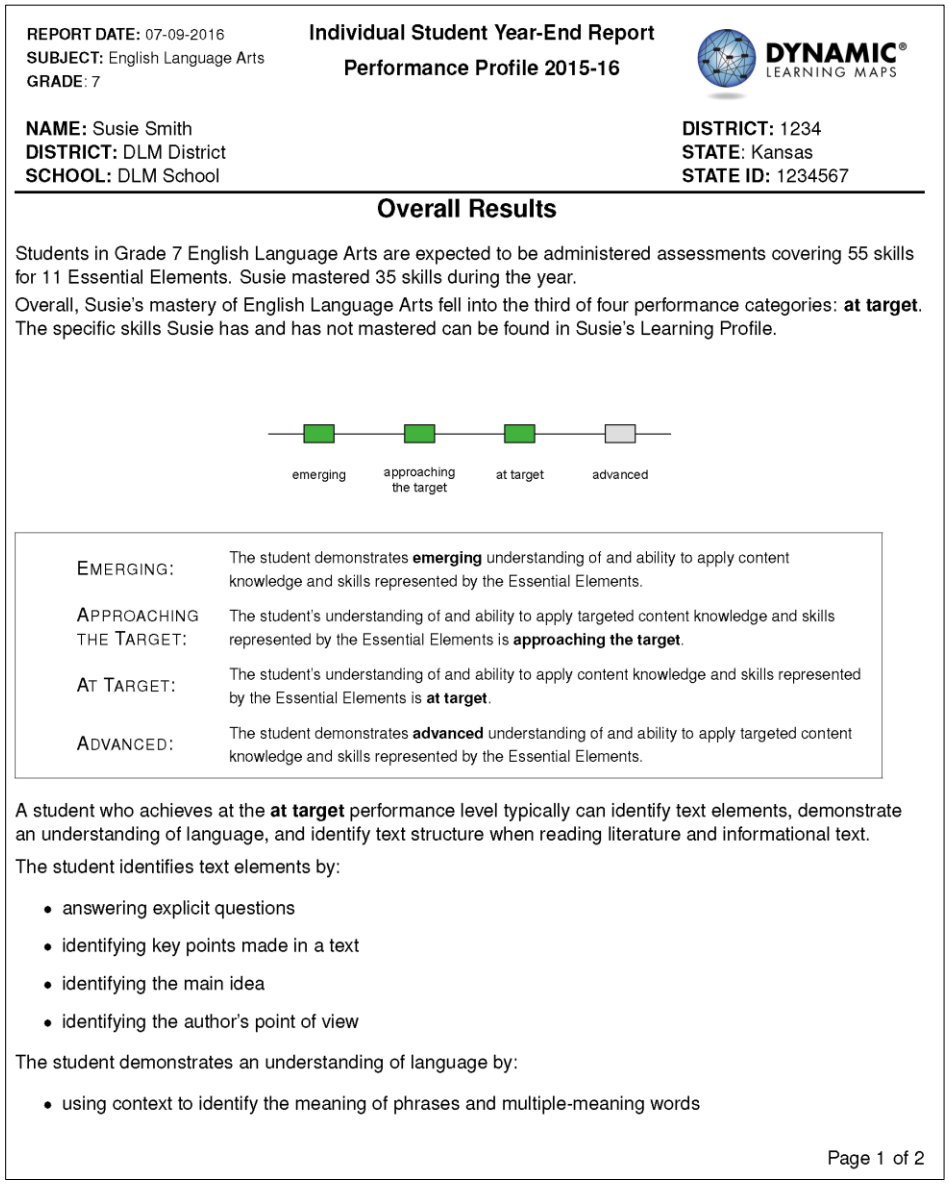
During the 2015–2016 year, minor changes were made to the Individual Student Score Reports.

One change to the content of the Performance Profile was the inclusion of grade and content performance level descriptors (PLDs). These grade and content PLDs replaced the bulleted list of skills mastered used in 2014–2015. The grade and content PLDs were developed after standard setting was conducted in 2015 to describe the types of skills typically mastered by students in a given performance level.

At the December 2015 governance meeting, year-end model state partners voted to remove the Learning Profile portion from the 2016 Individual Student Score Reports. This decision was made because of the limited number of items informing the mastery classification for each linkage level and the concern about the reliability of interpretations made on such limited data.

Minor changes were also made to the display of information found in the header of the Individual Student Score Reports. Additionally, 2016 Individual Student Score Reports were

delivered via the Educator Portal platform rather than through the secure file transfer platform used in 2015 (for more information on Educator Portal, see Chapter IV of DLM Consortium, 2016). A sample Individual Student Score Report reflecting the 2016 changes is provided in Figure 12.



REPORT DATE: 07-09-2016
SUBJECT: English Language Arts
GRADE: 7

Individual Student Year-End Report
Performance Profile 2015-16

DYNAMIC LEARNING MAPS

NAME: Susie Smith
DISTRICT: DLM District
SCHOOL: DLM School

DISTRICT: 1234
STATE: Kansas
STATE ID: 1234567

Overall Results

Students in Grade 7 English Language Arts are expected to be administered assessments covering 55 skills for 11 Essential Elements. Susie mastered 35 skills during the year.

Overall, Susie's mastery of English Language Arts fell into the third of four performance categories: **at target**. The specific skills Susie has and has not mastered can be found in Susie's Learning Profile.

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

A student who achieves at the **at target** performance level typically can identify text elements, demonstrate an understanding of language, and identify text structure when reading literature and informational text.

The student identifies text elements by:

- answering explicit questions
- identifying key points made in a text
- identifying the main idea
- identifying the author's point of view

The student demonstrates an understanding of language by:

- using context to identify the meaning of phrases and multiple-meaning words

Page 1 of 2

Figure 12. Page 1 of the performance profile for 2015–2016.

VII.5. QUALITY CONTROL PROCEDURES FOR DATA FILES AND SCORE REPORTS

In 2016, quality control procedures were updated to include automated procedures following a spring 2016 quality control audit. No changes were made to manual quality control checks for 2016. For a complete description of quality control procedures, see Chapter VII of the *2014-2015 Technical Manual – Year-End Model* (DLM Consortium, 2016).

VII.5.A. Quality Control Audit

An audit of the quality control processes was held on March 25, 2016. Attendees included DLM psychometric staff; the director of the Dynamic Learning Maps project; the director of the Center for Educational Testing and Evaluation (CETE), which houses the DLM project; CETE psychometric staff; and the director of the Achievement and Assessment Institute, which houses CETE. Process documentation was created to ensure that established quality control procedures were clearly outlined and easily comprehensible. The audit meeting concluded that the quality control procedures currently in place were acceptable, though several enhancements were suggested for the 2015–2016 reporting cycle. Suggested changes included creation of automated checks using the R programming language, use of networked workstations to coordinate score report generation and review, and the addition of reasonableness checks to ensure that data retrieved from the database did not contain any unexpected values.

VII.5.B. Automated Quality Control Checks

To allow quality control checks to be performed more rapidly and efficiently, R programs were developed to perform quality control procedures on the GRF and on Individual Student Score Reports.

VII.5.B.i. GRF Automated Quality Control Program

The first program written to perform automated checks was designed to perform quality control on the GRF. This program conducts a series of checks that can be organized into four main steps.

1. Check the data for reasonableness (checks detailed below).
2. Ensure that the number of linkage levels mastered for each student is less than or equal to the maximum possible value for that grade and subject.
3. Check all EE scores against the original scoring file.
4. Verify that students participating in End-of-Instruction assessments are displayed with one row per course.

The automated program checks each row of data in the GRF and generates errors for review by the psychometrics team.

The reasonableness checks ensure that the GRF column names accurately match the data dictionary provided to states and additional check the following columns to ensure that data match defined parameters: Student ID, State Student Identifier, Current Grade Level, Course, Student Legal First Name, Student Legal Middle Name, Student Legal Last Name, Generation Code, Username, First Language, Date of Birth, Gender, Comprehensive Race, Hispanic Ethnicity, Primary Disability Code, ESOL Participation Code, School Entry Date, District Entry Date, State Entry Date, State, District Code, District, School Code, School, Educator First Name, Educator Last Name, Educator Username, Final ELA Band, and Final Math Band. If invalid

values are found, they are corrected as necessary by DLM staff and/or state partners during their two-week review period.

VII.5.B.ii. Student Score Reports Automated Quality Control Program

An automated program was developed to support manual review of Individual Student Score Reports. The program was written to check key values used to generate the Individual Student Score Reports. As the score reporting program generates reports, it creates a *proofreader* file containing the values that are used to create each score report. These values are then checked against the GRF to ensure that they are being accurately populated into score reports.

Demographic values including student name, school, district, grade level, state, and state student identifier are checked to ensure a precise match. Values of skills mastered, performance levels, conceptual areas tested and mastered, and EEs mastered and tested are also checked to ensure the correct values are populated, and values referring to the total number of skills, EEs, or conceptual areas available are checked to ensure they are the correct value for that grade, subject, and content area.

VIII. RELIABILITY

The Dynamic Learning Maps® (DLM®) Alternate Assessment System uses nontraditional psychometric models (i.e., diagnostic classification models) to produce student score reports. As such, evidence for the reliability of scores[5] 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 "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. Values closer to 1.0 indicate variation in test scores comes from individual differences in the trait, while values closer to 0.0 indicate variation in test scores comes from random error.

Many traditional measures of reliability exist; their differences are due to assumptions each 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 stricter definition of Spearman–Brown, then alpha will be equal to Spearman–Brown. As a

[5]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.

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 (DCMs) 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 content strands, alternate achievement standards, and levels within standards in which each student was tested. In DLM terms, each content strand is called a conceptual area, which is made up of standards called Essential Elements (EEs). Each EE is divided into five linkage levels of complexity: Initial Precursor (IP), Distal Precursor (DP), Proximal Precursor (PP), Target (T), and Successor (S).

For each linkage level embedded within each EE, DLM testlets were written with items measuring the listed linkage level. Because of the DLM administration design, students took testlets on a single linkage level within an EE. Therefore, a linkage-level model was used to estimate examinee proficiency. (See Chapter V of this manual for more information.)

The DCMs used in psychometric analyses of student test data produced student-level posterior probabilities for each linkage level for which a student was tested, with a threshold of 0.8 specified for demonstrating mastery. (See Chapter VI of the *2014-2015 Technical Manual – Year-End Model*.) To guard against the model being overly influential, two additional scoring rules were applied. Students could additionally demonstrate mastery by providing correct responses to at least 80% of the items measuring the EE and linkage level. Furthermore, because students did not test at more than one linkage level within an EE, students who did not meet mastery status for the tested linkage level were assigned mastery status for the linkage level two levels below the level in which they were tested (unless the level tested was either the IP or DP level, in which case students were considered nonmasters of all linkage levels within the EE).

Linkage-level results are aggregated for EEs within each conceptual area on DLM score reports. Score reports also summarize overall performance in each content area with a performance level classification. The classification is determined by summing all linkage levels mastered in the content area and comparing the value with cut points determined during standard setting. For more information on cut points, see Chapter VI of *2014-2015 Technical Manual – Year-End Model* (2016). For more information on score reports, see Chapter VII in this manual.

Consistent with the levels at which DLM results are reported, this chapter provides six types of reliability evidence: (a) classification to overall performance level (performance-level reliability);

(b) the total number of linkage levels mastered within a content area (content-area reliability; provided for ELA and mathematics); (c) the number of linkage levels mastered within each conceptual area (conceptual-area reliability); (d) the number of linkage levels mastered within each EE (EE reliability); (d) the classification accuracy of each linkage level within each EE (linkage-level reliability); and (e) classification accuracy summarized for the five linkage levels (conditional evidence by linkage level). 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.

Each type of reliability evidence provides various correlation coefficients. Correlation estimates mirror estimates of reliability from contemporary measures such as the Guttman–Cronbach alpha. For performance level and EE reliability, the polychoric correlation estimates the relationship between two ordinal variables: true performance level or number of linkage levels mastered, and estimated value. For content-area reliability and conceptual-area/claim reliability, the Pearson correlation estimates the relationship between the true and estimated numbers of linkage levels mastered. Finally, for linkage level and conditional evidence by linkage-level reliability, the tetrachoric correlation estimates the relationship between true and estimated linkage-level mastery statuses. The tetrachoric correlation is a special case of the polychoric in which the variables are discrete. Both the polychoric and tetrachoric correlations are intended to provide more accurate estimates of relationships between ordinal and discrete variables that would otherwise be attenuated using the traditional correlation (i.e., the Pearson coefficient).

Each type of reliability evidence produces correct classification rates (raw and chance corrected), which indicate the proportion of estimated classifications that match true classifications. The chance-corrected classification rate is labeled kappa and represents the proportion of error reduced above chance. Values of kappa above 0.6 indicate substantial-to-perfect agreement between estimated and true values (Landis & Koch, 1977).

With the classification methods of DCMs 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). In other words, because diagnostic classification modeling does not produce a total score or scale score, traditional methods of calculating conditional standard errors of measurement are not appropriate. Rather, an alternative method is presented whereby reliability evidence is summarized for each linkage level. Since linkage levels are intended to represent varying levels of skills and abilities, reliability provided at each level is analogous to conditional reliability evidence.

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).

Because the DLM psychometric model produces complex mastery results summarized at multiple levels of reporting (performance level, content area, conceptual area, EE, and linkage levels) rather than a traditional raw or scale score value, methods for evaluating reliability 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 DCMs, simulation-based reliability has been used in a number of studies (e.g., Bradshaw, Izsák, Templin, & Jacobson, 2014; Templin & Bradshaw, 2013). For a simulation-based method of computing reliability, 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 assessments, 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. Because the simulation is based on a replication of the exact same items that were administered to examinees, the two administrations are perfectly parallel. However, the use of simulation produces approximate estimates of reliability, which are contingent on the accuracy of the current scoring model.

VIII.1.A.i. Reliability Sampling Procedure

The simulation design that was used to obtain 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 operational testing data (spring window). 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[6] for the items of the testlet, conditional on the profile of linkage-level mastery or nonmastery for the student.
3. Score the simulated item responses using the operational DLM scoring procedure (described in Chapter V),[7] 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.
5. Repeat Steps 1 through 4 for 2,000,000 simulated students.

Figure 13 presents Steps 1 through 4 of the simulation process as a flow chart.

[6]Calibrated-model parameters were treated as true and fixed values for the simulation.

[7]To be consistent with the operational scoring procedure, all three scoring rules were included when scoring the simulated responses. The scoring rules are described further in Chapter V.

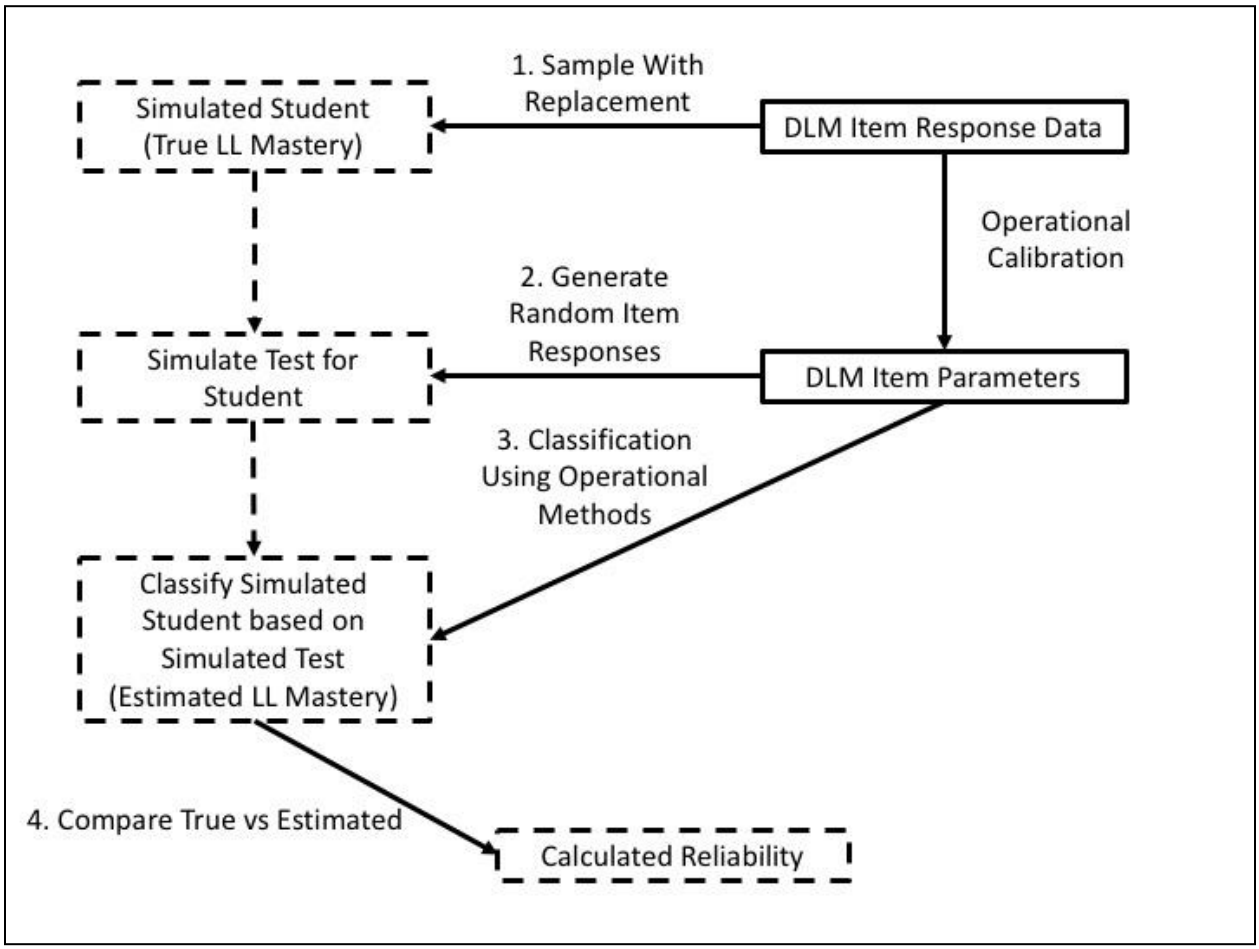


Figure 13. Simulation process for creating reliability evidence.

Note: LL = linkage level.

VIII.2. 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 six levels of data, each important in the DLM testing design: (a) performance-level reliability, (b) content-area reliability, (c) conceptual-area reliability, (d) EE reliability, (e) linkage-level reliability, and (f) conditional reliability by linkage level. With 255 EEs, each with five linkage levels, a total of 1,275 analyses were conducted to summarize reliability. Due to the number of analyses, the reported evidence will be summarized in this chapter. Full reporting of reliability evidence for all 1,210 linkage levels and 242 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 six 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 six levels ensures that these reliability-estimation procedures meet Standard 2.5.

VIII.2.A. Performance-Level Reliability Evidence

Results from DLM assessments are reported using four performance levels. The total of linkage levels mastered in each content area is summed, and cut points are applied to distinguish between performance categories.

Performance-level reliability provides evidence for how reliably students were classified into the four performance levels for each content area and grade level. Because performance level is based on total linkage levels mastered, large fluctuations in the number of linkage levels mastered or fluctuation around the cut points could impact how reliably students are classified to performance categories. The performance-level reliability evidence is based on the true and estimated performance level (based on estimated total number of linkage levels mastered and predetermined cut points) for a given content area. Three statistics are included to provide a comprehensive summary of results; the specific metrics were chosen due to their interpretability.

1. The polychoric correlation between the true and estimated performance level within a grade and content area
2. The correct classification rate between the true and estimated performance level within a grade and content area
3. The correct classification kappa between the true and estimated performance level within a grade and content area

Table 39 shows this information across all grades and content areas. Polychoric correlations between true and estimated performance levels range from .840 to .925. Correct classification rates range from 0.813 and 0.961, and Cohen's kappa values are between 0.945 and 0.992. These results indicate that the DLM scoring procedure of assigning and reporting performance levels

based on total linkage levels mastered results in reliable classification of students to performance-level categories.

Table 39. *Summary of Performance-Level Reliability Evidence*

Grade/ Course	Content Area	Polychoric Correlation	Correct Classification Rate	Cohen's Kappa
3	English language arts	0.915	0.946	0.981
3	Mathematics	0.881	0.929	0.983
4	English language arts	0.917	0.955	0.986
4	Mathematics	0.888	0.947	0.989
5	English language arts	0.925	0.961	0.988
5	Mathematics	0.870	0.939	0.986
6	English language arts	0.897	0.950	0.992
6	Mathematics	0.877	0.936	0.989
7	English language arts	0.896	0.946	0.989
7	Mathematics	0.896	0.919	0.984
8	English language arts	0.906	0.955	0.992
8	Mathematics	0.915	0.938	0.990
9	English language arts	0.894	0.940	0.988
9	Mathematics	0.880	0.913	0.984
10	English language arts	0.912	0.919	0.962
10	Mathematics	0.840	0.820	0.953
11	English language arts	0.918	0.938	0.985
11	Mathematics	0.872	0.831	0.955
Algebra 1	Mathematics	0.877	0.949	0.992
Algebra 2	Mathematics	0.859	0.918	0.991
Geometry	Mathematics	0.846	0.933	0.987
English 2	English language arts	0.900	0.813	0.945
English 3	English language arts	0.846	0.921	0.985

VIII.2.B. CONTENT-AREA RELIABILITY EVIDENCE

Content-area reliability provides consistency evidence for the number of linkage levels mastered across all EEs for a given content area and grade level. Because 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 analogous to the number of items answered correctly (e.g., total score) in a different type of testing program.

Content-area reliability evidence compares the true and estimated number of linkage levels mastered across all tested levels for a given content area. Reliability is reported with three summary numbers.

1. The Pearson correlation between the true and estimated number of linkage levels mastered within a content area
2. The correct classification rate for which linkage levels were mastered as averaged across all simulated students
3. The correct classification kappa for which linkage levels were mastered as averaged across all simulated students

Table 40 shows the three summary values for each grade and content area. Classification-rate information is provided in accordance with Standard 2.16. The two summary statistics included in Table 40 also meet Standard 2.19.

Table 40. *Summary of Content-Area Reliability Evidence*

Grade/ Course	Content Area	Linkage Levels Mastered Correlation	Average Student Correct Classification	Average Student Cohen's Kappa
3	English language arts	0.991	0.968	0.906
3	Mathematics	0.978	0.974	0.917
4	English language arts	0.994	0.967	0.898
4	Mathematics	0.988	0.963	0.879
5	English language arts	0.994	0.971	0.914
5	Mathematics	0.988	0.966	0.883
6	English language arts	0.991	0.962	0.891
6	Mathematics	0.984	0.971	0.912
7	English language arts	0.992	0.961	0.890
7	Mathematics	0.984	0.969	0.905
8	English language arts	0.993	0.968	0.904
8	Mathematics	0.988	0.974	0.931
9	English language arts	0.991	0.965	0.904
9	Mathematics	0.979	0.986	0.971
10	English language arts	0.987	0.961	0.890
10	Mathematics	0.968	0.986	0.968
11	English language arts	0.991	0.970	0.919
11	Mathematics	0.965	0.990	0.979
Algebra 1	Mathematics	0.984	0.982	0.960
Algebra 2	Mathematics	0.987	0.988	0.980
Geometry	Mathematics	0.983	0.986	0.974
English 2	English language arts	0.977	0.976	0.943
English 3	English language arts	0.987	0.967	0.923

It is evident from Table *40* that content-area reliability, as demonstrated by the correlation between true and estimated number of linkage levels mastered, ranges from .965 to .994. These

values indicate the DLM scoring procedure of reporting the number of linkage levels mastered provides reliable results of student performance.

VIII.2.C. Conceptual-Area Reliability Evidence

Within each content area, students are assessed on multiple content strands. These strands of related EEs are called conceptual areas and describe the overarching sections of the learning map model upon which DLM assessments are developed (see Chapter II of the *2014-2015 Technical Manual – Year-End Model* for more information). Because student score reports summarize the number and percentage of linkage levels students mastered for each conceptual area (see Chapter VII of this manual for more information), reliability evidence is provided for each conceptual area.

Conceptual-area reliability provides consistency evidence for the number of linkage levels mastered across all EEs in each conceptual area for each grade and content area. Because conceptual-area reporting summarizes the total linkage levels a student mastered, the statistics reported for conceptual-area reliability are the same as described for content-area reliability.

Conceptual-area reliability evidence compares the true and estimated number of linkage levels mastered across all tested levels for each conceptual area. Reliability is reported with three summary numbers.

1. The Pearson correlation between the true and estimated number of linkage levels mastered within a conceptual area
2. The correct classification rate for which linkage levels were mastered as averaged across all simulated students for each conceptual area
3. The correct classification kappa for which linkage levels were mastered as averaged across all simulated students for each conceptual area

Table 41 and Table 42 show the three summary values for each conceptual area, by grade, for English language arts (ELA) and mathematics respectively. Values range from 0.591 to 0.999 in ELA and from 0.504 to 0.999 in mathematics, indicating that overall the DLM method of reporting the total and percentage of linkage levels mastered by conceptual area results in values that can be reliably reproduced.

Table 41. *Summary of English Language Arts Conceptual-Area Reliability Evidence*

Grade/ Course	Conceptual Area	Linkage Levels Mastered Correlation	Average Student Correct Classification	Average Student Cohen's Kappa
3	ELA.C1.1	0.978	0.986	0.972
3	ELA.C1.2	0.969	0.984	0.967
3	ELA.C1.3	0.916	0.996	0.994
3	ELA.C2.1	0.920	0.996	0.995
4	ELA.C1.1	0.984	0.986	0.970
4	ELA.C1.2	0.976	0.976	0.941
4	ELA.C1.3	0.936	0.999	0.999
4	ELA.C2.1	0.992	0.998	0.997
5	ELA.C1.1	0.965	0.996	0.994
5	ELA.C1.2	0.986	0.981	0.956
5	ELA.C1.3	0.968	0.993	0.988
5	ELA.C2.1	0.978	0.998	0.997
6	ELA.C1.1	0.591	0.997	0.997
6	ELA.C1.2	0.984	0.966	0.910
6	ELA.C1.3	0.955	0.993	0.990
6	ELA.C2.1	0.981	0.999	0.998
7	ELA.C1.1	0.750	0.997	0.997
7	ELA.C1.2	0.981	0.972	0.929
7	ELA.C1.3	0.964	0.989	0.981
7	ELA.C2.1	0.969	0.993	0.988
8	ELA.C1.2	0.986	0.970	0.912
8	ELA.C1.3	0.936	0.989	0.981
8	ELA.C2.1	0.994	0.998	0.997
9	ELA.C1.2	0.981	0.970	0.923
9	ELA.C1.3	0.939	0.990	0.983
9	ELA.C2.1	0.974	0.997	0.996
9	ELA.C2.2	0.979	0.998	0.998
10	ELA.C1.2	0.982	0.966	0.911
10	ELA.C1.3	0.929	0.987	0.978
10	ELA.C2.1	0.973	0.997	0.997
10	ELA.C2.2	0.981	0.999	0.999
11	ELA.C1.2	0.978	0.981	0.960

Grade/ Course	Conceptual Area	Linkage Levels Mastered Correlation	Average Student Correct Classification	Average Student Cohen's Kappa
11	ELA.C1.3	0.953	0.983	0.966
11	ELA.C2.1	0.991	0.998	0.997
11	ELA.C2.2	0.942	0.998	0.997
English 2	ELA.C1.2	0.973	0.986	0.971
English 2	ELA.C1.3	0.865	0.991	0.987
English 2	ELA.C2.1	0.975	0.998	0.997
English 2	ELA.C2.2	0.981	0.999	0.999
English 3	ELA.C1.2	0.951	0.984	0.970
English 3	ELA.C1.3	0.883	0.991	0.987
English 3	ELA.C2.1	0.984	0.996	0.994
English 3	ELA.C2.2	0.911	0.994	0.991

Table 42. *Summary of Mathematics Conceptual-Area Reliability Evidence*

Grade/ Course	Conceptual Area	Linkage Levels Mastered Correlation	Average Student Correct Classification	Average Student Cohen's Kappa
3	M.C1.1	0.929	0.995	0.993
3	M.C1.3	0.852	0.998	0.998
3	M.C2.2	0.785	0.998	0.998
3	M.C3.1	0.910	0.996	0.994
3	M.C3.2	0.822	0.998	0.998
3	M.C4.1	0.930	0.996	0.994
3	M.C4.2	0.653	0.998	0.997
4	M.C1.1	0.874	0.997	0.996
4	M.C1.2	0.837	0.994	0.991
4	M.C1.3	0.903	0.999	0.998
4	M.C2.1	0.941	0.993	0.990
4	M.C2.2	0.939	0.999	0.999
4	M.C3.1	0.949	0.995	0.992
4	M.C3.2	0.764	0.998	0.998
4	M.C4.1	0.904	0.995	0.993
4	M.C4.2	0.611	0.997	0.996
5	M.C1.1	0.820	0.994	0.992

Grade/ Course	Conceptual Area	Linkage Levels Mastered Correlation	Average Student Correct Classification	Average Student Cohen's Kappa
5	M.C1.2	0.946	0.994	0.990
5	M.C1.3	0.911	0.995	0.993
5	M.C2.1	0.957	0.997	0.996
5	M.C2.2	0.971	0.999	0.999
5	M.C3.1	0.951	0.994	0.991
5	M.C3.2	0.871	0.998	0.998
5	M.C4.2	0.680	0.997	0.997
6	M.C1.1	0.864	0.998	0.998
6	M.C1.2	0.893	0.994	0.992
6	M.C1.3	0.935	0.996	0.994
6	M.C2.2	0.934	0.996	0.995
6	M.C3.2	0.860	0.998	0.998
6	M.C4.1	0.908	0.992	0.989
7	M.C1.1	0.896	0.996	0.994
7	M.C1.2	0.804	0.998	0.998
7	M.C1.3	0.928	0.994	0.991
7	M.C2.1	0.939	0.996	0.994
7	M.C2.2	0.839	0.999	0.998
7	M.C3.2	0.914	0.997	0.996
7	M.C4.1	0.820	0.999	0.998
7	M.C4.2	0.831	0.998	0.998
8	M.C1.1	0.579	0.996	0.995
8	M.C1.2	0.798	0.998	0.997
8	M.C1.3	0.942	0.997	0.997
8	M.C2.1	0.950	0.994	0.991
8	M.C2.2	0.916	0.999	0.999
8	M.C3.2	0.868	0.998	0.998
8	M.C4.1	0.843	0.998	0.998
8	M.C4.2	0.931	0.990	0.983
9	M.C1.3	0.938	0.994	0.991
9	M.C2.1	0.912	0.996	0.995
9	M.C2.2	0.845	0.999	0.998
9	M.C4.1	0.781	0.996	0.994
10	M.C1.3	0.804	0.998	0.998

Grade/ Course	Conceptual Area	Linkage Levels Mastered Correlation	Average Student Correct Classification	Average Student Cohen's Kappa
10	M.C2.1	0.856	0.999	0.999
10	M.C3.1	0.896	0.999	0.999
10	M.C3.2	0.898	0.997	0.996
10	M.C4.1	0.866	0.997	0.997
10	M.C4.2	0.890	0.997	0.997
11	M.C1.3	0.938	0.994	0.991
11	M.C1.3	0.898	0.998	0.998
11	M.C2.1	0.869	0.999	0.999
11	M.C3.2	0.866	0.999	0.999
Algebra 1	M.C4.2	0.947	0.995	0.991
Algebra 1	M.C1.3	0.951	0.990	0.983
Algebra 1	M.C3.1	0.857	0.998	0.998
Algebra 1	M.C3.2	0.944	0.997	0.996
Algebra 2	M.C4.1	0.935	0.998	0.997
Algebra 2	M.C1.3	0.817	0.993	0.991
Algebra 2	M.C3.2	0.504	0.995	0.993
Algebra 2	M.C4.1	0.734	0.983	0.974
Geometry	M.C4.2	0.985	0.997	0.996
Geometry	M.C2.1	0.935	0.992	0.988
Geometry	M.C2.2	0.694	0.998	0.998

VIII.2.D. Essential-Element Reliability Evidence

Moving from higher-level aggregation to EEs, the reliability evidence shifts slightly. That is, because EEs are collections of linkage levels with an implied order, the highest linkage level mastered per EE, rather than for the whole content area, is examined. If content-area scores are regarded as total scores from an entire assessment, evidence at the EE level is more fine grained than reporting at a content-area strand level, which is commonly reported for other testing programs. EEs are the specific standards within the content area itself.

Three statistics are used to summarize reliability evidence for EEs.

1. The polychoric correlation between true and estimated numbers of linkage levels mastered within an EE
2. The correct classification rate for the number of linkage levels mastered within an EE

3. The correct classification kappa for the number of linkage levels mastered within an EE

Because there are 242 EEs, the summaries reported herein are based on the number and proportion of EEs that fall within a given range of an index value. Results are given in both tabular and graphical form. Table 43 and Figure 14 provide proportions and the number of EEs, respectively, falling within prespecified ranges of values for the three reliability summary statistics (i.e., correct classification rate, kappa, and correlation). In general, the reliability summaries for number of linkage levels mastered within EEs show strong evidence of reliability.

Table 43. *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**
Polychoric Correlation	0.000	0.000	0.004	0.004	0.012	0.029	0.153	0.500	0.298
Correct Classification Rate	0.000	0.000	0.000	0.033	0.103	0.335	0.364	0.132	0.033
Kappa	0.000	0.004	0.000	0.017	0.037	0.128	0.310	0.393	0.112

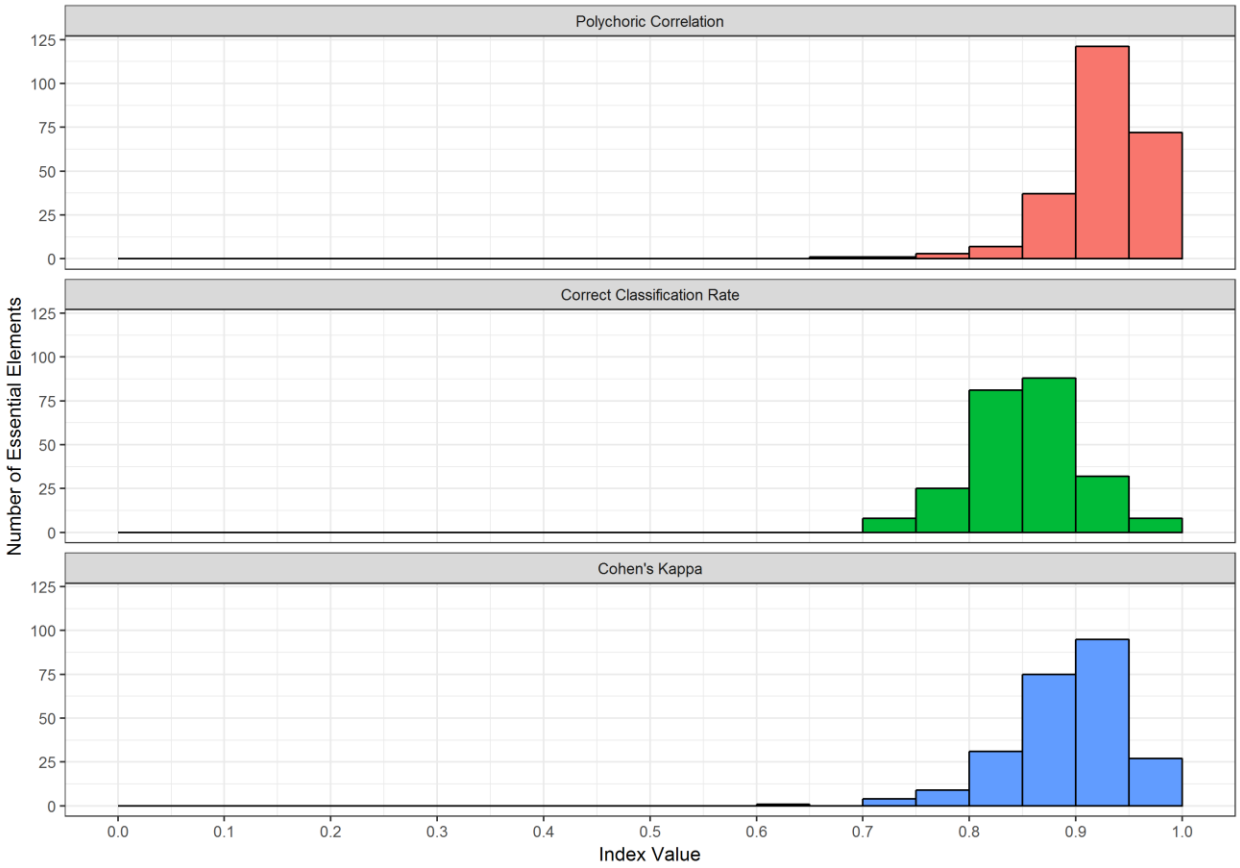


Figure 14. Number of linkage levels mastered within EE reliability summaries.

VIII.2.E. LINKAGE-LEVEL RELIABILITY EVIDENCE

Evidence at the linkage level comes from the comparison of true and estimated mastery statuses for each of the 1,210 linkage levels in the operational DLM assessment.[8] This level of reliability reporting is even more fine grained than at the EE level. While it does not have a comparable classical test theory or item response theory analog, its inclusion is important because it is the level at which mastery classifications are made for DLM assessments.

As one example, Table 44 shows a simulated table from the PP linkage level of the EE, M.EE.MD.3.4.

[8]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 linkage levels during assessment, see Chapter 3 – Pilot Administration: Initialization and Chapter 4 – Adaptive Delivery in the *2014-2015 Technical Manual, Year-End*.

Table 44. *Example of True and Estimated Mastery Status from Reliability Simulation for Proximal Precursor Linkage Level of Essential Element M.EE.MD.3.4*

		Estimated Mastery Status	
		Nonmaster	**Master**
True Mastery Status	**Nonmaster**	574	235
	Master	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 1,210 linkage levels. Three summary statistics are presented.

1. The tetrachoric correlation between estimated and true mastery status
2. The correct classification rate for the mastery status of each linkage level
3. The correct classification kappa for the mastery status of each linkage level

As there are 1,210 total linkage levels across all 242 EEs, the summaries reported herein are based on the proportion and number of linkage levels that fall within a given range of an index value. Results are given in both tabular and graphical form. Table 45 and Figure 15 provide proportions and number of linkage levels, respectively, that fall within prespecified ranges of values for the three reliability summary statistics (i.e., correct classification rate, kappa, and correlation). The kappa value for 38 linkage levels could not be computed because all students were labeled as masters or nonmasters of the linkage level.

The correlations and correct classification rates show reliability for the classification of mastery at the linkage level. Across all linkage levels, a total of three had tetrachoric correlation values below 0.6, zero had a correct classification rate below 0.6, and 40 had a kappa value below 0.6.

Table 45. *Reliability Summaries Across All Linkage Levels: Proportion of Linkage Levels 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
Tetrachoric Correlation	0.003	0.003	0.000	0.004	0.003	0.016	0.022	0.125	0.824
Correct Classification Rate	0.000	0.000	0.000	0.000	0.003	0.015	0.093	0.335	0.555
Kappa	0.034	0.019	0.034	0.071	0.131	0.174	0.225	0.158	0.154

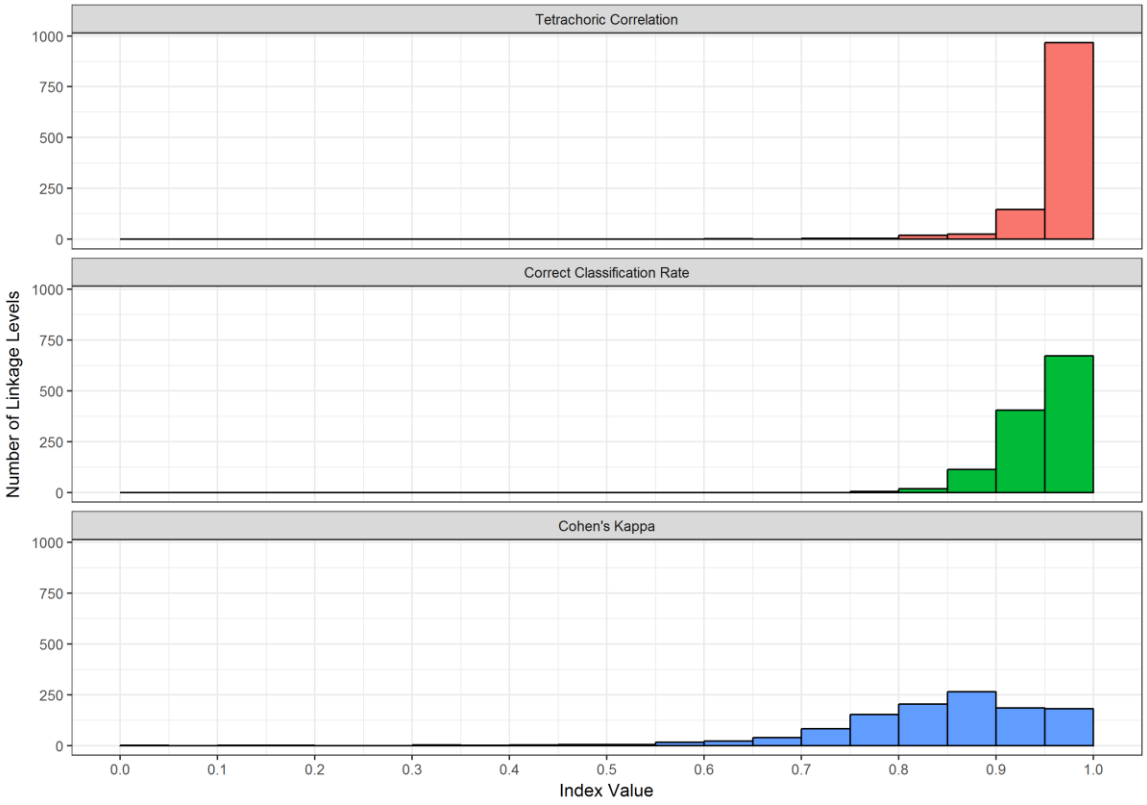


Figure 15. Linkage-level reliability summaries.

VIII.2.F. Conditional-Reliability Evidence by Linkage Level

Traditional assessment programs often report conditional standard errors of measurement to indicate how the precision of measurement differs along the score continuum. The DLM assessment system does not report total or scale score values. However, because DLM

assessments were designed to span the continuum of students' varying skills and abilities as defined by the five linkage levels, evidence of reliability can be summarized for each linkage level to approximate conditional evidence over all EEs, similar to a conditional standard error of measurement for a total score.

Conditional-reliability evidence by linkage level is based on the true and estimated mastery statuses for each linkage level, summarized by each of the five levels. Results are reported using the same three statistics used for the overall linkage-level reliability evidence (i.e., tetrachoric correlation, correct classification rate, and kappa).

Figure 16 provides the number of linkage levels that fall within prespecified ranges of values for the three reliability summary statistics (i.e., tetrachoric correlation, correct classification rate, and kappa). The correlations and correct classification rates generally indicate that all three linkage levels provide reliable classifications of student mastery, with fairly consistent results across all linkage levels for each of the three statistics reported.

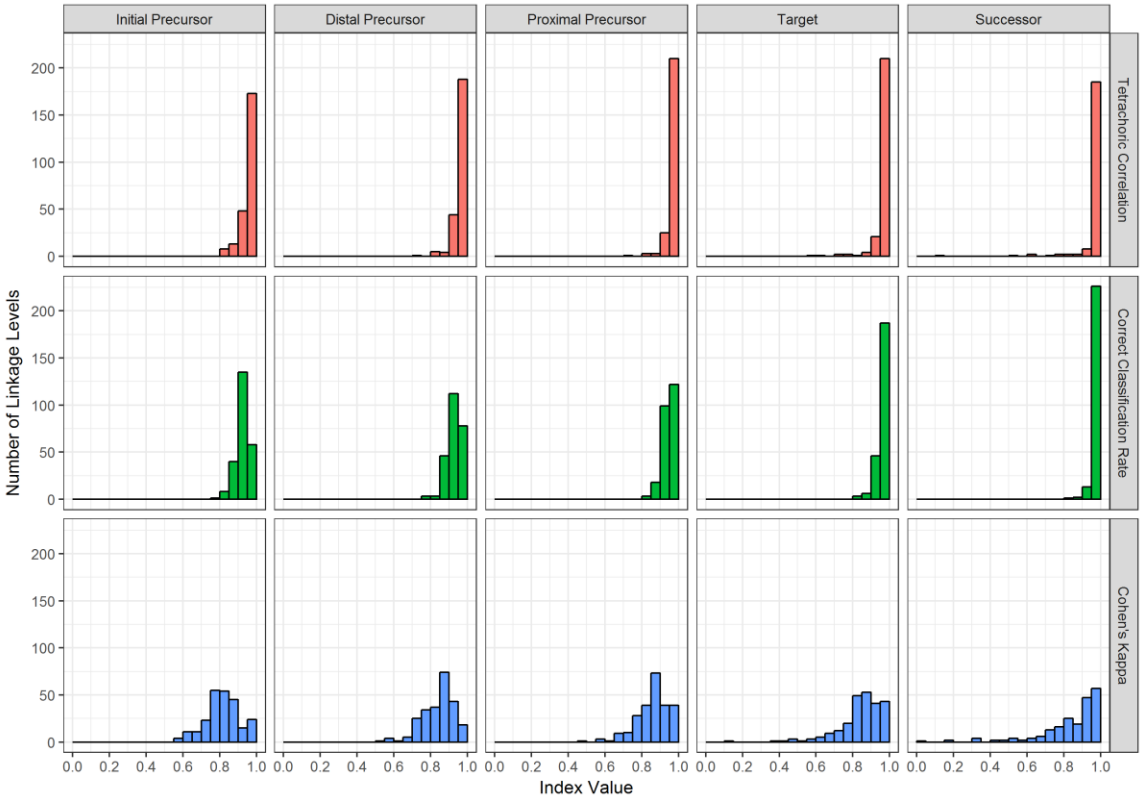


Figure 16. Conditional-reliability evidence summarized by linkage level.

VIII.3. CONCLUSION

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 diagnostic classification modeling and yielded evidence to support the argument for internal consistency of the program for each level of reporting. Because the reliability results are dependent upon

the model used to calibrate and score the assessment, any changes to the model or evidence obtained when evaluating model fit would also impact reliability results. As with any selected methodology for evaluating reliability, the current results assume that the model and model parameters used to score DLM assessments are correct. However, unlike other traditional measures of reliability that often require unattainable assumptions about equivalent test forms, the simulation method described in this chapter provides a replication of the exact same test items (i.e., perfectly parallel forms) which theoretically reduces the amount of variance that may be found in test scores across administrations. Furthermore, while results in general may be higher than those observed for some traditionally scored assessments, research suggests that DCMs have higher reliability with fewer items (e.g., Templin & Bradshaw, 2013), suggesting the results are expected.

IX. VALIDITY STUDIES

The preceding chapters and the *2014-2015 Technical Manual – Year-End Model* (DLM Consortium, 2016) 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 collected during 2015–2016 for three of the four critical sources of evidence as described in *Standards for Educational and Psychological Testing* (AERA et al., 2014): evidence based on test content, internal structure, and consequences of testing. Evidence for the fourth source, response process, along with additional evidence for the other three sources, can be found in Chapter IX of the *2014-2015 Technical Manual – Year-End Model* (DLM Consortium, 2016).

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 validity study presented in this section provides data collected in spring 2016 regarding student opportunity to learn the assessed content. For additional evidence based on test content, including the alignment of test content to content standards via the DLM maps (which underlie the assessment system), see Chapter IX of the *2014-2015 Technical Manual – Year-End Model* (DLM Consortium, 2016).

IX.1.A. Opportunity to Learn

After completing administration of the spring 2016 operational assessments, teachers were invited to complete a survey about the assessment administration process, which included the same items that were available in the spring 2015 survey. All educators who had administered a DLM assessment during the spring 2016 window (N = 20,112) were invited to respond to the survey. State partners announced the availability of the survey and encouraged teachers' participation. A total of 2,320 teachers responded, yielding an overall response rate of 11.5%, a decrease of 1.2 percentage points from 2015. Future teacher surveys are planned for administration within the Kansas Interactive Testing Engine (KITE®), which will allow for examining the representativeness of the sample of responding teachers and improve ease of responding, which should improve responses rates in 2017 and beyond.

The survey served several purposes.[9] One item 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 survey asked teachers to indicate whether they judged the test content, across all testlets, to be aligned with their instruction. Table 46 reports the results. Overall, the frequency distribution ranged from no

[9]Results for other items are reported in Chapter IV in this manual and later in this chapter.

testlets matching instruction to five or more testlets matching in both mathematics and English language arts (ELA). More specific measures of instructional alignment are planned.

Table 46. *Number of Testlets That Matched Instruction, Spring 2016*

	ELA		Mathematics	
Number of Testlets	***n***	**%**	***n***	**%**
0	112	5.8	110	5.9
1	208	10.7	260	13.9
2	250	12.9	331	17.6
3	306	15.8	334	17.8
4	341	17.6	294	15.7
5 or more	723	37.3	547	29.1

Note. Students receive up to seven testlets during the spring window.

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 et al., 2014). The validity study presented in this section provides survey data collected in spring 2016 regarding teacher feedback on students' abilities to respond to testlets. For additional evidence based on response process, including studies on student and teacher behaviors during testlet administration, and evidence of fidelity of administration, see Chapter IX of the *2014-2015 Technical Manual – Year-End Model* (DLM Consortium, 2016).

IX.2.A. Evaluation of Test Administration

Teachers provided feedback after administering spring operational assessments in 2016. Survey data that inform evaluations of assumptions regarding response processes include teacher perceptions of student ability to respond as intended, free of barriers, and teacher perceptions of the ease of administering teacher-administered testlets.[10]

The spring 2016 teacher survey included three items about students' ability to respond. Teachers were asked to rate statements from *Strongly Disagree* to *Strongly Agree* for the students with the best and worst experiences. Results are combined in the summary presented in Table 47. The majority of teachers agreed or strongly agreed that their students (a) responded to items to the best of their knowledge ability; (b) were able to respond regardless of disability, behavior, or health concerns; and (c) had access to all necessary supports to participate. These results are similar to those observed in 2015.

[10]Recruitment and response information for this survey was provided earlier in this chapter.

Table 47. *Teacher Perceptions of Student Experience with Testlets, Spring 2016*

Statement	Strongly Disagree		Disagree		Agree		Strongly Agree	
	n	%	*n*	%	*n*	%	*n*	%
Student responded to items to the best of his/her knowledge and ability.	304	8.2	436	11.8	2,024	54.6	944	25.5
Student was able to respond regardless of his/her disability, behavior, or health concerns.	552	15.0	643	17.4	1,860	50.4	636	17.2
Student had access to all necessary supports to participate.	197	5.3	293	7.9	2,108	57.1	1,094	29.6

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

DIF analyses for 2016 followed the same procedure as 2015, including data from both the 2014–2015 year and the 2015–2016 year to flag items for evidence of DIF. As additional data are 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 DIF analyses based on minimum sample-size requirements for the two gender subgroups: male and female. 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 retained from 2015 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 DIF analyses described here because they are scored at the option level rather than item level and include nonindependent response options (see Chapter III of this manual for more information). Only operational content meeting sample-size thresholds was included in the DIF analyses.

Two additional criteria were included for the 2015–2016 year to prevent estimation errors from occurring. Items with an overall *p* value (or proportion correct) greater than .95 were removed from the analyses. Additionally, items in which one gender group had a *p* value greater than .97 were also removed from the analyses.

Using the above criteria for inclusion, 2,848 (45%) items on Year-End model testlets were selected for inclusion in the analysis. In the year-end model (multi-EE testlets), the number of items evaluated by grade level and content area ranged from 116 items in seventh grade for ELA to 214 items in sixth-grade mathematics. Item sample sizes were between 267 and 7,962.

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 composed of the student's total linkage levels mastered in the content area of the item and a group membership variable, with females coded 0 as the focal group and males coded 1 as the reference group. An interaction term was included to evaluate whether nonuniform 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. When nonuniform 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,

$$M_0: logit(\pi_i) = \alpha + \beta X + \gamma_I + \delta_i X$$

$$M_1: logit(\pi_i) = \alpha + \beta X + \gamma_I$$

$$M_2: logit(\pi_i) = \alpha + \beta X,$$

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 and Thomas (1997) and Jodoin and Gierl (2001) indices for reflecting a negligible, moderate, or large effect. The Zumbo and Thomas thresholds for classifying DIF effect size are based on Cohen's (1992) guidelines for identifying a small, medium, or large effect. The thresholds for each level are 0.13 and 0.26 for distinguishing negligible, moderate, and large effects, whereby items with an effect size less than 0.13 are classified as having a negligible effect, values between 0.13 and 0.26 are classified as having moderate effect, and values of 0.26 or greater are classified as having a large effect.

The Jodoin and Gierl approach expanded on the Zumbo and 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 nonuniform DIF and makes use of classification guidelines that are based on the widely accepted ETS Mantel–Haenszel classification guidelines. The Jodoin and Gierl threshold values for distinguishing negligible, moderate, and large DIF are more stringent than the Zumbo and Thomas approach, with lower threshold values of 0.035 and 0.07 to distinguish negligible, moderate, and large effects. Similar to the ETS method, negligible effect is classified with an A, moderate effect with a B, and large effect with a C, for both methods.

Jodoin and 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

IX.3.A.ii.a Uniform DIF Model

A total of 70 items were flagged for evidence of uniform DIF when comparing M_1 to M_2. Table 48 summarizes the total number of items flagged for evidence of uniform DIF by content area and grade for each model. The percentage of items flagged for uniform DIF for each grade and content area ranged from 5.9 to 19.3.

Table 48. *Items Flagged for Evidence of Uniform DIF, Spring 2016*

Content Area	Grade	Items Flagged	Total Items	% Flagged	Number with Moderate or Large Effect Size
ELA	3	19	147	12.9	0
	4	12	152	7.9	0
	5	11	167	6.6	0
	6	21	137	15.3	0
	7	10	116	8.6	0
	8	20	123	16.3	0
	9	22	122	18.0	0
	10	10	134	7.5	0
	11	23	155	14.8	0
Math	3	27	153	17.6	0
	4	27	176	15.3	0
	5	30	186	16.1	0
	6	20	214	9.3	1
	7	32	188	17.0	0
	8	37	192	19.3	0
	9	21	164	12.8	0
	10	9	153	5.9	0
	11	19	169	11.2	0

Using the Zumbo and Thomas (1997) effect-size classification criteria, all 70 items were found to have a negligible change in effect size after adding the gender term to the regression equation.

Using the Jodoin and Gierl (2001) effect-size classification criteria, one item was found to have a moderate effect-size change and the remaining 69 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 49. The one mathematics item that had a moderate effect-size value is represented by a value of B. The γ values in Table *49* indicate which group was favored on the item after holding total linkage levels mastered constant, with positive values indicating that the focal group (females) had a higher probability of success on the item. The one item favored male students.

Table 49. *Item Flagged for Uniform DIF with Moderate or Large Effect Size, Spring 2016*

Content area	Grade	Item ID	EE	χ^2	*p* value	γ	R^2	Z & T effect size	J & G effect size
Math	6	57643	6.RP.1	6.89	0.01	-1.11	0.04	A	B

Note. Z& T = Zumbo and Thomas, J & G = Jodoin and Gierl.

Combined Model. A total of 436 items were flagged for evidence of DIF when both the gender and interaction terms were included in the regression equation. Table 50 summarizes the number of items flagged by content area and grade. The percentage of items flagged for each grade and content area ranged from 7.2% to 24.2%.

Table 50. *Items Flagged for Evidence of DIF for the Combined Model, Spring 2016*

Content Area	Grade	Items Flagged	Total Items	% Flagged	Number Moderate or Large Effect Size
ELA	3	20	147	13.6	1
	4	18	152	11.8	0
	5	12	167	7.2	0
	6	22	137	16.1	0
	7	16	116	13.8	0
	8	25	123	20.3	0
	9	22	122	18.0	0
	10	15	134	11.2	0
	11	31	155	20.0	0
Math	3	37	153	24.2	1
	4	37	176	21.0	0
	5	29	186	15.6	0
	6	30	214	14.0	1
	7	41	188	21.8	0
	8	41	192	21.4	0
	9	24	164	14.6	0
	10	12	153	7.8	0
	11	31	169	18.3	0

Using the Zumbo and Thomas (1997) effect-size classification criteria, two items were found to have a large change in effect size. The remaining 434 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 and Gierl (2001) effect-size classification criteria, one item was found to have a moderate change in effect size, two items were found to have a large change in effect size, and the remaining 433 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 51 and Table 52 for ELA and mathematics respectively. One ELA item and two mathematics items had moderate or large changes in effect-size values, as represented by a value of B or C respectively. Two items favored the male group (as indicated by a negative γ value) and one item favored the female group (as indicated by a positive γ value).

Table 51. *ELA Item Flagged for DIF with Moderate or Large Effect Size, Spring 2016*

Grade	Item ID	EE	χ^2	*p* value	γ	$\delta_i X$	R^2	Z & T *	J & G*
3	25604	RL.3.5	6.20	.05	0.01	0.03	0.93	C	C

Note. *Effect-size measure. Z& T = Zumbo and Thomas, J & G = Jodoin and Gierl.

Table 52. *Mathematics Items Flagged for DIF with Moderate or Large Effect Size, Spring 2016*

Grade	Item ID	EE	χ^2	*p* value	γ	$\delta_i X$	R^2	Z & T *	J & G*
3	24724	3.OA.4	11.84	<.01	-0.14	0.06	0.93	C	C
6	57643	6.RP.1	7.10	.03	-2.35	0.03	0.04	A	B

Note. *Effect-size measure. *Note.* Z& T = Zumbo and Thomas, J & G = Jodoin and Gierl.

A comparison of results from 2014-2015 to 2015-2016 indicates no items flagged in 2015 were also flagged in 2016, after the collection of an additional year's worth of data.

Appendix F includes plots 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.

IX.3.A.iii. Test-Development Team Review of Flagged Items

The test-development teams for each content area were provided with data files that listed all items flagged with a moderate or large effect size. Files provided to the test-development teams did not indicate which group was favored, so as not to bias their review of the items.

During their review of the flagged items, test-development teams were asked to consider facets of each item that may lead one gender group to provide correct responses at a higher rate than the other. Because DIF is closely related to issues of fairness, the bias and sensitivity external

review criteria (see Chapter III of the *2014-2015 Technical Manual – Year-End Model*) were provided to the test-development teams for their consideration as they reviewed the items. After reviewing the flagged item and considering its context in the testlet, including the ELA text and the engagement activity in mathematics, content teams were asked to provide one of three decision codes for each item.

1. Accept: No evidence of bias favoring one group or the other. Leave content as is.
2. Minor revision: Clear indication that a fix will correct the item, if the edit can be made within the allowable edit guidelines.
3. Reject: There is evidence the item favors one gender group over the other. There is not an allowable edit to correct the issue. The item is slated for retirement.

After their review, all items flagged for moderate or large effect size were provided with a decision code of 1: Accept. No evidence could be found in any of the items indicating the content favored one gender group over the other.

IX.4. 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). To establish sound score interpretations, the assessment must measure important content that informs instructional choices and goal setting.

During spring 2016, one additional source of evidence was collected via teacher survey responses. Additional consequential evidence will be collected in subsequent years.

IX.4.A. Teacher Survey Responses

Teachers were asked two questions on the spring 2016 survey[11] that assessed their perceptions of the assessment contents. Teachers 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 administered a DLM assessment to only one student responded only once. Table 53 summarizes the responses across all students: best experience, worst experience, and only student. Teachers 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), teachers' responses may reflect an awareness that DLM assessments contain challenging content. However, they are divided on its importance in the educational programs of students with the most significant cognitive disabilities.

[11]Recruitment and sampling described earlier in this chapter.

Table 53.*Teacher Perceptions of Student Experience with Testlets, Spring 2016*

Statement	Strongly Disagree		Disagree		Agree		Strongly Agree	
	n	%	*n*	%	*n*	%	*n*	%
Content measures important academic skills.	764	20.6	965	26.0	1,741	46.8	247	6.7
Content reflects high expectations for this student.	424	11.4	782	21.1	1,997	53.8	511	13.8

IX.5. 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, 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 the *2014-2015 Technical Manual – Year-End Model* (DLM Consortium, 2016), Chapter XI, references evidence presented through the 1014-2015 technical manual, including Chapter IX, and expands the discussion of the overall validity argument. The chapter also provides areas for further inquiry and ongoing evaluation of the DLM Alternate Assessment System.

X. TRAINING AND PROFESSIONAL DEVELOPMENT

Chapter X describes the training that was offered in 2015–2016 for state and local education agency staff, the required test administrator training, and the optional professional development provided. Participation rates and evaluation results from 2015–2016 instructional professional development are included in this chapter (see Table 54 and Table 55 at the end of the chapter).

For a complete description of training and professional development for Dynamic Learning Maps® (DLM®) assessments, including a description of training for state and local education agency staff, along with descriptions of facilitated and self-directed training, see Chapter X of the *2014–2015 Technical Manual – Year-End* (DLM Consortium, 2016).

X.1. REQUIRED TRAINING FOR TEST ADMINISTRATORS

Training is required annually for educators who serve as test administrators and administer the DLM alternate assessments. In 2015–2016, training was available in two formats: facilitated training (in-person training with quizzes in Moodle) and self-directed training (all content and quizzes within Moodle). The switch to Moodle (from Educator Portal in 2014–2015) was implemented due to ease of use for test administrators and the ability to more effectively manage data captured by the system.

All new test administrators were required to successfully complete modules before beginning testing; they were not allowed access to 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 four modules and pass all four posttests with a score of 80% or higher. Test administrators were able to retake posttests as many times as needed to pass all parts of the training.

Returning test administrators had to successfully complete a single combined module with a score of 80% or higher on each of four posttests before being allowed access to their students' log-in information. Training time was estimated at less than 1 hour. If the module posttest was not successfully completed on the first attempt, additional training was required. The additional training could take an extra 30 minutes to 4 hours, depending on the areas in which the test administrator was not successful on the first attempt.

For a complete description of required training for test administrators, see Chapter X of the *2014-2015 Technical Manual – Year-End Model* (DLM Consortium, 2016).

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 17 illustrates the differences between the two training formats.

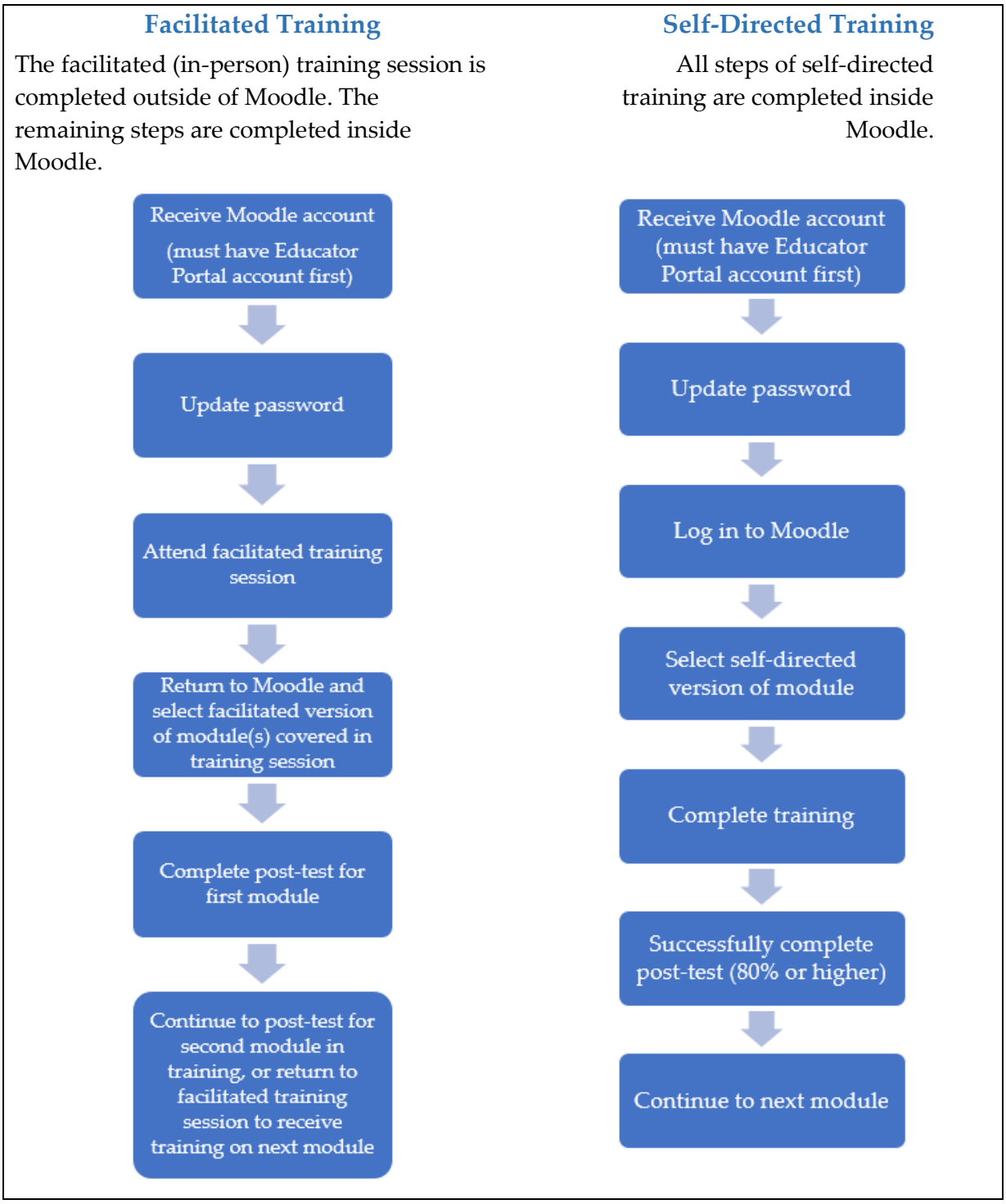


Figure 17. Required training process flows for facilitated and self-directed training.

X.1.A. Training Content

Training content was updated for 2015–2016 from the content available in 2014–2015. The seven modules available in 2014–2015 were reduced to four modules in 2015–2016. Module content

was combined, content was made more concise where possible, and unnecessary content was removed. The four resulting modules are described in the sections that follow.

X.1.A.i. Module 1: About the DLM System

Module 1 of the test administrator training provided an overview of the DLM system components and DLM test security. Topics included illustration and discussion of the DLM maps, claims and conceptual areas, Essential Elements (EEs), testlets, linkage levels, and the security demands of the DLM system. 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 EEs and differentiate them from functional skills. Participants were to be able to define claims and place them within the context of instructional practice. Finally, educators were expected to practice the security guidelines for assessments as outlined in Module 1.

Module 1 explained how the DLM testlets were developed. It also emphasized that Target-level testlets are aligned directly to the EE 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 (i.e., Initial Precursor, Distal Precursor, Proximal Precursor, Target, and Successor) in their assessment, whether ELA or mathematics. Participants were introduced to mini-maps that specifically detail the nodes that are assessed at each linkage level.

After viewing Module 1, participants were expected to know all DLM security standards, which 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.1.A.ii. Module 2: Accessibility by Design

Module 2 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, KITE Client, was introduced, along with explanation of its accessibility supports, including guidelines for selecting accessibility supports for the Personal Needs and Preferences Profile (PNP).

Participants were expected to demonstrate understanding of test accessibility supports, their purpose, student eligibility, and appropriate practice. In addition, participants were shown how to complete the PNP and how the PNP and First Contact (FC) survey responses combined to develop a personal learning profile to guide administration decisions for each student.

Module 2 also demonstrated how to actualize all accessibility supports for an individual student, both within KITE Client and through external supports, in conjunction with Testlet Information Pages (TIPs).

Module 2 addressed flexibility in the ways that students access the items and materials, including flexibility that is considered appropriate (e.g., test administrator adapts the physical arrangement of the response options) and flexibility that 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 accessibility supports that are specifically mentioned in the child's IEP.

X.1.A.iii. Module 3: Understanding and Delivering Testlets in the DLM System

Module 3 focused on participants' understanding and delivery of content through testlets within KITE Client. Topics included testlet structure, item types, completing testlets, standard test administration process, allowable practices, and practices to avoid.

The third module provided participants with focused information on how the assessments are delivered via computer. Content 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.

Module 3 also addressed teacher-administered testlets, including the specific structures used and the processes for completing testlets by administering them outside KITE Client. The module also covered how the test administrator entered responses into 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).

X.1.A.iv. Module 4: Preparing to Administer the Assessment

Module 4 prepared participants for their role as test administrators. They learned to check data, complete the FC, use practice activities and released 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 2015-2016* (DLM Consortium, 2015). Participants reviewed the DLM assessment components, which are accessed through the Educator Portal (e.g., FC 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 FC, which is completed before testing begins. It uses test administrator responses to questions about student communication and academic skills to determine at which linkage level it is best to start students the first time they encounter the DLM assessments. The FC is completed online, but test administrators also have access to all the questions in advance in an appendix to the *Test Administration Manual 2015-2016.* The FC 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.

The module also addressed planning and scheduling the assessments. Prior to the assessments, test administrators were directed to allow their students taking the assessments to complete practice activities to expose them to KITE Client. Test administrators were advised to retrieve TIPs, determine the appropriate length of each assessment session, and consider the schedules according to their states' requirements. Test administrators were also instructed to arrange a space for assessments that is quiet, clear from distractions, and able to accommodate students' accessibility needs.

X.2. Instructional Professional Development

The DLM Professional Development System includes approximately 50 modules, including 20 focused on English language arts instruction, 25 focused on mathematics instruction, and five others that address individual education programs, the DLM claims and conceptual areas, Universal Design for Learning, DLM EEs, and the Common Core State Standards. The complete list of module titles is included in Table 55. The modules are available in two formats, self-directed and facilitated, and are accessed at http://dlmpd.com.

The self-directed modules were designed to meet the needs of all educators, especially those in rural and remote areas, to offer educators just-in-time, on-demand training. The self-directed modules are available online via an open-access, interactive portal and combine videos, text, student work samples, 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. Each module ends with a posttest, and educators who achieve a score of 80% or higher on the posttest receive a certificate via email.

The facilitated modules are intended for use with groups. This version of the modules was designed to meet the need for face-to-face training without requiring a train-the-trainers approach. Instead of requiring trainers to themselves be subject matter experts in content related to academic instruction and the population of students with the most significant cognitive disabilities, the facilitated training is delivered via recorded video created by subject matter experts. Facilitators are provided with an agenda, a detailed guide, handouts, and other supports required to facilitate meaningful, face-to-face training. By definition, they are facilitating training developed and provided by members of the DLM professional development team.

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 and deliver the training via video.

X.2.A. Professional Development Participation and Evaluation

As reported in Table 54, a total of 92,439 modules were completed in the self-directed format from fall 2012, when the first module was launched, until September 30, 2016. This is an increase of 14,120 modules since September 30, 2015 (78,319 modules completed). Data are 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, and West Virginia) use the facilitated modules extensively.

Table 54. *Number of Self-Directed Modules Completed by Educators in DLM States and Other Localities through September 2016.*

State	Total Self-Directed Modules Completed (*n*)
Missouri	21,377
Kansas	16,778
Mississippi	14,040
New Jersey	8,995
Colorado	4,781
Wisconsin	4,225
North Carolina	2,950
Utah	2,395
Illinois	2,162
Oklahoma	1,733
Vermont	1,139
Iowa	1,017
Pennsylvania	822
New Hampshire	671
Alaska	607
North Dakota	447

State	Total Self-Directed Modules Completed (*n*)
New York	330
West Virginia	162
Non-DLM states and other locations	7,808
Total	**92,439**

To evaluate educator perceptions of the utility and applicability of the modules, DLM staff asked educators to respond to a series of evaluation questions upon completion of each self-directed module. Through September 2016, on average, educators completed the evaluation questions 77% of the time. The responses are consistently positive, as Table 55 illustrates.

Table 55. *Response Rates and Average Ratings on Self-Directed Module Evaluation Questions*

Note: The first three questions use a 4-point scale. The final question has three response options: No, Maybe, and Yes.

Module name	Total modules completed (*n*)	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?	11,589	0.41	3.44	3.12	3.27	2.77
1: Common Core Overview	6,069	0.35	3.16	2.91	3.09	2.67
2: Dynamic Learning Maps Essential Elements	9,480	0.41	3.33	3.21	3.19	2.74
3: Universal Design for Learning	5,765	0.41	3.34	3.24	3.26	2.75
4: Principles of Instruction in English Language Arts	5,193	0.46	3.30	3.21	3.21	2.76
5: Standards of Mathematics Practice	7,520	0.24	3.25	3.21	3.22	2.72
6: Counting and Cardinality	3,813	0.50	3.36	3.30	3.29	2.75

Module name	Total modules completed (*n*)	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.
7: IEPs Linked to the DLM Essential Elements	4,642	0.43	3.28	3.21	3.22	2.71
8: Symbols	3,362	0.28	3.36	3.29	3.32	2.73
9: Shared Reading	4,598	0.53	3.43	3.35	3.29	2.78
10: DLM Claims and Conceptual Areas	2,706	0.70	3.25	3.10	3.11	2.66
11: Speaking and Listening	2,782	0.50	3.33	3.24	3.23	2.74
12: Writing: Text Types and Purposes	2,875	0.61	3.23	3.16	3.12	2.68
13: Writing: Production and Distribution	1,371	0.92	3.25	3.20	3.19	2.70
14: Writing: Research and Range of Writing	1,646	0.70	3.23	3.19	3.16	2.70
15: The Power of Ten-Frames	1,258	0.93	3.26	3.24	3.20	2.67
16: Writing with Alternate Pencils	1,558	0.91	3.37	3.31	3.29	2.68

Module name	Total modules completed (*n*)	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.
17: DLM Core Vocabulary and Communication	1,618	0.91	3.43	3.37	3.40	2.76
18: Unitizing	860	0.88	3.19	3.13	3.13	2.61
19: Forms of Number	1,033	0.86	3.14	3.10	3.09	2.58
20: Units and Operations	809	0.89	3.13	3.09	3.07	2.57
21: Place Value	835	0.87	3.14	3.10	3.07	2.53
22: Fraction Concepts and Models Part I	690	0.89	3.15	3.12	3.10	2.55
23: Fraction Concepts and Models Part II	582	0.90	3.16	3.13	3.10	2.58
24: Composing, Decomposing, and Comparing Numbers	787	0.84	3.19	3.16	3.16	2.58
25: Basic Geometric Shapes and Their Attributes	748	0.89	3.18	3.14	3.10	2.57
26: Writing Information and Explanation Texts	596	0.92	3.17	3.16	3.16	2.63

Module name	Total modules completed (*n*)	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.
27: Calculating Accurately with Addition	555	0.90	3.17	3.14	3.09	2.58
28: Measuring and Comparing Lengths	332	0.91	3.15	3.10	3.07	2.53
29: Emergent Writing	1,049	0.91	3.37	3.32	3.33	2.73
30: Predictable Chart Writing	493	0.94	3.36	3.31	3.33	2.74
31: Calculating Accurately with Subtraction	341	0.90	3.15	3.13	3.09	2.55
32: Teaching Text Comprehension: Anchor-Read-Apply	589	0.88	3.33	3.27	3.28	2.69
33: Generating Purposes for Reading	421	0.88	3.27	3.23	3.25	2.66
34: Exponents and Probability	214	0.87	3.11	3.11	3.08	2.50
35: Beginning Communicators	973	0.92	3.46	3.31	3.36	2.76
36: Time and Money	358	0.92	3.27	3.20	3.20	2.67

Module name	Total modules completed (*n*)	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.
37: DR-TA and Other Text Comprehension Approaches	369	0.87	3.29	3.26	3.25	2.69
38: Supporting Participation in Discussions	368	0.86	3.29	3.26	3.24	2.66
39: Algebraic Thinking	403	0.92	3.25	3.17	3.17	2.58
40: Composing and Decomposing Shapes and Areas	278	0.90	3.22	3.18	3.17	2.56
41: Writing: Getting Started with Writing Arguments	169	0.91	3.09	3.12	3.07	2.51
42: Calculating Accurately with Multiplication	209	0.86	3.23	3.14	3.12	2.54
43: Perimeter, Volume, and Mass	167	0.89	3.10	3.09	3.05	2.51
44: Writing: Getting Started in Narrative Writing	135	0.92	3.17	3.14	3.10	2.57
45: Patterns and Sequence	139	0.90	3.04	2.98	2.94	2.40

Module name	Total modules completed (*n*)	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	95	0.82	3.00	3.03	3.00	2.38
47: Calculating Accurately with Division	157	0.87	3.26	3.23	3.20	2.59
48: Organizing and Using Data to Answer Questions	89	0.81	3.30	3.28	3.24	2.64
49: Strategies and Formats for Presenting Ideas	176	0.79	3.33	3.29	3.31	2.64
50: Properties of Lines and Angles						
Total	**92,864**					
Average		**0.77**	**3.24**	**3.18**	**3.18**	**2.64**

In addition to the modules, the DLM instructional professional development system includes a variety of other instructional resources and supports. These include DLM EEs unpacking documents; links to dozens of texts that are at an appropriate level of complexity for students who take DLM assessments and are linked to the texts that are listed in Appendix B of the Common Core State Standards; vignettes that illustrate shared reading with students with the most complex needs across the grade levels; supports for augmentative and alternative communication for students who do not have a comprehensive, symbolic communication system; alternate "pencils" for educators to download and use with students who cannot use a standard pen, pencil, or computer keyboard; and links to Pinterest boards and other online supports. The team is currently working on new supports to help teachers understand the Initial and Distal Precursor LLs and how they relate cognitively to the target nodes and DLM EEs.

Finally, the DLM instructional professional development system includes a virtual community of practice that is open to educators, related service providers, families, and others who are seeking support in teaching students with the most significant cognitive disabilities in achieving academic standards. The virtual community of practice allows registered users to create and join groups, post and answer questions, and share instructional resources and materials. The virtual community of practice is monitored and seeded by the DLM professional development team at the University of North Carolina at Chapel Hill.

XI. CONCLUSION AND DISCUSSION

The Dynamic Learning Maps® (DLM®) Alternate Assessment System is based on the core belief that all students should have access to challenging, grade-level academic content. Therefore, the DLM assessments provide students with the most significant cognitive disabilities the opportunity to demonstrate what they know and can do. It is designed to map students' learning throughout the year with items and tasks that are embedded in day-to-day instruction.

The DLM system completed its second operational administration year in 2015–2016. This technical manual provides updated evidence from the 2015–2016 year 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. The contents of this manual address the information summarized in Table 56. For a complete summary of evidence collected for the DLM theory of action, see the *2014-2015 Technical Manual- Year-End Model* (DLM Consortium, 2016).

Table 56. *Review of Technical Manual Contents*

Chapter(s)	Contents
I	Provides an overview of information updated for the 2015–2016 year.
II	Not updated for 2015–2016.
III, IV, X	Provides procedural evidence collected during 2015–2016 of test content development and administration, including field-test information, teacher survey results, and professional development module use.
V	Describes the statistical model used to produce scores based on student responses.
VI	Not updated for 2015–2016.
VII, VIII	Describes results and analysis of the second operational administration's data, evaluating student performance on the assessment, score distributions, aggregated and disaggregated results, and analysis of the internal consistency of student responses.
IX	Provides additional studies from 2015–2016 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 discusses future research studies as part of ongoing and iterative processes of program responsiveness, validation, and evaluation.

XI.1. VALIDITY EVIDENCE SUMMARY

The accumulated evidence available by the end of the 2015–2016 year provides additional support for the validity argument. Each proposition is addressed by evidence in one or more of the categories of validity evidence, as summarized in Table 57. While many sources of evidence support multiple propositions, Table 57 lists the primary associations. For example, Proposition 4 is indirectly supported by content-related evidence described for Propositions 1 through 3. Table 58 shows the titles and sections for the chapters cited in Table 57. A complete summary of evidence can be found in Chapter XI of *2014-2015 Technical Manual – Year End Model* (DLM Consortium, 2016).

Table 57. *DLM Alternate Assessment System Propositions and Sources of Updated Evidence for 2015–2016*

Proposition	Sources of Evidence*				
	Test Content	**Response Processes**	**Internal Structure**	**Relations with Other Variables**	**Consequences of Testing**
Scores represent what students know and can do.	1, 2, 3, 4, 5, 6, 7, 8, 13, 15	9, 16, 19	10, 14, 15, 17		11, 18
Achievement level descriptors provide useful information about student achievement.			14		12
Inferences regarding student achievement, progress, and growth can be drawn at the conceptual area level.	12		14		
Assessment scores provide useful information to guide instructional decisions.					18

Note. *See Table 58 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 58. *Evidence Sources Cited in Previous Table*

Evidence #	Chapter	Section(s)
1	III	English Language Arts Blueprint Coverage
2	III	Item Writer Characteristics
3	III	English Language Arts Passage Development
4	III	English Language Arts Writing Testlets
5	III	Selection of Accessible Graphics for Testlets
6	III	External Reviews
7	III	Operational Assessment Items for 2015–2016
8	III	Field Testing
9	IV	Implementation Evidence
10	V	All
11	VII	Student Performance
12	VII	Score Reports
13	VII	Quality Control Procedures for Data Files and Score Reports
14	VIII	All
15	IX	Evidence Based on Test Content
16	IX	Evidence Based on Response Process
17	IX	Evidence Based on Internal Structure
18	IX	Evidence Based on Consequences of Testing
19	X	Required Training for Test Administrators

XI.2. CONTINUOUS IMPROVEMENT

XI.2.A. OPERATIONAL ASSESSMENT

As noted previously in this manual, 2015–2016 was the second year the DLM Alternate Assessment System was operational. While the 2015–2016 assessments were carried out in a manner that supports the validity of the proposed uses of the DLM information for the intended purposes, the DLM 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 2016–2017. This section describes significant changes from the first to second year of operational administration, as well as examples of improvements to be made during the 2016–2017 year.

Overall, there were no significant changes to learning map models, item-writing procedures, item flagging outcomes, test administration, or the modeling procedure used to calibrate and score assessments from 2014–2015 to 2015–2016.

However, performance differences were observed across years. Specifically, the percentage of students classified to the At Target or Advanced performance levels decreased from 2014–2015 to 2015–2016 in some grades and subjects, after including only states who participated in the year-end model in both years. Prior to delivery of data files and score reports to state partners, the DLM psychometric team ruled out potential sources of error, including scoring issues or systematic changes to the population. Upon discussion of the finding with DLM TAC members and state partners, it was suggested that implementation within each state could be a potential source of the change in performance. Additional explanations included greater fidelity of administration and understanding of allowable practices during the second year of administration and a history of unreliability in performance for students taking alternate assessments, which predates Dynamic Learning Maps. Results will be compared again following the 2016–2017 administration to determine if a trend is evident across three years of results.

Survey results obtained from the spring 2016 survey administration provided feedback and areas for improvement on test development and the Kansas Interactive Testing Engine (KITE®) functionality. Improvements to test development procedures for 2016–2017 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.

Improvements to the 2016–2017 test administration procedures will focus on ensuring a high-quality assessment experience for teachers using Educator Portal. Improvements will be made to the interface to increase usability based on teacher survey feedback collected during 2015–2016.

The validity evidence collected in 2015–2016 expands upon the evidence collected in the first operational year for three of the four critical sources of evidence as described in *Standards for Educational and Psychological Testing* (AERA et al., 2014): evidence based on test content, internal structure, and consequences of testing. Specifically, analysis of blueprint coverage and opportunity to learn contribute to the evidence collected based on test content. Additional teacher survey responses further contributed to the body of evidence collected based on response process. Evaluation of item-level bias via differential item functioning analysis, along with item pool statistics, provided additional evidence collected based on internal structure. Evidence for the fourth source, response process, was not collected during the 2015–2016 year.

Teacher survey responses also provided evidence based on consequences of testing, although further research is still needed to collect additional evidence.

XI.2.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.

A score report interpretation study is planned for 2016–2017 to collect information about how teachers read and interpret DLM score report information. The planned study provides an online, on-demand tutorial for teachers to view to aid in understanding report contents and their instructional uses.

Collection of teacher survey data is also planned for spring 2017 to provide additional longitudinal data as further validity evidence.

In addition, a long-term research plan has been outlined and is underway with the ultimate purpose of designing a data collection and statistical modeling plan that will support node-based estimation. The goal of the approach is to model the relationships and interconnections across nodes such that information about mastery on one tested node can propagate information to other untested nodes based on known relationships represented in the learning map models. This research agenda is being guided by a technical subcommittee of DLM TAC members.

Several initiatives and studies are also planned or underway to support and improve the current linkage-level scoring model. These projects include model-fit analyses that are planned to evaluate how well the response data from the DLM assessments fit the selected latent class statistical model. Model fit will be evaluated using both relative-fit and absolute-fit indices. Also in development are plans to flag items for evidence of misfit, which the test-development team will use to make operational decisions.

Other research is also anticipated as sample sizes increase across subsequent years of operational delivery. For example, DIF analyses, which expanded from 2014–2015 but still did not evaluate all items, may be replicated with different focal and reference groups after the 2016–2017 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, state partners will also begin collaborating to collect additional, state-level validity evidence. For example, states may collect data (e.g., online progress monitoring) that would be appropriate for use in evaluating the relationship between 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.

All future studies will be guided by advice from the DLM TAC and the state partners, using processes established over the life of the DLM Consortium.

XII. REFERENCES

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian Networks in Educational Assessment*. New York: Springer.

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

Bartholomew, D., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (3rd edition). Hoboken, NJ: Wiley.

Bradshaw, L., Izsák, A, & Templin, J., Jacobson, E. (2014). Diagnosing teachers' understanding of rational number: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice, 33 (1)*, 2-14. doi: 10.1111/emip.12020

Camilli, G, & Shepard, L.A. (1994). *Methods for identifying biased test items* (4th ed.). Thousand Oaks, CA: Sage.

Clark, A., Karvonen, M., & Wells Moreaux, S. (2016). *Summary of results from the 2014 and 2015 field test administrations of the Dynamic Learning Maps™ Alternate Assessment System* (Technical Report No. 15-04). Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation.

Clark, A., Swinburne Romine, R., Bell, B., & Karvonen, M. (2015). *Results from external review during the 2015–2016 academic year* (Technical Report No. 15-01). Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation.

Clark, A., Beitling, B., Bell, B., & Karvonen, K. (2016). *Results from external review during the 2015–2016 academic year* (Technical Report No. 16-05). Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155-159.

Cronbach, L. J. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. doi: 10.1177/0146621612445470

Dynamic Learning Maps Consortium. (2015). *Test Administration Manual 2015–2016.* Lawrence, KS: University of Kansas.

Dynamic Learning Maps Consortium. (2016). *2014-2015 Technical Manual – Year-End Model*. Lawrence, KS: University of Kansas.

Dynamic Learning Maps Consortium. (2017). *2015-2016 Technical Manual—Science.* Lawrence, KS: University of Kansas.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282. doi: 10.1007/BF02288892

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement, 29*(4), 262-277.

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

Karvonen, M., Wakeman, S. Y., Browder, D. M., Rogers, M. A., & Flowers, C. (2011). Academic Curriculum for Students with Significant Cognitive Disabilities: Special Education Teacher Perspectives a Decade after IDEA 1997. from ERIC database

Landis, J. R., & Koch G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. doi: 10.2307/2529310

Leighton, J.P., & Gierl, M.J. (Eds.) (2007). *Cognitive diagnostic assessment for education: Theory and practices*. New York: Cambridge University Press.

Li, H. H. & Stout, W.F. (1996). A new procedure for detection of crossing DIP. *Psychometrika,* 61, 647-677.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, *2*(2), 99-120.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64, 197-212.

Mislevy, R. J., & Gitomer, D. H. (1995). The role of probability-based inference in an intelligent tutoring system. *User modeling and user-adapted interaction*, *5* (*3-4*), 253-282.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems.* San Mateo, CA: Morgan Kaufmann.

R Core Team. (2013). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rupp, A., & Templin, J. (2008). Unique characteristics of diagnostic models: a review of the current state-of-the-art. *Measurement, 6*, 219-262.

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

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79,* 317-339.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30*(2), 251–275. doi: 10.1007/s00357-013-9129-4

Templin, J., & Henson, R. (2008, March). Understanding the impact of skill acquisition: relating

diagnostic assessments to measureable outcomes. Paper presented at the 2008 American Educational Research Association conference in New York, New York.

Zumbo, B. D. and Thomas, D. R. (1997). A measure of effect size for a model-based approach for studying DIF. Prince George, Canada: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science.