DYNAMIC®
LEARNING MAPS

2018–2019 Technical Manual Update

Year-End Model

December 2019

Dynamic Learning Maps Consortium. (2019, December). *2018–2019 Technical Manual Update—Year-End Model*. Lawrence, KS: University of Kansas, Accessible Teaching, Learning, and Assessment Systems (ATLAS).

Acknowledgements

The publication of this technical manual update builds upon the documentation presented in the *2014–2015 Technical Manual—Year-End Model* and annual technical manual updates. This document 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. Dynamic Learning Maps® (DLM®) staff who made significant writing contributions to this technical manual update are listed below with gratitude.

W. Jake Thompson, Ph.D., *Psychometrician*
Brianna Beitling, *Psychometrician Assistant*
Amy Clark, Ph.D., *Associate Director for Operational Research*
Brooke Nash, Ph.D., *Associate Director for Psychometrics*

The authors wish to acknowledge Zachary Hopper, Elizabeth Kavitsky, Teri Millar, Noelle Pablo, Courtney Rodgers, Russell Swinburne Romine, and Michelle Shipman for their contributions to this update. For a list of 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, please see the *2014–2015 Technical Manual—Year-End Model*, and the subsequent annual technical manual updates.

We are also grateful for the contributions of the members of the DLM Technical Advisory Committee who graciously provided their expertise and feedback. Members of the Technical Advisory Committee during the 2018–2019 operational year include:

Russell Almond, Ph.D., *Florida State University*
Greg Camilli, Ph.D., *Rutgers University*
Karla Egan, Ph.D., *EdMetric*
Claudia Flowers, Ph.D, *University of North Carolina-Charlotte*
Robert Henson, Ph.D., *University of North Carolina-Greensboro*
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*

Contents

List of Tables

List of Figures

1. Introduction

The 2018–2019 academic year was the fifth operational administration of the Dynamic Learning Maps® (DLM®) Alternate Assessment System. Assessments measured student achievement in mathematics, English language arts (ELA), and science for students with the most significant cognitive disabilities in grades 3 through 8 and high school. Because science was initially implemented on an independent timeline from ELA and mathematics, a separate technical manual update was prepared for science for 2018–2019 (see Dynamic Learning Maps Consortium [DLM Consortium], 2019).

The purpose of the DLM system is to improve academic experiences and outcomes for students with the most significant cognitive disabilities by setting high and actionable academic expectations and providing appropriate and effective supports to educators. Results from the DLM alternate assessment are intended to support interpretations about what students know and are able to do and to support inferences about student achievement in the given subject. Results provide information that can guide instructional decisions as well as information for use with state accountability programs.

The DLM 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 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 after the first operational administration in 2014–2015. After each annual administration, a technical manual update is provided to summarize updated information. The current technical manual provides updates for the 2018–2019 administration. Only sections with updated information are included in this manual. For a complete description of the DLM assessment system, refer to previous technical manuals, including the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016).

1.1. Background

In 2018–2019, DLM assessments were administered to students in 19 states and one Bureau of Indian Education school: Alaska, Arkansas, Colorado, Delaware, District of Columbia, Illinois, Iowa, Kansas, Maryland, Miccosukee Indian School, Missouri, New Hampshire, New Jersey, New York, North Dakota, Oklahoma, Rhode Island, Utah, West Virginia, and Wisconsin.

Two DLM Consortium partners, District of Columbia and Maryland, did not administer operational assessments in ELA or mathematics in 2018–2019.

In 2018–2019, Accessible Teaching, Learning, and Assessment Systems (ATLAS) at the University of Kansas (KU) 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 KU. The project was also supported by a Technical Advisory Committee.

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

1.2. Assessment

Assessment blueprints consist of the Essential Elements (EEs) prioritized for assessment by the DLM Consortium. To achieve blueprint coverage, each student is administered a series of testlets. Each testlet is delivered through an online platform, Kite® Student Portal. 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 (i.e., 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, summative results are based on mastery estimates for linkage levels for each EE (including performance on all instructionally embedded and spring testlets). The pools of operational assessments for the instructionally embedded and spring windows are separate. In 2018–2019, the states participating in the instructionally embedded model included Arkansas, Iowa, Kansas, Missouri, and North Dakota.
- **Year-end model.** During a single operational testing window in the spring, all students take testlets that cover the whole blueprint. Each student is assessed at one linkage level per EE. The linkage level for each testlet varies according to 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 2018–2019, the states participating in the year-end model included Alaska, Colorado, Delaware, Illinois, Miccosukee Indian School, New Hampshire, New Jersey, New York, Oklahoma, Rhode Island, Utah, West Virginia, and Wisconsin.

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

1.3. Technical Manual Overview

This manual provides evidence collected during the 2018–2019 administration to evaluate the DLM Consortium's assertion of technical quality and the validity of assessment claims.

Chapter 1 provides a brief overview of the assessment and administration for the 2018–2019 academic year and a summary of contents of the remaining chapters. While subsequent chapters describe the individual components of the assessment system separately, several key topics are addressed throughout this manual, including accessibility and validity.

Chapter 2 was not updated for 2018–2019; no changes were made to the learning map models used for operational administration of DLM assessments. See the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016) for a description of the DLM map-development process.

Chapter 3 outlines evidence related to test content collected during the 2018–2019 administration, including a description of test development activities and the operational and field test content available.

Chapter 4 provides an update on test administration during the 2018–2019 year. The chapter provides updated Personal Needs and Preferences Profile selections, a summary of administration time, an updated writing testlet assignment process, updated adaptive routing analyses and teacher survey results regarding educator experience and system accessibility.

Chapter 5 provides a brief summary of the psychometric model used in scoring DLM assessments. This chapter includes a summary of 2018–2019 calibrated parameters and mastery assignment for students. For a complete description of the modeling method, see *2015–2016 Technical Manual Update—Year-End Model* (DLM Consortium, 2017a).

Chapter 6 was not updated for 2018–2019; no changes were made to the cut points used in scoring DLM assessments. See the *2014–2015 Technical Manual—Year-End Model* (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.

Chapter 7 reports the 2018–2019 operational results, including student participation data. The chapter details the percentage of students at each performance level; subgroup performance by gender, race, ethnicity, and English-learner status; and the percentage of students who showed mastery at each linkage level. Finally, the chapter provides descriptions of changes to data files during the 2018–2019 administration.

Chapter 8 summarizes reliability evidence for the 2018–2019 administration, including a brief overview of the methods used to evaluate assessment reliability and results by performance level, subject, conceptual area, EE, linkage level, and conditional linkage level. For a complete description of the reliability background and methods, see *2015–2016 Technical Manual Update—Year-End Model* (DLM Consortium, 2017a).

Chapter 9 describes additional validation evidence collected during the 2018–2019 administration not covered in previous chapters. The chapter provides study results for four of the five critical sources of evidence: test content, internal structure, response process, and consequences of testing.

Chapter 10 describes the professional development offered across the DLM Consortium in 2018–2019, including participation rates and evaluation results. There were no updates to training in 2018–2019.

Chapter 11 synthesizes the evidence from the previous chapters. It also provides future directions to support operations and research for DLM assessments.

2. 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 refine the DLM maps in light of the Common Core State Standards and the needs of the student population, see Chapter 2 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016).

3. Item and Test Development

Chapter 3 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016) describes item and test development procedures. This chapter provides an overview of updates to item and test development for the 2018–2019 academic year. The first portion of the chapter provides an overview of 2018–2019 item writers' characteristics. The next portion of the chapter describes the pool of operational and field test testlets administered during spring 2019.

For a complete description of item and test development for 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).

3.1. Items and Testlets

This section describes information pertaining to items and testlets administered as part of the DLM assessment system, including a brief summary of item writer demographics and duties for the 2018–2019 year and an analysis of answer option selection. For a complete summary of item and testlet development procedures that began in 2014–2015 and were implemented through 2018–2019, see Chapter 3 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016).

3.1.1. Item Writing

For the 2018–2019 year, items were written to replenish the pool. The item writing process for 2018–2019 began with an on-site event in January 2019. Following this initial event, item writing continued remotely via a secure online platform. A single pool of item writers was trained to write both single-Essential Element (EE) and multi-EE testlets to expand the operational pool. A total of 202 single-EE testlets were written for English language arts (ELA), and 234 were written for mathematics.

3.1.1.1. Item Writers

An item writer survey was used to collect demographic information about the teachers and other professionals who were hired to write DLM testlets. In total, 27 item writers wrote testlets for the 2018–2019 year, including 13 for mathematics and 14 for ELA. The median and range of years of teaching experience is shown in Table 3.1. The median years of experience was at least 10 years for item writers of both ELA and mathematics testlets in pre-K–12 and special education.

Table 3.1. Item Writers' Years of Teaching Experience

	English language arts		Mathematics	
Area	**Median**	**Range**	**Median**	**Range**
Pre-K–12	18	7-39	12	6-25
English Language Arts	16	8-29	8	2-25
Mathematics	14	5-29	9	2-25
Special Education	10	5-28	10	1-15

The level and types of degrees held by item writers are shown in Table 3.2 and Table 3.3, respectively. All item writers who completed the demographics survey held at least a Bachelor's degree, with the most common field of study being education (*n* = 11; 42%), followed by special education (*n* = 6; 23%). A majority (*n* = 25; 96%) also held a Master's degree, for which the most common field of study was special education (*n* = 13; 52%).

Table 3.2. Item Writers' Level of Degree

	English language arts		Mathematics	
Degree	***n***	**%**	***n***	**%**
Bachelor's	13	100.0	13	100.0
Master's	13	100.0	12	92.3
Missing	1	7.1	0	0.0

Table 3.3. Item Writers' Degree Type

	English language arts	Mathematics
Degree	***n***	***n***
Bachelor's Degree		
Education	4	7
Content Specific	1	0
Special Education	3	3
Other	4	1
Missing	1	2
Master's Degree		
Education	0	2
Content Specific	0	0
Special Education	7	6
Other	6	4
Missing	1	0

Item writers reported a range of experience working with students with different disabilities, as summarized in Table 3.4. Teachers collectively had the most experience working with students with a significant cognitive disability, mild cognitive disability, multiple disabilities, specific learning disability, or other health impairment.

Table 3.4. Item Writers' Experience with Disability Categories

	English language arts		Mathematics	
Disability Category	*n*	%	*n*	%
Blind/Low Vision	3	21.4	6	46.2
Deaf/Hard of Hearing	2	14.3	8	61.5
Emotional Disability	5	35.7	9	69.2
Mild Cognitive Disability	8	57.1	9	69.2
Multiple Disabilities	8	57.1	9	69.2
Orthopedic Impairment	1	7.1	5	38.5
Other Health Impairment	8	57.1	9	69.2
Significant Cognitive Disability	9	64.3	11	84.6
Specific Learning Disability	7	50.0	10	76.9
Speech Impairment	6	42.9	8	61.5
Traumatic Brain Injury	1	7.1	4	30.8
Not reported	5	35.7	2	15.4

3.1.2. *Items*

During 2018–2019, we answer-option selection was analyzed for the operational pool. All computer-delivered multiple-choice items contain three answer options, one of which is correct. Students may select only one answer option. Most answer options are words, phrases, or sentences. For items that evaluate certain learning targets, answer options are images. All teacher-administered items contain five answer options, and educators select the option that best describes the student's behavior in response to the item.

Items typically begin with a stem, which is the question or task statement itself. Each stem is followed by the answer options, which vary in format depending on the nature of the item. Answer options are presented without labels (e.g., A, B, C) and allow students to directly indicate their chosen responses. Computer-delivered testlets use multiple-choice items. Answer options for computer-delivered multiple-choice items are ordered according to the following guidelines:

- Single-word answer options are arranged in alphabetical order.
- Answer options that are phrases or sentences are arranged by logic (e.g., order as appears in a passage, stanza, or paragraph; order from key, chart, or table; chronological order; atomic number from periodic table; etc.), or, if no logical alternative is available, by length from shortest to longest.
- The order may be rearranged to avoid creating a pattern if following these guidelines results in consistently having the first (or the second or the third) option as the key for all items in a testlet.

Teacher-administered item answer options are presented in a multiple-choice format often called a Teacher Checklist. These checklists typically follow the outline below:

- The first answer option is the key.
- The second answer option reflects an incorrect option.
- The third answer option reflects the student choosing both answer options (i.e., the key and the incorrect option).
- The second-to-last answer option usually is "Attends to other stimuli."
- The last answer option usually is "No response."

Refer to Chapter 3 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016) for a complete description of the design of computer-delivered and teacher-administered testlets.

We evaluated the current operational item pool[2] to determine the number of items for which each answer option (A, B, or C) was the correct option, also called the key. As mentioned, the first answer option is always the key for all teacher-administered items (i.e., items measuring the initial linkage level); therefore, Table 3.5 shows the number and percentage of items for which each answer option is the key for computer-administered items. Across items, the key was fairly evenly distributed between the three answer options.

Table 3.5. Number and Percentage of Computer-Delivered Items by Answer Key

	Distal Prescursor		**Proximal Precursor**		**Target**		**Successor**	
Answer Key	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
A	608	38.8	498	32.2	493	31.0	250	32.6
B	488	31.1	519	33.5	554	34.8	269	35.1
C	472	30.1	530	34.3	544	34.2	248	32.3

An additional analysis was conducted to determine if item difficulty differed by answer key. A weighted *p*-value was calculated for items with each answer option as the key, weighted by each item's sample size. Table 3.6 presents the weighted *p*-values for computer-delivered three-option multiple-choice items. Results suggest that items that have B as the answer key may be, on average, slightly more difficult than items where A or C is the key. Because *p*-values are sample-dependent; therefore, values are not directly comparable to one another. In other words, fluctuations in *p*-values may also reflect differences in the samples of students who took the items.

Table 3.6. Weighted *p*-values by Answer Key for Computer-Delivered Items

	Distal Precursor		**Proximal Precursor**		**Target**		**Successor**	
Answer Key	***p*-value**	**SE**	***p*-value**	**SE**	***p*-value**	**SE**	***p*-value**	**SE**
A	0.77	0.001	0.82	0.001	0.89	0.001	0.93	0.001
B	0.71	0.001	0.82	0.001	0.90	0.001	0.93	0.001
C	0.73	0.001	0.86	0.001	0.91	0.001	0.95	0.001

[2] These analyses include items that were in the operational item pool and administered during the testing window.

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

The process for external review 2017–2018 was updated from external reviews in the previous three review cycles. Changes included hosting an on-site event, the training process for external reviewers, and not having power reviewers.

3.2.1. Review Recruitment, Assignments, and Training

In April 2018, a volunteer survey was used to recruit external review panelists. Volunteers for the external review process completed the Qualtrics survey to capture demographic information as well as information about their education and experience. The candidates were screened by the implementation and test development teams to ensure they qualified. These data were then used to identify panel types (content, bias and sensitivity, and accessibility) for which the candidate would be eligible. A total of 17 individuals from integrated-model states and 22 individuals from year-end states were placed on external review panels for ELA and mathematics. All panelists reviewed single-EE testlets.

Each reviewer was assigned to one of the three panel types. There were 20 ELA reviewers: 6 on accessibility panels, 11 on content panels, and 3 on bias and sensitivity panels. There were 19 mathematics reviewers: 7 on accessibility panels, 9 on content panels, and 3 on bias and sensitivity panels.

Panelists completed 6 rounds of reviews. Each round consisted of 1 collection of testlets that ranged from 6 testlets to 26 testlets, dependent on the panel type. Content panels had the smallest number of testlets per collection, and bias and sensitivity panels had the largest number of testlets per collection.

The professional roles reported by the 2018–2019 reviewers are shown in Table 3.7. Reviewers who reported "Other" roles included state education agency (SEA) staff and specialized teachers.

Table 3.7. Professional Roles of External Reviewers

	English language arts		**Mathematics**	
Role	***n***	**%**	***n***	**%**
Classroom Teacher	12	60.0	15	78.9
District Staff	4	20.0	1	5.2
Instructional Coach	1	5.0	0	0.0
Other	3	15.0	3	15.8

Reviewers had varying experience teaching students with the most significant cognitive disabilities. ELA reviewers had a median of 14 years of experience, with a minimum of 5 and a maximum of 34 years of experience. Mathematics reviewers had a median of 11 years of experience teaching students with the most significant cognitive disabilities, with a minimum of 1 and a maximum of 25 years of experience. The population density of schools in which reviewers taught or held a position is reported in Table 3.8. 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 3.8. Population Density for Schools of External Reviewers

	English language arts		Mathematics	
Population Density	*n*	%	*n*	%
Rural	5	25.0	6	31.6
Suburban	12	60.0	7	36.8
Urban	2	10.0	5	26.3
Not Applicable	1	5.0	1	5.3

Prior to attending the on-site external review event, panelists completed an advance training course. The course included two modules that all panelists had to complete: DLM Overview and External Review Process. After each module, the panelists had to complete a quiz and receive a score of at least 80% to continue to the next module. After completing the first two modules and quizzes, each panelist was then directed to a module and quiz that was catered towards their subject and panel type. While the bias and sensitivity and accessibility modules were universal for all subjects, each content module was subject-specific. Panelists were required to complete advance training prior to reviewing any testlets at the event.

Review of testlets was completed during the two day on-site training. The panelists reviewed each testlet on their own and then reviewed them together as a group. Each group came to a consensus for each item and testlet, and the facilitator recorded that recommendation for the test development teams to consider.

3.2.2. *Results of Reviews*

Most of the externally reviewed content was included in the 2019 fall and 2020 spring windows. For ELA, the percentage of items and testlets rated as ***accept*** across grades, panels, and rounds of review ranged from 65% to 98% and 55% to 93%, respectively. The percentage of items and testlets rated as ***revise*** across grades, panels, and rounds of review ranged from 2% to 32% and 7% to 36%, respectively. The rate at which items and testlets were recommended for rejection ranged from 0% to 3% and 0% to 9%, respectively, across grades, panels, and rounds of review.

For mathematics, the percentage of items and testlets rated as ***accept*** ranged from 66% to 92% and 61% to 89%, respectively. The percentage of items and testlets rated as ***revise*** ranged from 8% to 33% and 11% to 38%, respectively. The rate at which both items and testlets were recommended for rejection ranged from 0% to 0.01% across grades, panels, and rounds of review.

3.2.3. *Test Development Decisions*

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

DLM test development teams documented the decision category applied by external reviewers to each item and testlet. Following this process, test development teams made a final decision to accept, revise, or reject each of the items and testlets. The ELA test development team retained 98% of items and testlets sent out for external review. Of the items and testlets that were revised, most required only minor changes (e.g., minor rewording but concept remained unchanged), as opposed to major changes (e.g., stem or option replaced). The ELA team made 46 minor revisions to items and 5 minor revisions to testlets. The mathematics test development team retained 100% of items and testlets sent out for external review. As with ELA, most revisions made to items and testlets were minor. The mathematics team made 143 minor revisions to items and 9 minor revisions to testlets.

3.3. Operational Assessment Items for Spring 2019

A total of 1,149,460 operational test sessions were administered during the spring testing window. One test session is one testlet taken by one student. Only test sessions that were complete at the close of each testing window counted toward the total sessions.

Testlets were made available for operational testing in 2018–2019 based on the 2017–2018 operational pool and the promotion of testlets field-tested during 2017–2018 to the operational pool following their review. Table 3.9 summarizes the total number of operational testlets for 2018–2019 for ELA and mathematics. There were 746 operational testlets available across grades and subjects. This total included 87 (0 mathematics, 87 ELA) EE/linkage level combinations for which both a general version and a version for students who are blind or visually impaired or read braille were available.

Table 3.9. 2019 Operational Testlets, by Subject (N = 746)

Grade	English language arts (*n*)	Mathematics (*n*)
3	53	37
4	53	44
5	49	41
6	41	41
7	39	38
8	31	41
9	36	47
10	38	43
11	30	44

Similar to prior years, the proportion correct (p-value) was calculated for all operational items to summarize information about item difficulty.

Figure 3.1 and Figure 3.2 include the p-values for each operational item for ELA and mathematics, respectively. To prevent items with small sample sizes from potentially skewing the results, the sample size cutoff for inclusion in the p-value plots was 20. In general, ELA items were easier than mathematics items, as evidenced by the presence of more items in the higher bin (*p*-value) ranges.

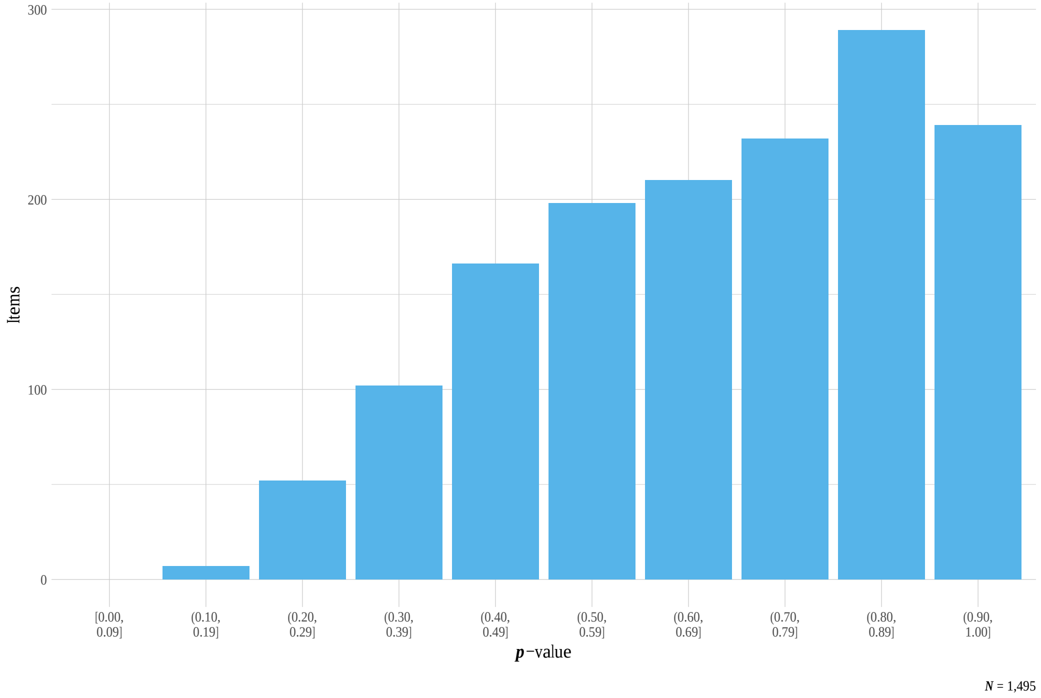


Figure 3.1. *p*-values for ELA 2019 operational items. *Note*. Items with a sample size of less than 20 were omitted.

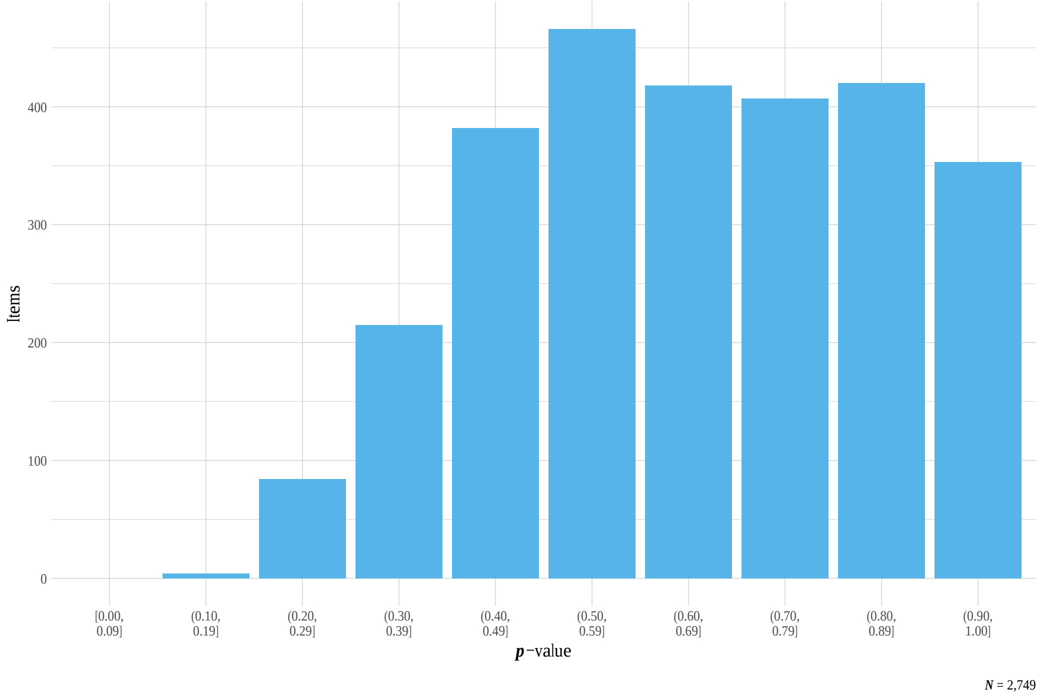


Figure 3.2. *p*-values for mathematics 2019 operational items. *Note*. Items with a sample size of less than 20 were omitted.

Standardized difference values were also calculated for all operational items, with a student sample size of at least 20 to required to compare the *p*-value for the item to all other items measuring the same EE and linkage level. The standardized difference values provide one source of evidence of internal consistency. See Chapter 9 in this manual for additional information.

Figure 3.3 and Figure 3.4 summarize the standardized difference values for operational items for ELA and mathematics, respectively. Most items fell within two standard deviations of the mean of all items measuring the EE and linkage level. As additional data are collected and decisions are made regarding item pool replenishment, test development teams will consider item standardized difference values, along with item misfit analyses when determining which items and testlets are recommended for retirement.

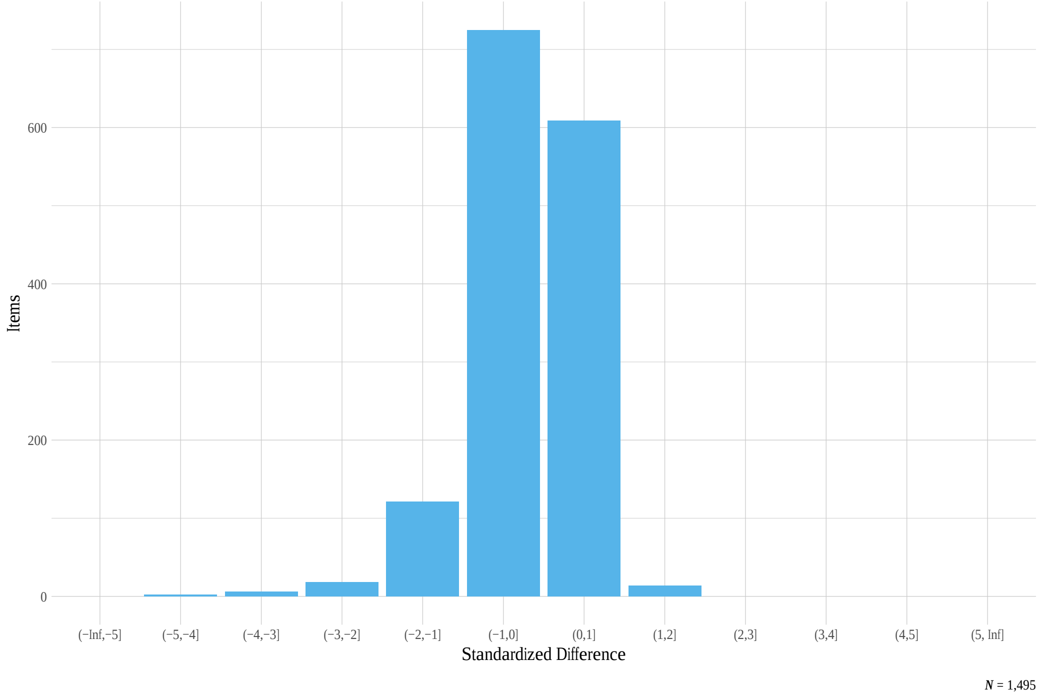


Figure 3.3. Standardized difference *z*-scores for ELA 2019 operational items. *Note.* Items with a sample size of less than 20 were omitted.

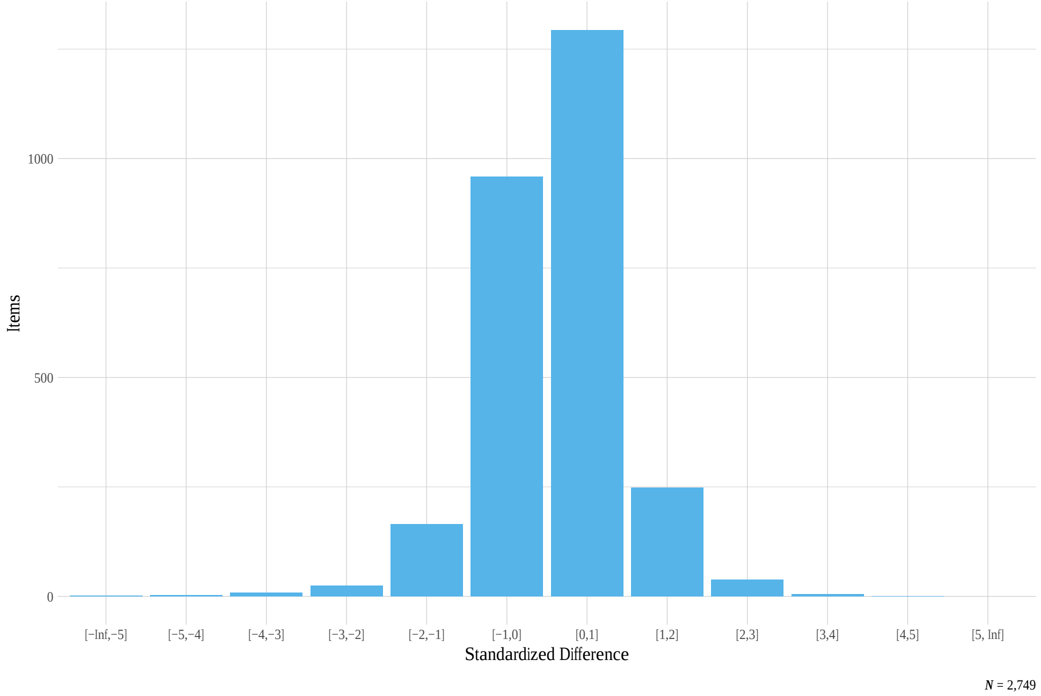


Figure 3.4. Standardized difference *z*-scores for mathematics 2019 operational items. *Note*. Items with a sample size of less than 20 were omitted.

Figure 3.5 summarizes the standardized difference values for operational items for both ELA and mathematics by linkage level. Most items fell within two standard deviations of the mean of all items measuring the respective EE and linkage level, and the distributions are consistent across linkage levels.

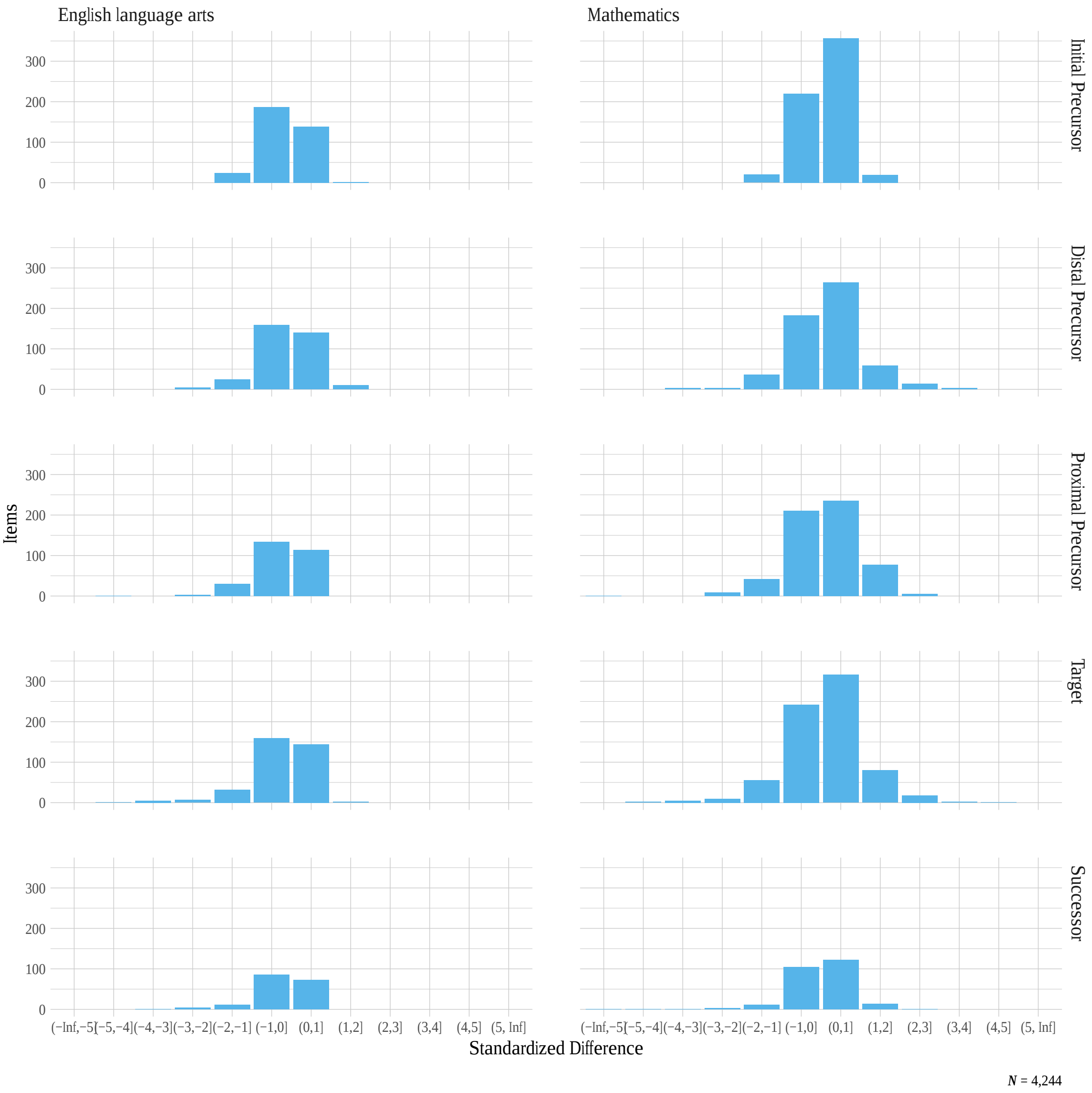


Figure 3.5. Standardized difference *z*-scores for ELA and mathematics 2018–2019 operational items by linkage level. *Note*. Items with a sample size of less than 20 were omitted.

3.4. Field Testing

During the spring 2019 administration, DLM field tests were administered to collect student data on linkage levels adjacent to those taken during the operational assessment. By collecting this data, we are better able to empirically evaluate the relationships between linkage levels.

A 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, et al., 2016), and in Chapter 3 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016) and subsequent annual DLM technical manual updates.

3.4.1. Description of Field Tests

Field test testlets were administered during the spring window. Students received a field test testlet for each subject upon completion of all operational testlets.

The spring field test administration was designed to ensure collection of data for each participating student at more than one linkage level for an EE to support future modeling development (see Chapter 5 of this manual). As such, the field test testlet for each subject was assigned at one linkage level above or below the linkage level that was assessed for the given EE during the spring assessment. In order to reduce the amount of missing data to further support modeling development, all spring field test content came from the existing single-EE testlet spring operational pool.

One ELA and two mathematics EEs were selected for field test from each grade (3–11 in ELA and mathematics). In the single-EE operational pool from which the field test content was drawn, ELA EEs are banded in grades 9 and 10. Therefore, one EE was selected from the grade band, which was administered to both grade 9 and grade 10 students in ELA. This resulted in a total of 26 EEs being selected for the field test. Although two mathematics EEs were selected for field testing, both EEs were administered on a single form. Table 3.10 shows the number of field test testlets that were available for each grade and subject. There were five testlets available for each grade, corresponding with the five linkage levels of the selected EEs for each grade and subject. Because there were two mathematics EEs selected in each grade, there were two testlets for each linkage level, corresponding to the two EEs.

Table 3.10. Spring 2019 Field Test Testlets Available

Grade	English language arts	Mathematics
3	5	10
4	5	10
5	5	10
6	5	10
7	5	10
8	5	10
9	5	10
10	—	10
11	5	10

Note. ELA is grade banded in grades 9–10.

Participation in spring field testing was not required, but teachers were encouraged to administer all available testlets to their students. Participation rates for ELA and mathematics in spring 2019 are shown in Table 3.11. In total, 74% of students in ELA and 73% of students in mathematics took at least one field test form. High participation rates allowed for a significant increase in the amount of cross-linkage level data, furthering modeling research into the structure of the EEs (see Chapter 5 of this manual for future directions). Because the purpose of the spring field test was to collect additional cross-linkage-level data and used currently available operational testlets, test development team review of items included in the field test was not necessary.

Table 3.11. Students Who Completed a Field Test Testlet, by Subject

Subject	*n*	%
English language arts	53,870	74.1
Mathematics	52,686	72.6

3.5. Conclusion

During the 2018–2019 academic year, the test development teams conducted events for both item writing and external review. Overall, over 400 testlets were written for ELA and mathematics. Additionally, during external review, 98% of ELA testlets and 100% of mathematics testlets were retained with no or minor changes. Of the content already in the operational pool, most items had a *p*-values within two standard deviations of the average for the the EE and linkage level. Field testing in 2018–2019 focused on collecting data from students at linkage levels adjacent to those administered during the operational assessment to support future modeling work. Field testing in 2019–2020 will be focused on collecting data for the content that was retained during the external review event described in this chapter.

4. Test Administration

Chapter 4 of the Dynamic Learning Maps® (DLM®) Alternate Assessment System *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016) describes general test administration and monitoring procedures. This chapter describes updated procedures and data collected in 2018–2019, including a summary of administration time, writing testlet assignment, adaptive routing, 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 instructionally embedded testlets, spring operational administration of testlets, the use of adaptive delivery during the spring window, and the availability of accessibility supports.

For a complete description of test administration for DLM assessments, including information on available resources and materials and information on monitoring assessment administration, see the *2014–15 Technical Manual—Year-End Model* (DLM Consortium, 2016).

4.1. Overview of Key Administration Features

This section describes the testing windows for DLM test administration for 2018–2019. For a complete description of key administration features, including information on assessment delivery, Kite Student Portal, and linkage level selection, see Chapter 4 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016). Additional information about administration can also be found in the *Test Administration Manual 2018–2019* (DLM Consortium, 2018c) and the *Educator Portal User Guide* (DLM Consortium, 2018b).

4.1.1. Test Windows

Instructionally embedded assessments were available for teachers to optionally administer between September 19 and December 19, 2018, and between January 2 and February 27, 2019. During the consortium-wide spring testing window, which occurred between March 11 and June 7, 2019, students were assessed on each Essential Element (EE) on the blueprint. Each state sets its own testing window within the larger consortium spring window.

4.2. Administration Evidence

This section describes evidence collected during the spring 2019 operational administration of the DLM alternate assessment. The categories of evidence include data relating to administration time, assignment of writing testlets, the adaptive delivery of testlets in the spring window, user experience, and accessibility.

4.2.1. Administration Time

Estimated administration time varies by student and subject. During the spring testing window, the estimated total testing time is 60–75 minutes per student in English language arts and 35–50 minutes in mathematics.

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

Kite Student Portal captured start and end dates and time stamps for every testlet. To calculate the actual testing time per testlet, the difference between these start and end times was calculated for the spring 2019 operational administration. Table 4.1 summarizes the distribution of test times per testlet. Most testlets took around 5 minutes or less to complete, with mathematics testlets generally taking less time than English language arts testlets. Testlets time out after 90 minutes.

Table 4.1. Distribution of Response Times per Testlet in Minutes

Grade	Min	Median	Mean	Max	25Q	75Q	IQR
English language arts							
3	0.10	3.85	4.82	88.98	2.48	5.90	3.42
4	0.12	4.07	5.12	88.73	2.65	6.30	3.65
5	0.10	4.10	5.03	89.10	2.63	6.27	3.64
6	0.07	4.18	5.19	89.83	2.73	6.45	3.72
7	0.10	4.12	5.24	88.45	2.72	6.33	3.61
8	0.12	4.53	5.52	87.80	3.07	6.75	3.68
9	0.15	4.42	5.78	89.57	2.95	6.83	3.88
10	0.17	4.50	5.62	83.75	3.07	6.80	3.73
11	0.15	5.13	6.69	89.75	3.42	7.88	4.46
12	0.65	6.77	8.07	32.57	4.82	10.01	5.19
Mathematics							
3	0.07	1.95	3.16	89.25	0.95	3.95	3.00
4	0.08	1.78	2.66	89.28	1.05	3.12	2.07
5	0.05	2.02	3.04	88.68	1.10	3.73	2.63
6	0.07	2.22	2.99	86.33	1.27	3.73	2.46
7	0.08	2.05	3.04	89.78	1.22	3.68	2.46
8	0.07	1.87	2.70	88.13	1.12	3.22	2.10
9	0.07	1.80	2.86	87.18	0.95	3.43	2.48
10	0.05	1.98	2.87	82.52	1.03	3.65	2.62
11	0.05	1.97	2.94	82.62	1.05	3.67	2.62
12	0.27	2.88	4.09	42.38	1.50	5.18	3.68

Note. 25Q = lower quartile; 75Q = upper quartile; IQR = interquartile range.

4.2.2. *Adaptive Delivery*

During the spring 2019 test administration, the ELA and mathematics assessments were adaptive between testlets, following the same routing rules applied in prior years. 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 and skill to the appropriate linkage level content.

- 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, a percentage of items answered correctly was calculated for each group of items measuring the same EE. The minimum of these values was then used to determine the next linkage level, based on the above thresholds.

The linkage level of the first testlet assigned to a student was based on First Contact survey responses. The correspondence between the First Contact complexity bands and first assigned linkage levels are shown in Table 4.2.

Table 4.2. Correspondence of Complexity Bands and Linkage Level

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 4 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016).

Following the spring 2019 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, subject, and complexity band. The aggregated results can be seen in Table 4.3 and Table 4.4 for ELA and mathematics, respectively.

Overall, results were similar to those found in the previous years. For the majority of students across all grades who were assigned to the Foundational Complexity Band by the First Contact survey, testlets did not adapt to a higher linkage level after the first assigned testlet (ranging from 65.8% to 93.3% across both subjects). Consistent patterns were not as apparent for students who were assigned Complexity Band 1, Complexity Band 2, or Complexity Band 3. Distributions across the three categories were more variable across grades and subjects.

The 2018–2019 results build on earlier findings from the pilot study and the previous years of operational assessment administration (see Chapter 3 and Chapter 4 of the *2014–2015 Technical Manual—Year-End Model*, respectively, as well as Chapter 3 and Chapter 4 of the annual technical manual updates) and suggest that the First Contact survey complexity band assignment is an effective tool for assigning students content at appropriate linkage levels. Results also indicate that linkage levels of students assigned to higher complexity bands are more variable with respect to the direction in which students move between the first and second testlets. Several factors may help explain these results, including more variability in student characteristics within this group and content-based differences across grades and subjects. Further exploration is needed in this area.

Table 4.3. Adaptation of Linkage Levels Between First and Second English Language Arts Testlets (*N* = 72,671)

	Foundational		Band 1			Band 2			Band 3		
Grade	**Adapted Up (%)**	**Did Not Adapt (%)**	**Adapted Up (%)**	**Did Not Adapt (%)**	**Adapted Down (%)**	**Adapted Up (%)**	**Did Not Adapt (%)**	**Adapted Down (%)**	**Adapted Up (%)**	**Did Not Adapt (%)**	**Adapted Down (%)**
Grade 3	20.2	79.8	33.2	37.2	29.6	69.2	15.5	15.3	87.7	4.7	7.6
Grade 4	34.2	65.8	18.5	42.8	38.7	34.0	41.9	24.1	53.0	44.3	2.8
Grade 5	22.4	77.6	25.3	30.2	44.5	60.5	27.3	12.2	66.8	31.2	2.0
Grade 6	13.1	86.9	23.4	10.0	66.5	39.8	22.9	37.3	35.1	22.3	42.6
Grade 7	15.7	84.3	20.6	30.4	49.0	31.3	36.4	32.3	42.1	30.1	27.8
Grade 8	33.4	66.6	31.0	40.6	28.4	50.2	39.8	10.0	86.9	10.2	2.9
Grade 9	10.9	89.1	16.5	9.8	73.7	28.1	14.2	57.7	43.2	10.3	46.5
Grade 10	6.7	93.3	10.6	37.5	51.9	24.6	46.5	28.9	45.8	45.8	8.4
Grade 11	10.2	89.8	4.2	25.7	70.0	24.8	41.1	34.1	38.3	44.9	16.9

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

Table 4.4. Adaptation of Linkage Levels Between First and Second Mathematics Testlets (N = 72,618)

	Foundational		Band 1			Band 2			Band 3		
Grade	**Adapted Up (%)**	**Did Not Adapt (%)**	**Adapted Up (%)**	**Did Not Adapt (%)**	**Adapted Down (%)**	**Adapted Up (%)**	**Did Not Adapt (%)**	**Adapted Down (%)**	**Adapted Up (%)**	**Did Not Adapt (%)**	**Adapted Down (%)**
Grade 3	6.9	93.1	6.5	29.6	63.9	13.9	26.1	59.9	5.6	53.4	40.9
Grade 4	14.6	85.4	51.3	12.5	36.2	62.1	18.0	19.9	46.3	26.7	27.0
Grade 5	22.6	77.4	11.7	17.4	70.9	16.6	9.1	74.3	55.6	8.4	36.0
Grade 6	12.2	87.8	12.9	25.8	61.3	17.2	34.0	48.9	44.1	27.3	28.6
Grade 7	11.1	88.9	8.9	17.3	73.8	34.1	33.2	32.7	37.6	8.9	53.5
Grade 8	16.1	83.9	14.6	6.0	79.3	3.8	11.7	84.6	12.6	17.7	69.6
Grade 9	13.2	86.8	7.2	29.6	63.3	5.3	39.4	55.3	19.5	49.3	31.2
Grade 10	6.9	93.1	0.6	33.2	66.2	1.7	18.7	79.6	17.0	49.6	33.4
Grade 11	7.7	92.3	2.7	24.3	73.1	2.4	25.0	72.6	8.7	54.7	36.7

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

4.2.3. *Writing Testlet Assignment*

Student assignment to emergent and conventional writing testlets was adjusted for the spring 2019 administration to improve the match between student writing skills and complexity of the writing testlet. For a complete description of the two types of writing testlets, please see Chapter 3 of the *2016–2017 Technical Manual Update—Year-End Model* (DLM Consortium, 2017b).

Prior to the spring 2019 assessment administration, each student's spring writing testlet level was assigned via adaptive routing[3] based on performance on the preceding English language arts (ELA) reading testlet. Beginning in spring 2019, teacher responses to a First Contact Survey[4] item about students' writing skills were used to assign students to a writing testlet. The seven-option, multiple-choice item asked teachers to indicate the answer that most closely matched the student's highest level of writing skill, with responses ranging from "Scribbles or randomly writes/selects letters or symbols" to "Writes paragraph-length text without copying using spelling (with or without word prediction)." Teachers most frequently responded that the student "scribbles or randomly writes/selects letters or symbols" (27%), followed by "writes by copying words or letters" (24%), and "writes words or simple phrases without copying using spelling" (17%). The full results are summarized in Table 4.5.

Table 4.5. Responses to Writing First Contact Item

Statement	***n***	**%**
Scribbles or randomly writes/selects letters or symbols	20,434	26.8
Writes by copying words or letters	18,556	24.4
Writes using word banks or picture symbols	5,455	7.2
Writes words using letters to accurately reflect some of the sounds	7,290	9.6
Writes words or simple phrases without copying using spelling	12,861	16.9
Writes sentences or complete ideas without copying using spelling	9,259	12.2
Writes paragraph-length text without copying using spelling	2,328	3.1

First Contact responses were used to assign the two types of writing testlets based on a review of emergent and conventional writing testlet content and prior student performance data. Students whose teachers indicated they wrote by scribbling, copying or using word bands, or writing words corresponding to some sounds received an emergent-level testlet. Students whose teacher indicated they wrote words or simple phrases, sentences or complete ideas, or paragraph-length text without copying and using spelling received the conventional writing testlet. The number and percentage of students assigned to each level of writing testlet by grade in spring 2019 is summarized in Table 4.6. Overall, 68% of students were assigned to an emergent writing testlet and 32% of students were assigned to a conventional writing testlet.

[3] For a complete description of adaptive routing, please see Chapter 4 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2017b)

[4] For a complete description of the First Contact Survey, please see Chapter 4 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016) and the First Contact census report (Nash et al., 2015)

Table 4.6. Students Assigned to Each Writing Testlet Level by Grade

	Emergent		Conventional	
Grade	*n*	%	*n*	%
Grade 3	7,278	84.5	1,338	15.5
Grade 4	7,436	79.4	1,926	20.6
Grade 5	7,018	75.0	2,341	25.0
Grade 6	6,541	69.6	2,857	30.4
Grade 7	6,085	64.7	3,326	35.3
Grade 8	5,559	59.7	3,749	40.3
Grade 9	3,529	53.6	3,055	46.4
Grade 10	1,818	54.2	1,536	45.8
Grade 11	3,071	51.3	2,918	48.7
Grade 12	11	44.0	14	56.0

4.2.4. *Administration Incidents*

As in all previous operational years, testlet assignment during the spring 2019 assessment window was monitored to ensure students were correctly assigned to testlets. Only two incidents were observed in 2018–2019 that had the potential to impact scoring.

The first incident involved a mathematics testlet that was administered with an incorrectly sized graphic due to a technology glitch. The size of the graphic may have impacted student responses to the item. Upon discovery, the item in question was immediately corrected. However, prior to the correction, 47 students had taken the item. Because the size of the graphic may be impacted their answer selection, the total correct responses on the testlet may have impacted routing to the subsequent testlet. The second incident involved an ELA testlet in which an item was placed in the incorrect order within a text. Upon discovery this testlet was immediately removed from the window and replaced with an alternative testlet. Prior to this switch, 110 students had taken the out-of-order testlet. Because the ordering of the testlet may have impacted student responses, the total correct responses on the testlet may have impacted routing to the subsequent testlet. For both incidents, state partners were given the option to revert students to the end of the testlet completed immediately prior to the the testlet on which the incident occurred and resume testing, or to let students proceed forward as usual.

As in previous years, an Incident File was delivered to state partners with the General Research File (see Chapter 7 of this manual for more information), which provided the list of all students potentially affected by either issue. States were able to use this file during the two-week review period to make decisions about invalidation of records at the student level based on state-specific accountability policies and practices. Assignment to testlets will continue to be monitored in subsequent years to track any potential incidents and report them to state partners.

4.3. Implementation Evidence

This section describes evidence collected during the spring 2019 operational implementation of the DLM alternate assessment. The categories of evidence include survey data relating to user experience and accessibility.

4.3.1. User Experience With the DLM System

User experience with the spring 2019 assessments was evaluated through the spring 2019 survey, which was disseminated to teachers who had administered a DLM assessment during the spring window. In 2019, the survey was distributed to teachers in Kite Student Portal, where students completed assessments. Each student was assigned a survey for their teacher to complete. The survey included three sections. The first and third sections were fixed across all students, while the second section was spiraled across students, with teachers responding to a block of questions pertaining to accessibility, Educator Portal and Kite Student Portal, the relationship of assessment content to instruction by subject, and score reports.

A total of 12,613 teachers in year-end model states responded to the survey (with a response rate of 77%) for 42,106 students.

Participating teachers responded to surveys for a median of two students. Teachers reported having an average of 11 years of experience in ELA, 11 years in mathematics, and 9 years with students with significant cognitive disabilities. The median response to the number of years of experience in ELA was 10 years, the median experience in mathematics was 9 years, and the median experience with students with significant cognitive disabilities was 7 years. Approximately 26% indicated they had experience administering the DLM assessment in all five operational years.

The following sections summarize user experience with the system and accessibility. Additional survey results are summarized in Chapter 9 (Validity Studies). For responses to the priors years' surveys, see Chapter 4 and Chapter 9 in the respective technical manuals (DLM Consortium, 2017a, 2017b, 2018a).

4.3.1.1. Educator Experience

Survey respondents were asked to reflect on their own experience with the assessments as well as their comfort level and knowledge administering them. Most of the questions required teachers to respond on a four-point scale: *strongly disagree, disagree, agree,* or *strongly agree.* Responses are summarized in Table 4.7.

Nearly all teachers (97%) agreed or strongly agreed that they were confident administering DLM testlets. Most respondents (90%) agreed or strongly agreed that the required test administrator training prepared them for their responsibilities as test administrators. Most teachers also responded that they had access to curriculum aligned with the content that was measured by the assessments (86%) and that they used the manuals and the Educator Resources page (89%).

Table 4.7. Teacher Responses Regarding Test Administration

	SD		D		A		SA		A+SA	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
I was confident in my ability to deliver DLM testlets.	108	0.9	264	2.3	4,419	38.3	6,737	58.4	11,156	96.7
Required test administrator training prepared me for the responsibilities of a test administrator.	298	2.6	884	7.7	5,560	48.4	4,754	41.4	10,314	89.8
I have access to curriculum aligned with the content measured by DLM assessments.	339	2.9	1,211	10.5	5,775	50.2	4,183	36.3	9,958	86.5
I used manuals and/or the DLM Educator Resource Page materials.	255	2.2	1,004	8.7	6,192	53.8	4,062	35.3	10,254	89.1

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

4.3.1.1.1. Kite System

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

The software used for the administration of DLM testlets is Kite Student Portal. Teachers were asked to consider their experiences with Kite Student Portal and respond to each question on a four-point scale: *very hard, somewhat hard, somewhat easy*, or *very easy*. Table 4.8 summarizes teacher responses to these questions.

Respondents found it to be either *somewhat easy* or *very easy* to log in to the system (95%), to navigate within a testlet (96%), to record a response (97%), to submit a completed testlet (97%), and to administer testlets on various devices (93%). Open-ended survey response feedback indicated testlets were easy to administer and that technology had improved compared to previous years.

Table 4.8. Ease of Using Kite Student Portal

	VH		SH		SE		VE		SE+VE	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Enter the site	89	0.8	460	4.3	3,392	32.0	6,658	62.8	10,050	94.8
Navigate within a testlet	73	0.7	372	3.5	3,213	30.4	6,923	65.4	10,136	95.8
Record a response	51	0.5	296	2.8	2,939	27.8	7,278	68.9	10,217	96.7
Submit a completed testlet	51	0.5	251	2.4	2,800	26.6	7,431	70.5	10,231	97.1
Administer testlets on various devices	126	1.2	571	5.4	3,683	35.0	6,157	58.4	9,840	93.4

Note. VH = very hard; SH = somewhat hard; SE = somewhat easy; VE = very easy; SE+VE = somewhat easy and very easy.

Educator Portal is an area of the Kite System used to store and manage student data and enter PNP and First Contact information. To address teachers' feedback from prior administrations, the appearance and functionality of Educator Portal was updated during the summer of 2018. The update focused on the improvement of user experience, accessibility, and a general improvement to the look, feel, and functionality of Educator Portal without causing undue disruption to how educators use the application. Updates made to Educator Portal during the summer of 2018 include: updating the user interface to be more intuitive, have a more logical flow, display auto-populated fields, and restrict users from saving incomplete records; reordering tabs to be more intuitive, rewriting data upload error messages in nontechnical language instead of programming language, and updating the color scheme to be consistent across the application.

Teachers were asked to assess the ease of navigating and using Educator Portal for its intended purposes. The data are summarized in Table 4.9 using the same scale used to rate experiences with Kite Student Portal. Overall, the improvements made to Educator Portal during summer 2018 are reflected in the respondents' favorable feedback. A majority of teachers found it to be either *somewhat easy* or *very easy* to navigate the site (89%), enter PNP and First Contact information (92%), manage student data (89%), manage their accounts (91%), or manage tests (89%). The percentages of respondents responding *somewhat easy* or *very easy* increased from 2017–2018, reflecting the improvements made to the system (DLM Consortium, 2016).

Table 4.9. Ease of Using Educator Portal

	VH		SH		SE		VE		SE+VE	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Navigate the site	161	1.5	1,005	9.5	4,307	40.6	5,147	48.5	9,454	89.1
Enter Access Profile and First Contact information	123	1.2	727	6.9	4,291	40.5	5,450	51.5	9,741	92.0
Manage student data	173	1.6	1,026	9.7	4,546	42.9	4,841	45.7	9,387	88.6
Manage my account	132	1.2	854	8.1	4,559	43.0	5,052	47.7	9,611	90.7
Manage tests	161	1.5	958	9.0	4,332	40.9	5,143	48.5	9,475	89.4

Note. VH = very hard; SH = somewhat hard; SE = somewhat easy; VE = very easy; SE+VE = somewhat easy and very easy.

Finally, respondents were asked to rate their overall experience with Kite Student Portal and Educator Portal on a four-point scale: *poor, fair, good,* and *excellent*. Results are summarized in Table 4.10. The majority of respondents reported a positive experience with Kite Student Portal. A total of 91% of respondents rated their Kite Student Portal experience as *good* or *excellent*, while 84% rated their overall experience with Educator Portal as *good* or *excellent.*

Table 4.10. Overall Experience With Kite Student Portal and Educator Portal

	Poor		Fair		Good		Excellent		Good + Excellent	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Student Portal	168	1.6	829	7.8	4,950	46.6	4,676	44.0	9,626	90.6
Educator Portal	328	3.1	1,320	12.4	5,455	51.3	3,529	33.2	8,984	84.5

Overall, feedback from teachers indicated that Kite Student Portal was easy to navigate and user friendly. Teachers also provided useful feedback about how to continue to improve the Educator Portal user experience, which will be considered for technology development for 2019–2020 and beyond.

4.3.2. *Accessibility*

Accessibility supports provided in 2018–2019 were the same as those available in previous years. The *DLM Accessibility Manual* (DLM Consortium, 2017c), distinguishes accessibility supports that are provided in Kite Student Portal via the Personal Needs and Preferences Profile, require additional tools or materials, or are provided by the test administrator outside the system.

Table 4.11 shows selection rates for the three categories of accessibility supports. The most commonly selected supports were human read aloud, test administrator enters responses for student, and individualized manipulatives. For a complete description of the available accessibility supports, see Chapter 4 in the *2014–15 Technical Manual—Year-End Model* (DLM Consortium, 2016).

Table 4.11. Accessibility Supports Selected for Students (*N* = 71,932)

Support	*n*	%
Supports provided in Kite Student Portal		
Spoken audio	11,529	16.0
Magnification	7,703	10.7
Color contrast	6,017	8.4
Overlay color	3,658	5.1
Invert color choice	2,357	3.3
Supports requiring additional tools/materials		
Individualized manipulatives	34,634	48.1
Calculator	23,058	32.1
Single-switch system	2,332	3.2
Alternate form - visual impairment	1,709	2.4
Two-switch system	1,048	1.5
Uncontracted braille	27	0.0
Supports provided outside the system		
Human read aloud	64,143	89.2
Test administrator enters responses for student	40,725	56.6
Partner assisted scanning	7,050	9.8
Language translation of text	1,410	2.0
Sign interpretation of text	1,157	1.6

Table 4.12 describes teacher responses to survey items about the accessibility supports used during administration. Teachers were asked whether the student was able to effectively use available accessibility supports and whether the accessibility supports were similar to the ones used for instruction. The majority of teachers agreed that students were able to effectively use accessibility supports (94%), while responses to whether the accessibility supports were similar to ones students used for instruction were mixed (60%). While states and districts have differing policies for whether to include accessibility supports on the student's IEP, most (66%) indicated supports were included.

Table 4.12. Teacher Report of Student Accessibility Experience

	Disagree		**Agree**	
Statement	*n*	%	*n*	%
Student was able to effectively use accessibility features.	748	6.4	11,005	93.6
Accessibility features were similar to ones student uses for instruction.	285	40.3	423	59.7

Of the teachers who reported that their student was unable to effectively use the accessibility supports (6%), the most commonly reported reason was that the student could not provide a response even with the support provided (64%).

Table 4.13. Reason Student was Unable to Effectively Use Available Accessibility Supports

Reason	*n*	%
Even with support, the student could not provide a response	432	64.4
The student needed a support that wasn't available or allowed	147	21.9
The student refused the support during testing	130	19.4
The student was unfamiliar with the support	115	17.1
There was a technology problem (e.g., KITE display, AAC device)	42	6.3

Teachers have several allowable options for flexibility while assessing students. Of these options for flexibility, teachers most frequently reported using breaks (62%), reinforcement (42%), or individualized student response mode (33%). Additionally, 37% of teachers reported adapting or substituting materials.

Table 4.14. Options for Flexibility Teachers Reported Utilizing for a Student

Option	*n*	%
Breaks	6919	61.53
Use of reinforcement	4708	41.87
Individualized student response mode	3677	32.70
Blank paper	2433	21.64
None of these	1952	17.36
Navigation across screens	1671	14.86
Alternate representation of answer options	1639	14.58
Generic definitions	1140	10.14
Special equipment for positioning	831	7.39
Display testlet on interactive whiteboard	562	5.00
Graphic organizer	482	4.29

While overall these data support the conclusion that the accessibility supports of the DLM alternate assessment were effectively used by students, additional data will be collected during spring 2020 to determine whether additional improvements can be made to ensure all students can access DLM assessments.

4.4. Conclusion

During the 2018–2019 academic year, the DLM system was available during two testing windows: an optional instructionally embedded window and the spring window. Implementation evidence was collected in the form of teacher survey responses regarding user experience, accessibility, and Personal Needs and Preferences Profile selections. Results from the teacher survey indicated that teachers felt confident administering testlets in the system, that Kite Student Portal was easy to use, and that Educator Portal had improved since the prior year.

5. Modeling

Chapter 5 of the Dynamic Learning Maps® (DLM®) Alternate Assessment System *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016) described the basic psychometric model that underlies the DLM assessment system, while the *2015–2016 Technical Manual Update—Year-End Model* (DLM Consortium, 2017a) provides a complete, detailed description of the process used to estimate item and student parameters from student assessment data. This chapter provides a high-level summary of the model used to calibrate and score assessments, along with a summary of updated modeling evidence from the 2018–2019 administration year.

For a complete description of the psychometric model used to calibrate and score the DLM assessments, including the psychometric background, the structure of the assessment system, suitability for diagnostic modeling, and a detailed summary of the procedures used to calibrate and score DLM assessments, see the *2015–2016 Technical Manual Update—Year-End Model* (DLM Consortium, 2017a).

5.1. Overview of the Psychometric Model

Learning map models, which are networks of sequenced learning targets, are at the core of the DLM assessments in English language arts (ELA) and mathematics. Because of the underlying map structure and the goal of providing 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 latent class analysis, a form of diagnostic classification modeling, to provide information about student mastery of multiple skills measured by the assessment. Results are reported for each alternate content standard, called an Essential Element (EE), at the five levels of complexity for which assessments are available: Initial Precursor, Distal Precursor, Proximal Precursor, Target, and Successor.

Simultaneous calibration of all linkage levels within an EE is not currently possible because of the administration design, in which overlapping data from students taking testlets at multiple levels within an EE is uncommon. Instead, each linkage level was calibrated separately for each EE using separate latent class analyses. Also, because items were developed to meet a precise cognitive specification, all master and non-master 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.

A description of the DLM scoring model for the 2018–2019 administration follows. Using latent class analysis, a probability of mastery was calculated on a scale from 0 to 1 for each linkage level within each EE. Each linkage level within each EE was considered the latent variable to be measured. Students were then classified into one of two classes for each linkage level of each EE: master or non-master. As described in Chapter 6 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016), a posterior probability of at least .8 was required for mastery classification. Consistent with the assumption of item fungibility, a single set of probabilities of masters and non-masters providing a correct response was estimated for all items within a linkage level. Finally, a structural parameter, which is the proportion of masters for the linkage level (i.e., the analogous map parameter), was also estimated. In total, three parameters per linkage level are specified in the DLM scoring model: a fungible probability for non-masters, a fungible probability for masters, and the proportion of masters.

Following calibration, students' 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 had mastered a given linkage level within an EE, then the student was assumed to have mastered all lower levels within that EE.

In addition to the calculated posterior probability of mastery, students could be assigned mastery of linkage levels within an EE in two other ways: correctly answering 80% of all items administered at the linkage level or through the *two-down* scoring rule. The two-down scoring rule was implemented to guard against students assessed at the highest linkage levels being overly penalized for incorrect responses. When a student tested at more than one linkage level for the EE and did not demonstrate mastery at any level, the two-down rule was applied according to the lowest linkage level tested. For more information, see the Mastery Assignment section.

5.2. Calibrated Parameters

As stated in the previous section, the comparable *item parameters* for diagnostic assessments are the conditional probabilities of masters and non-masters providing a correct response to the item. Because of the assumption of fungibility, parameters are calculated for each of the 1,210 linkage levels across ELA and mathematics (5 linkage levels × 242 EEs). Parameters include a conditional probability of non-masters providing a correct response and a conditional probability of masters providing a correct response. Across all linkage levels, the conditional probability that masters will provide a correct response is generally expected to be high, while it is expected to be low for non-masters. In addition to the item parameters, the psychometric model also includes a structural parameter, which defines the base rate of mastery for each linkage level. A summary of the operational parameters used to score the 2018–2019 assessment is provided in the following sections.

5.2.1. Probability of Masters Providing Correct Response

When items measuring each linkage level function as expected, students who have mastered the linkage level have a high probability of providing a correct response to items measuring the linkage level. Using the 2019 operational calibration, Figure 5.1 depicts the conditional probability of masters providing a correct response to items measuring each of the 1,210 linkage levels. Because the point of maximum uncertainty is .5, masters should have a greater than 50% chance of providing a correct response. The results in Figure 5.1 demonstrate that most linkage levels (n = 1,192, 99%) performed as expected. Additionally, 96% of linkage levels (n = 1,161) had a conditional probability of masters providing a correct response over .6. Only a few linkage levels (n = 4, <1%) had a conditional probability of masters providing a correct response less than .4. Thus, a large majority of linkage levels performed consistent with expectations for masters of the linkage levels.

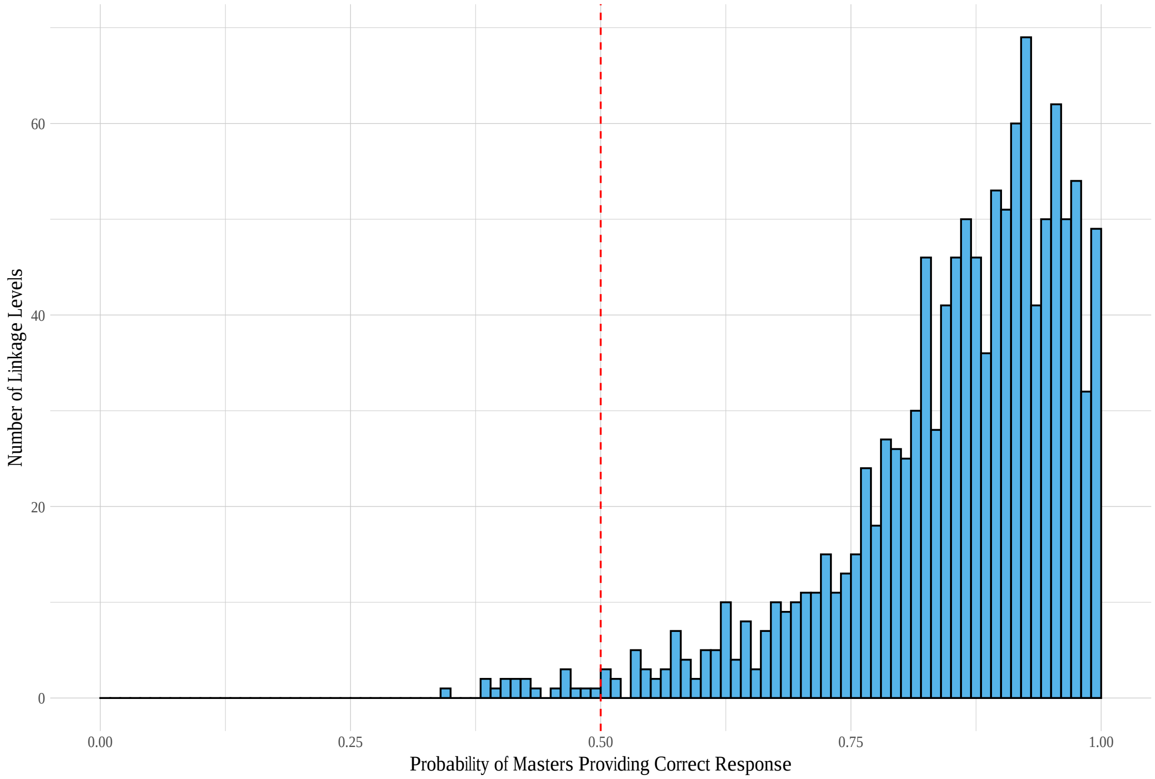


Figure 5.1. Probability of masters providing a correct response to items measuring each linkage level. *Note*. Histogram bins are shown in increments of .01. Reference line indicates .5.

5.2.2. Probability of Non-Masters Providing Correct Response

When items measuring each linkage level function as expected, non-masters of the linkage level have a low probability of providing a correct response to items measuring the linkage level. Instances where non-masters have a high probability of providing correct responses may indicate that the linkage level does not measure what it is intended to measure, or that the correct answers to items measuring the level are easily guessed. These instances may result in students who have not mastered the content providing correct responses and being incorrectly classified as masters. This outcome has implications for the validity of inferences that can be made from results and for teachers using results to inform instructional planning, monitoring, and adjustment.

Figure 5.2 summarizes the probability of non-masters providing correct responses to items measuring each of the 1,210 linkage levels. There is greater variation in the probability of non-masters providing a correct response to items measuring each linkage level than was observed for masters, as shown in Figure 5.2. While most linkage levels (*n* = 897, 74%) performed as expected, non-masters sometimes had a greater than chance (> .5) likelihood of providing a correct response to items measuring the linkage level. Although most linkage levels (*n* = 673, 56%) have a conditional probability of non-masters providing a correct response less than .4, 131 (11%) have a conditional probability for non-masters providing a correct response greater than .6, indicating there are many linkage levels non-masters are more likely than not to provide a correct response. This may indicate the items (and linkage level as a whole, since the item parameters are shared) were easily guessable or did not discriminate well between the two groups of students.

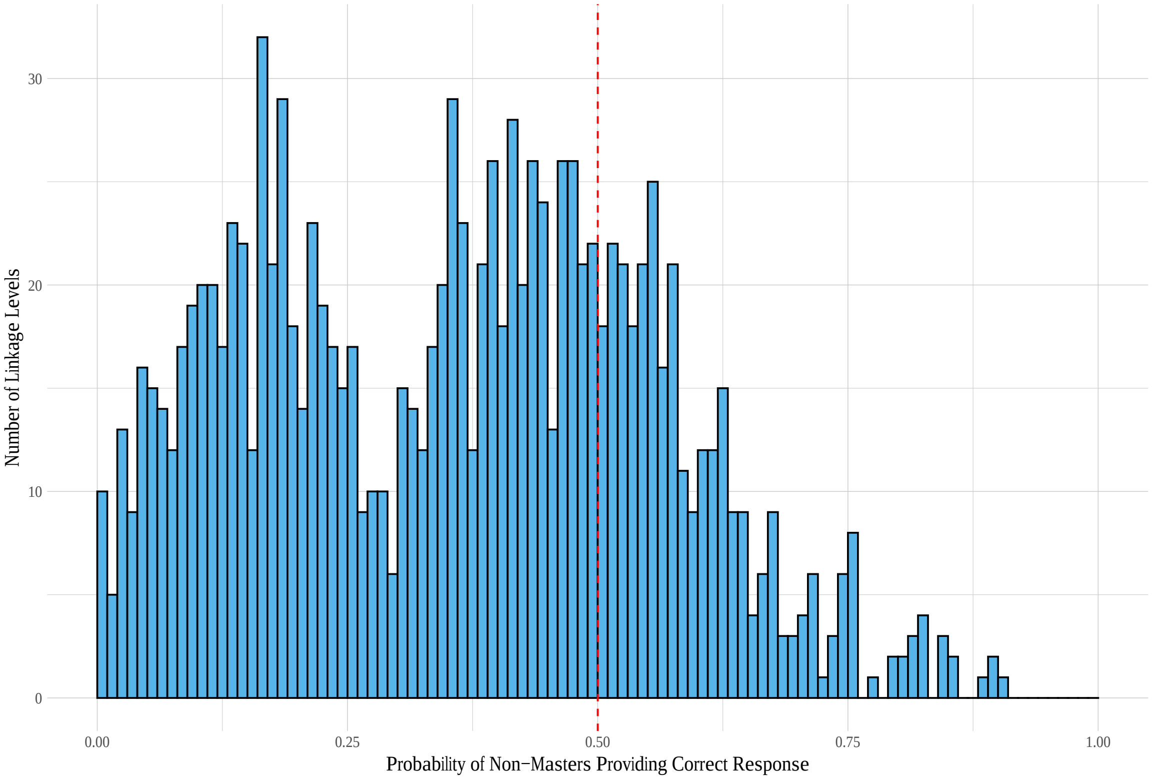


Figure 5.2. Probability of non-masters providing a correct response to items measuring each linkage level. *Note*. Histogram bins are in increments of .01. Reference line indicates .5.

5.2.3. Item Discrimination

The discrimination of a linkage level represents how well the items are able to differentiate masters and non-masters. For diagnostic models, this is assessed by comparing the conditional probabilities of masters and non-masters providing a correct response. Linkage levels that are highly discriminating will have a large difference between the conditional probabilities, with a maximum value of 1.0 (i.e., masters have a 100% chance of providing a correct response and non-masters a 0% chance). Figure 5.3 shows the distribution of linkage level discrimination values. Overall, 72% of linkage levels (n = 868) have a discrimination greater than .4, indicating a large difference between the conditional probabilities (e.g., .75 to .35, .9 to .5, etc.). However, there were 34 linkage levels (3%) with a discrimination of less than .1, indicating that masters and non-masters tend to perform similarly on items measuring these linkage levels.

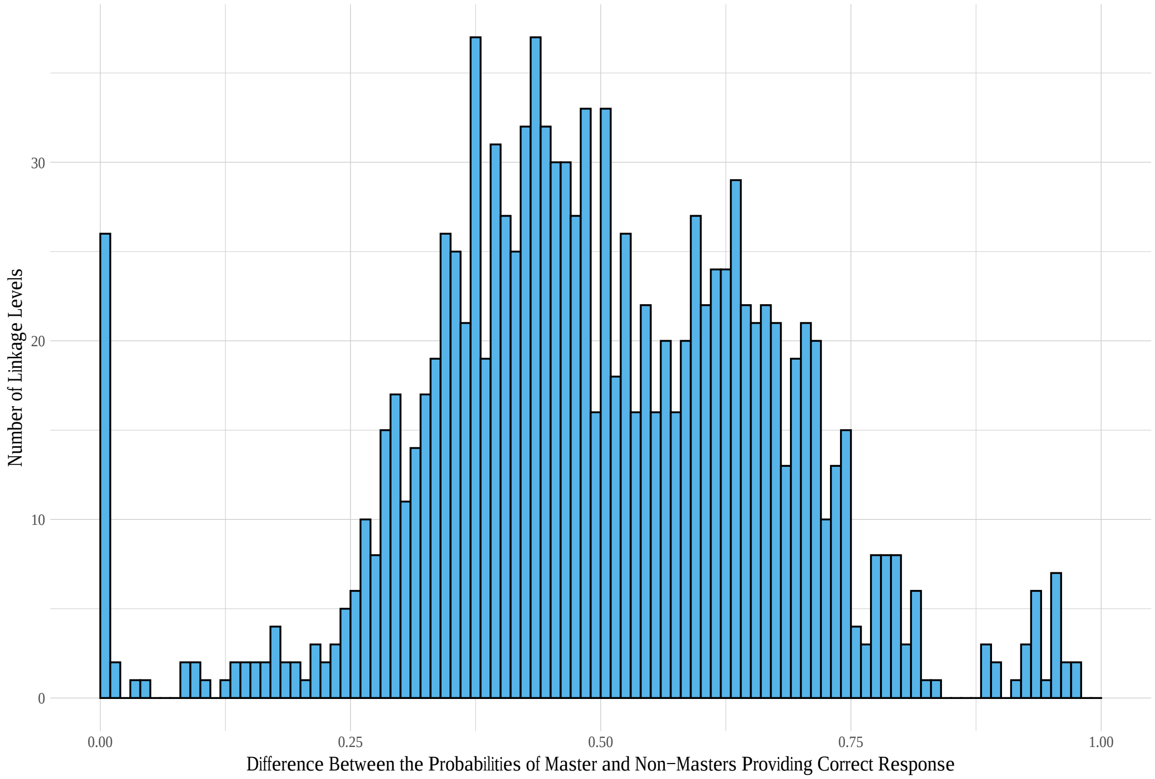


Figure 5.3. Difference between masters' and non-masters' probability of providing a correct response to items measuring each linkage level. *Note*. Histogram bins are in increments of .01. Reference line indicates .5.

5.2.4. Base Rate Probability of Mastery

The DLM assessments are designed to maximize the match of student knowledge and skill to the appropriate linkage level content. The base rate of mastery represents the estimated proportion of masters among students assessed on an EE and linkage level. A base rate of mastery close to .5 indicates that students assessed on a given linkage level are equally likely to be a master or non-master. Conversely a high base rate of mastery would indicate that nearly all students testing on a linkage level are classified as masters. Figure 5.4 depicts the distribution of the base rate of mastery probabilities. Overall, 69% of linkage levels (n = 830) had a base rate of mastery between .25 and .75. This indicates that most linkage levels are performing as expected. On the edges of the distribution, 87 linkage levels (7%) had a base rate of mastery less than .25, and 293 linkage levels (24%) had a base rate of mastery higher than .75. This indicates that students are more likely be assessed on linkage levels they have mastered than those they have not mastered.

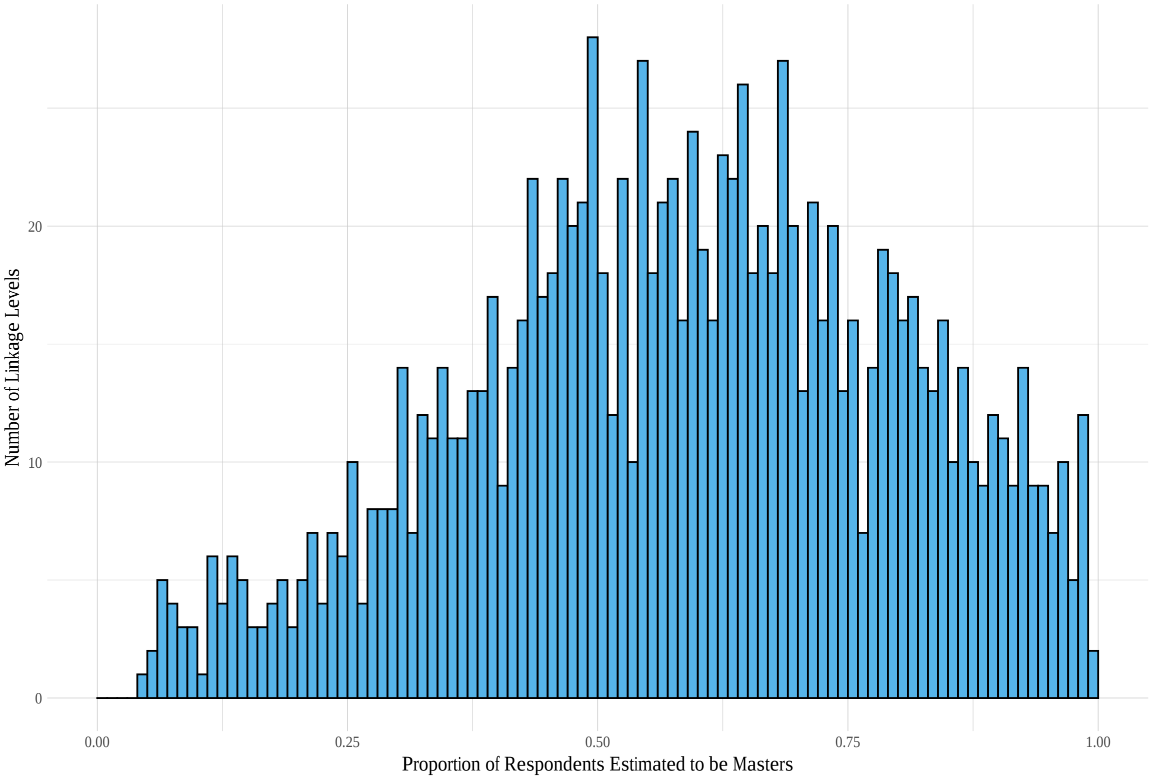


Figure 5.4. Base rate of linkage level mastery. *Note.* Histogram bins are shown in increments of .01.

5.3. Mastery Assignment

As mentioned, in addition to the calculated posterior probability of mastery, students could be assigned mastery of each linkage level within an EE in two additional ways: by correctly answering 80% of all items administered at the linkage level correctly or by the two-down scoring rule.

The two-down scoring rule is designed to avoid excessively penalizing students who do not show mastery of their tested linkage levels. This rule is used to assign mastery to untested linkage levels. Take, for example, a student who tested only on the Target linkage level of an EE. If the student demonstrated mastery of the Target linkage level, as defined by the .8 posterior probability of mastery cutoff or the 80% correct rule, then all linkage levels below and including the Target level would be categorized as mastered. If the student did not demonstrate mastery on the tested Target linkage level, then mastery would be assigned at two linkage levels below the tested linkage level (i.e., the Distal Precursor), rather than showing no evidence of mastery at all. When a student tested on multiple linkage levels and did not show mastery on any tested linkage level, the two-down rule was applied to the lowest tested linkage level. Theoretical evidence for the use of two-down rule is presented in Chapter 2 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016).

To evaluate the degree to which each mastery assignment rule contributed to students' linkage level mastery status during the 2018–2019 administration of DLM assessments, the percentage of mastery statuses obtained by each scoring rule was calculated, as shown in Figure 5.5. Posterior probability was given first priority. That is, if multiple scoring rules agreed on the highest linkage level mastered within an EE (e.g., the posterior probability and 80% correct both indicate the Target linkage level as the highest mastered), the mastery status was counted as obtained via the posterior probability. If mastery was not demonstrated by meeting the posterior probability threshold, the 80% scoring rule was imposed, followed by the two-down rule. Approximately 67% to 78% of mastered linkage levels were derived from the posterior probability obtained from the modeling procedure. Approximately 2% to 18% of linkage levels were assigned mastery status by the percentage correct rule. The remaining approximately 10% to 29% of mastered linkage levels were determined by the minimum mastery, or two-down rule.

Because correct responses to all items measuring the linkage level are often necessary to achieve a posterior probability above the .8 threshold, the percentage correct rule overlapped considerably (but was second in priority) with the posterior probabilities. The percentage correct rule did, however, provide mastery status in those instances where correctly responding to all or most items still resulted in a posterior probability below the mastery threshold. The agreement between these two methods was quantified by examining the rate of agreement between the highest linkage level mastered for each EE for each student. For the 2018–2019 operational year, the rate of agreement between the two methods was 84%. However, in instances where the two methods disagreed, the posterior probability method indicated a higher level of mastery (and was therefore was implemented for scoring) in 21% of cases. Thus, in some instances the posterior probabilities allowed students to demonstrate mastery when the percentage correct was lower than 80% (e.g., a student completed a four-item testlet and answered three of four items correctly).

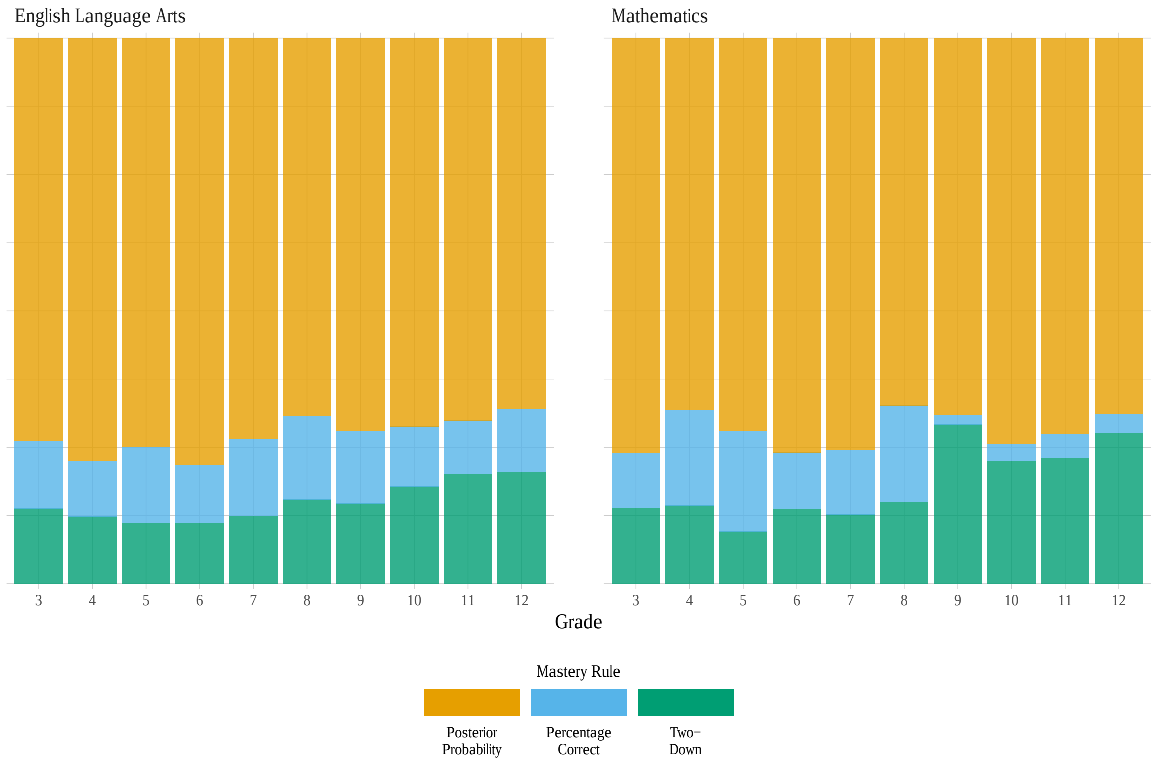


Figure 5.5. Linkage level mastery assignment by mastery rule for each subject and grade.

5.4. Model Fit

Model fit has important implications for the validity of inferences that can be made from assessment results. If the model used to calibrate and score the assessment does not fit the data well, results from the assessment may not accurately reflect what students know and can do. Relative and absolute model fit were compared following the 2017 administration. Model fit research was also prioritized during the 2017–2018 and 2018–2019 operational years, and frequent feedback was provided by the DLM Technical Advisory Committee (TAC) modeling subcommittee, a subgroup of TAC members focused on reviewing modeling-specific research. During the 2018–2019 year, the modeling subcommittee reviewed research related to Bayesian methods for assessing model and item-level fit using posterior predictive model checks (Gelman & Hill, 2006; Gelman et al., 1996), the effect of partial equivalency constraints on model parameters, and new methods for model comparisons (e.g., Vehtari et al., 2017).

For a complete description of the methods and process used to evaluate model fit, see Chapter 5 of the *2016–2017 Technical Manual Update—Year-End Model* (DLM Consortium, 2017b).

5.5. Conclusion

In summary, the DLM modeling approach uses well-established research in Bayesian inference networks and diagnostic classification modeling to determine student mastery of skills measured by the assessment. 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 masters and non-masters, owing 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 non-masters. To ensure students are not excessively penalized by the modeling approach, in addition to posterior probabilities of mastery obtained from the model, two additional scoring procedures are implemented: percentage correct at the linkage level and a two-down scoring rule. Analysis of the scoring rules indicates most students demonstrate mastery of the linkage level based on the posterior probability values obtained from the modeling results.

6. Standard Setting

The standard setting process for the Dynamic Learning Maps® (DLM®) Alternate Assessment System in English language arts (ELA) and mathematics derived cut points for assigning students to four performance levels based on results from the 2014–2015 DLM alternate assessments. For a description of the process, including the development of policy performance level descriptors, the 4-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 6 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016).

7. Assessment Results

Chapter 7 of the Dynamic Learning Maps® (DLM®) Alternate Assessment System *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 2018–2019 student participation data; the percentage of students achieving at each performance level; and subgroup performance by gender, race, ethnicity, and English learner (EL) status. This chapter also reports the distribution of students by the highest linkage level mastered during spring 2019. Finally, this chapter describes updates made to score reports and data files during spring 2019. For a complete description of score reports and interpretive guides, see Chapter 7 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016).

7.1. Student Participation

During spring 2019, assessments were administered to 72,840 students in 12 states and 1 Bureau of Indian Education (BIE) school adopting the year-end model. Counts of students tested in each state and BIE are displayed in Table 7.1. The assessments were administered by 18,852 educators in 9,813 schools and 3,143 school districts.

Table 7.1. Student Participation by State (*N* = 72,840)

State	Students (*n*)
Alaska	476
Colorado	5,071
Delaware	1,071
Illinois	14,676
Miccosukee Indian School	8
New Hampshire	834
New Jersey	11,485
New York	21,360
Oklahoma	5,863
Rhode Island	940
Utah	3,944
West Virginia	1,635
Wisconsin	5,477

Table 7.2 summarizes the number of students assessed in each grade. In grades 3 through 8, over 8,700 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 high school grade-level participation can be traced to differing state-level policies about the grade(s) in which students are assessed.

Table 7.2. Student Participation by Grade (*N* = 72,840)

Grade	Students (*n*)
3	8,786
4	9,529
5	9,542
6	9,573
7	9,576
8	9,449
9	6,747
10	3,460
11	6,152
12	26

Table 7.3 summarizes the demographic characteristics of the students who participated in the spring 2019 administration. The majority of participants were male (67%) and white (59%). About 6% of students were monitored or eligible for EL services.

Table 7.3. Demographic Characteristics of Participants (*N* = 72,840)

Subgroup	*n*	%
Gender		
Male	48,739	66.9
Female	24,101	33.1
Race		
White	42,680	58.6
African American	14,730	20.2
Two or more races	8,930	12.3
Asian	3,780	5.2
American Indian	2,203	3.0
Native Hawaiian or Pacific Islander	377	0.5
Alaska Native	140	0.2
Hispanic ethnicity		
No	55,531	76.2
Yes	17,309	23.8
English learner (EL) participation		
Not EL eligible or monitored	68,629	94.2
EL eligible or monitored	4,211	5.8

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 7.4 summarizes the number of students participating in instructionally embedded testing by state. A total of 503 students took at least one instructionally embedded testlet during the 2018–2019 academic year.

Table 7.4. Students Completing Instructionally Embedded Testlets by State (*N* = 503)

State	*n*
Colorado	76
Delaware	32
Illinois	19
New York	128
Oklahoma	233
Utah	7
West Virginia	8

Table 7.5 summarizes the number of instructionally embedded test sessions taken in ELA and mathematics. Across all states, students took 2,036 ELA testlets and 2,325 mathematics testlets.

Table 7.5. Number of Instructionally Embedded Test Sessions, by Grade

Grade	English language arts	Mathematics
3	259	307
4	355	404
5	321	333
6	224	279
7	314	347
8	365	436
9	59	60
10	19	13
11	120	146
Total	*2,036*	*2,325*

7.2. Student Performance

Student performance on DLM assessments is interpreted using cut points, determined during standard setting, which separate student results into four performance levels. For a full description of the standard-setting process, see Chapter 6 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016). A student's performance level is determined based on the total number of linkage levels mastered across the assessed Essential Elements (EEs).

For the spring 2019 administration, student performance was reported using the same four performance levels approved by the DLM Consortium for prior years:

- 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.

7.2.1. *Overall Performance*

Table 7.6 reports the percentage of students achieving at each performance level from the spring 2019 administration for English language arts (ELA) and mathematics. For ELA, the percentage of students who achieved at the At Target or Advanced levels ranged from approximately 22% to 40%. In mathematics, the percentage of students meeting or exceeding Target expectations ranged from approximately 7% to 31%.

Table 7.6. Percentage of Students by Grade and Performance Level

Grade	Emerging (%)	Approaching (%)	Target (%)	Advanced (%)	Target+ Advanced (%)
English language arts					
3 (*n* = 8,777)	61.4	16.9	20.2	1.4	21.6
4 (*n* = 9,518)	52.2	24.3	20.7	2.8	23.5
5 (*n* = 9,525)	50.8	21.4	24.1	3.7	27.8
6 (*n* = 9,559)	50.4	25.9	16.0	7.7	23.7
7 (*n* = 9,555)	36.4	28.8	25.8	9.0	34.8
8 (*n* = 9,432)	38.0	28.0	25.8	8.1	33.9
9 (*n* = 6,739)	37.6	32.8	21.9	7.7	29.6
10 (*n* = 3,454)	36.3	33.7	25.7	4.2	29.9
11 (*n* = 6,112)	35.5	34.4	26.4	3.7	30.1
12 (*n* = 25)	24.0	36.0	36.0	4.0	40.0
Mathematics					
3 (*n* = 8,763)	60.2	15.2	17.6	7.1	24.6
4 (*n* = 9,503)	53.7	15.2	22.0	9.0	31.0
5 (*n* = 9,526)	60.8	19.7	9.8	9.6	19.5
6 (*n* = 9,551)	58.1	25.1	9.8	7.0	16.9
7 (*n* = 9,544)	64.2	24.2	7.2	4.4	11.6
8 (*n* = 9,426)	56.4	31.1	9.9	2.7	12.6
9 (*n* = 6,736)	52.6	30.6	13.6	3.2	16.8
10 (*n* = 3,451)	64.6	27.6	7.5	0.3	7.9
11 (*n* = 6,117)	60.1	32.7	6.9	0.3	7.2
12 (*n* = 26)	53.8	34.6	11.5	0.0	11.5

7.2.2. *Subgroup Performance*

Data collection for DLM assessments includes demographic data on gender, race, ethnicity, and EL status. Table 7.7 and Table 7.8 summarize the disaggregated frequency distributions for ELA and mathematics, respectively, collapsed across all assessed grade levels. Although states each have their own rules for minimum student counts needed to support public reporting of results, small counts are not suppressed here because results are aggregated across states, and individual students cannot be identified.

Table 7.7. ELA Performance Level Distributions, by Demographic Subgroup (*N* = 72,696)

	Emerging		**Approaching**		**Target**		**Advanced**	
Subgroup	*n*	%	*n*	%	*n*	%	*n*	%
Gender								
Male	22,096	45.4	12,793	26.3	11,102	22.8	2,661	5.5
Female	10,952	45.5	6,400	26.6	5,354	22.3	1,338	5.6
Race								
White	19,273	45.3	10,996	25.8	9,833	23.1	2,484	5.8
African American	6,231	42.4	4,188	28.5	3,466	23.6	819	5.6
Two or more races	4,399	49.3	2,405	27.0	1,743	19.5	370	4.1
Asian	2,124	56.3	866	23.0	648	17.2	134	3.6
American Indian	772	35.1	603	27.4	660	30.0	166	7.5
Native Hawaiian or Pacific Islander	162	43.1	103	27.4	86	22.9	25	6.6
Alaska Native	87	62.1	32	22.9	20	14.3	1	0.7
Hispanic ethnicity								
No	25,066	45.2	14,706	26.5	12,588	22.7	3,050	5.5
Yes	7,982	46.2	4,487	26.0	3,868	22.4	949	5.5
English learner (EL) participation								
Not EL eligible or monitored	31,291	45.7	17,963	26.2	15,459	22.6	3,777	5.5
EL eligible or monitored	1,757	41.8	1,230	29.2	997	23.7	222	5.3

Table 7.8. Mathematics Performance Level Distributions, by Demographic Subgroup (*N* = 72,643)

	Emerging		Approaching		Target		Advanced	
Subgroup	*n*	%	*n*	%	*n*	%	*n*	%
Gender								
Male	27,967	57.5	11,476	23.6	6,162	12.7	2,993	6.2
Female	14,662	61.0	5,827	24.2	2,564	10.7	992	4.1
Race								
White	25,042	58.8	10,333	24.3	4,998	11.7	2,195	5.2
African American	8,181	55.7	3,588	24.4	1,940	13.2	979	6.7
Two or more races	5,646	63.3	1,976	22.2	944	10.6	347	3.9
Asian	2,430	64.6	729	19.4	385	10.2	219	5.8
American Indian	1,047	47.7	551	25.1	382	17.4	215	9.8
Native Hawaiian or Pacific Islander	186	49.5	92	24.5	69	18.4	29	7.7
Alaska Native	97	69.3	34	24.3	8	5.7	1	0.7
Hispanic ethnicity								
No	32,589	58.9	13,390	24.2	6,499	11.7	2,892	5.2
Yes	10,040	58.1	3,913	22.7	2,227	12.9	1,093	6.3
English learner (EL) participation								
Not EL eligible or monitored	40,503	59.2	16,270	23.8	8,045	11.8	3,623	5.3
EL eligible or monitored	2,126	50.6	1,033	24.6	681	16.2	362	8.6

7.2.3. *Linkage Level Mastery*

As described earlier in the chapter, overall performance in each subject is calculated based on the number of linkage levels mastered across all EEs. Results indicate the highest linkage level the student mastered for each EE. The linkage levels are (in order): Initial Precursor, Distal Precursor, Proximal Precursor, Target, and Successor. A student can be a master of zero, one, two, three, four, or all five linkage levels, within the order constraints. For example, if a student masters the Proximal Precursor level, they also master all linkage levels lower in the order (i.e., Initial Precursor and Distal Precursor). This section summarizes the distribution of students by highest linkage level mastered across all EEs. For each student, the highest linkage level mastered across all tested EEs was calculated. Then, for each grade and subject, the number of students with each linkage level as their highest mastered linkage level across all EEs was summed and then divided by the total number of students who tested in the grade and subject. This resulted in the proportion of students for whom each level was the highest level mastered.

Table 7.9 and Table 7.10 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, the Initial Precursor level was the highest level that students mastered 8% of the time. For ELA, the average percentage of students who mastered as high as the Target or Successor linkage level across all EEs ranged from approximately 43% in grade 3 to 64% in grade 12. For mathematics, the average percentage of students who mastered the Target or Successor linkage level across all EEs ranged from approximately 12% in grade 11 to 30% in grade 4.

Table 7.9. Students' Highest Linkage Level Mastered Across ELA EEs, by Grade

	Linkage Level					
Grade	No evidence (%)	IP (%)	DP (%)	PP (%)	T (%)	S (%)
3 (*n* = 8,777)	2.6	8.4	24.2	21.5	18.4	24.9
4 (*n* = 9,518)	3.0	6.7	23.7	12.3	16.1	38.2
5 (*n* = 9,525)	2.3	6.6	26.6	13.5	12.5	38.5
6 (*n* = 9,559)	2.5	6.7	28.3	19.0	8.9	34.6
7 (*n* = 9,555)	3.0	4.1	22.9	13.0	14.3	42.7
8 (*n* = 9,432)	3.5	4.9	23.0	14.7	16.1	37.8
9 (*n* = 6,739)	4.3	8.7	19.1	18.4	16.9	32.5
10 (*n* = 3,454)	4.1	11.2	20.6	17.2	15.1	31.7
11 (*n* = 6,112)	3.7	6.2	25.6	13.5	14.6	36.4
12 (*n* = 25)	0.0	4.0	24.0	8.0	8.0	56.0

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

Table 7.10. Students' Highest Linkage Level Mastered Across Mathematics EEs, by Grade

	Linkage Level					
Grade	No evidence (%)	IP (%)	DP (%)	PP (%)	T (%)	S (%)
3 (*n* = 8,763)	5.4	25.3	30.7	14.8	12.1	11.7
4 (*n* = 9,503)	2.3	16.0	24.1	27.4	16.2	14.0
5 (*n* = 9,526)	4.3	21.3	37.6	16.3	9.3	11.2
6 (*n* = 9,551)	6.4	19.9	23.2	28.1	10.6	11.8
7 (*n* = 9,544)	4.4	20.2	20.3	28.8	18.0	8.3
8 (*n* = 9,426)	4.3	11.2	23.6	34.2	15.2	11.4
9 (*n* = 6,736)	7.8	23.9	19.7	22.0	13.0	13.6
10 (*n* = 3,451)	8.7	36.1	30.2	11.5	10.4	3.1
11 (*n* = 6,117)	9.0	25.9	43.6	9.3	9.4	2.8
12 (*n* = 26)	3.8	26.9	38.5	15.4	15.4	0.0

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

7.3. Data Files

Data files were made available to DLM state partners following the spring 2019 administration. Similar to prior years, the General Research File (GRF) contained student results, including each student's highest linkage level mastered for each EE and final performance level for the subject for all students who completed any testlets. In addition to the GRF, the DLM Consortium delivered several supplemental files. Consistent with prior years, the Special Circumstances File provided information about which students and EEs were affected by extenuating circumstances (e.g., chronic absences), as defined by each state. State partners also received a supplemental file to identify exited students. The exited students file included all students who exited at any point during the academic year. In the event of observed incidents during assessment delivery, state partners are provided with an Incident File describing students impacted. For a description of incidents observed during the 2018–2019 administration, see Chapter 4 of this manual.

Consistent with prior delivery cycles, state partners were provided with a two-week review window following data file delivery to review the files and invalidate student records in the GRF. Decisions about whether to invalidate student records are informed by individual state policy. If changes were made to the GRF, state partners submitted final GRFs via Educator Portal. The final GRF was used to generate score reports.

In addition to the GRF and its supplemental files, states were provided with two additional de-identified data files: a teacher survey data file and a test administration observations data file. The teacher survey file provided state-specific teacher survey responses, with all identifying information about the student and educator removed. The test administration observations file provided test administration observation responses with any identifying information removed. For more information regarding teacher survey content and response rates, see Chapter 4 of this manual. For more information about test administration observation results, see Chapter 9 of this manual.

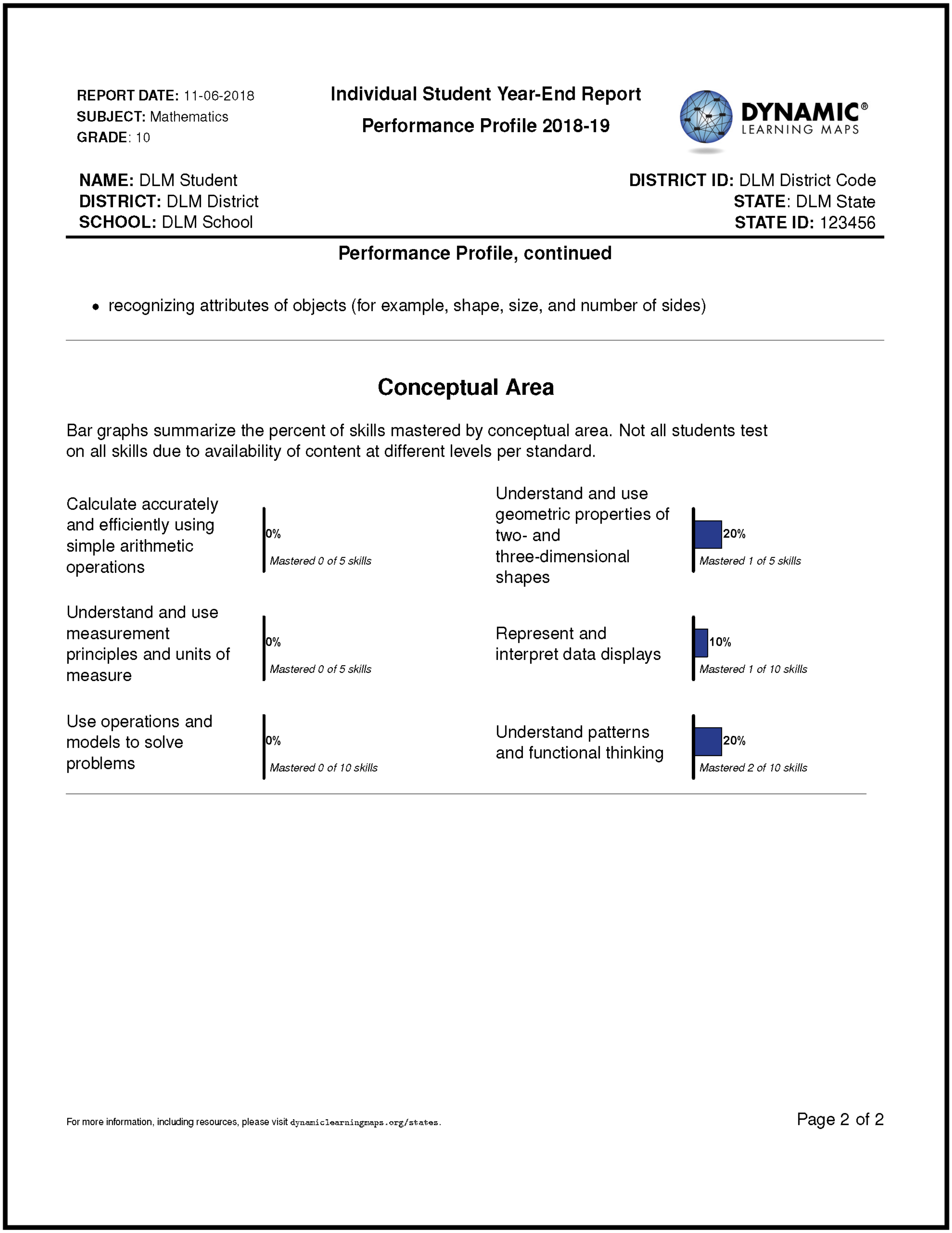
7.4. Score Reports

The DLM Consortium provides assessment results to all member states to report to parents/guardians, educators, and state and local education agencies. Individual Student Score Reports summarized student performance on the assessment by subject. Several aggregated reports were provided to state and local education agencies, including reports for the classroom, school, district, and state. No changes were made to the structure of aggregated reports during spring 2019. Changes to the Individual Student Score Reports are summarized below. For a complete description of score reports, including aggregated reports, see Chapter 7 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016).

7.4.1. Individual Student Score Reports

During the 2018–2019 year, minor changes were made to the Individual Student Score Reports. A website was added to the footnote of the report which linked to additional resources related to the DLM assessment and understanding student results. On the Performance Profile portion of the report, a text description of the bar graphs was added to aid in interpretation.

A sample Performance Profile portion of the report reflecting the 2019 changes is provided in Figure 7.1.



REPORT DATE: 11-06-2018
SUBJECT: Mathematics
GRADE: 10

Individual Student Year-End Report
Performance Profile 2018-19

DYNAMIC® LEARNING MAPS

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

DISTRICT ID: DLM District Code
STATE: DLM State
STATE ID: 123456

Performance Profile, continued

- recognizing attributes of objects (for example, shape, size, and number of sides)

Conceptual Area

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

Calculate accurately and efficiently using simple arithmetic operations	0% *Mastered 0 of 5 skills*	Understand and use geometric properties of two- and three-dimensional shapes	20% *Mastered 1 of 5 skills*
Understand and use measurement principles and units of measure	0% *Mastered 0 of 5 skills*	Represent and interpret data displays	10% *Mastered 1 of 10 skills*
Use operations and models to solve problems	0% *Mastered 0 of 10 skills*	Understand patterns and functional thinking	20% *Mastered 2 of 10 skills*

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

Page 2 of 2

Figure 7.1. Example page of the Performance Profile for 2018–2019.

7.5. Quality Control Procedures for Data Files and Score Reports

No changes were made to the manual or automated quality control procedures for spring 2019. For a complete description of quality control procedures, see Chapter 7 in the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016) and *2015–2016 Technical Manual—Year-End Model* (DLM Consortium, 2017a).

7.6. Conclusion

Following the spring 2019 administration, six data files were delivered to state partners: GRF, special circumstance code file, exited students file, incident file, teacher survey data file, and test administration observations file. Overall, between 7% and 40% of students achieved at the At Target or Advanced levels across all grades and subjects, which is consistent with prior years. An incident file was delivered describing the impact of the two reported incidents. Minor changes were made to score reports to aid in interpretation.

8. Reliability

Chapter 8 of the Dynamic Learning Maps® (DLM®) Alternate Assessment System *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016) describes the methods used to calculate reliability for the DLM assessment system and provided results at six levels, consistent with the levels of reporting. The *2015–2016 Technical Manual Update—Year-End Model* (DLM Consortium, 2017a) expands the description of the methods used to calculate reliability and provides results at six reporting levels. This chapter provides a high-level summary of the methods used to calculate reliability, along with updated evidence from the 2018–2019 administration year for six levels.

For a complete description of the simulation-based methods used to calculate reliability for DLM assessments, including the psychometric background, see the *2015–2016 Technical Manual Update—Year-End Model* (DLM Consortium, 2017a).

8.1. Background Information on Reliability Methods

The reliability information presented in this chapter adheres to guidance given in the *Standards for Educational and Psychological Testing* (American Educational Research Association et al. [AERA et al.], 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 with the design of the assessment and interpretations of results.

Consistent with the levels at which DLM results are reported, this chapter provides results for six types of reliability evidence. For more information on DLM reporting, see Chapter 7 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016). The types of reliability evidence for DLM assessments include (a) classification to overall performance level (performance level reliability); (b) the total number of linkage levels mastered within a subject (subject reliability; provided for ELA and mathematics); (c) the number of linkage levels mastered within each conceptual area for ELA and mathematics (conceptual area reliability); (d) the number of linkage levels mastered within each Essential Element (EE; EE reliability); (e) the classification accuracy of each linkage level within each EE (linkage level reliability); and (f) 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.

8.2. 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).

The simulation used to estimate reliability for DLM versions of scores and classifications considers the unique design and administration of DLM assessments. The use of simulation is necessitated by two factors: the assessment blueprint and the results that classification-based administrations give.

Because of the limited number of items students complete to cover the blueprint, students take only minimal items per EE. The reliability simulation replicates DLM classification-based scores from real examinees based upon the actual set of items each examinee took. Therefore, this simulation replicates the administered items for the examinees. Because the simulation is based on a replication of the same items administered to examinees, the two administrations are perfectly parallel.

8.2.1. 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 assessment data to generate simulated examinees. Using this process guarantees that the simulation takes on characteristics of the DLM operational assessment data that are likely to affect reliability results. For one simulated examinee, the process was as follows:

1. Draw with replacement the student record of one student from the operational assessment data (i.e., spring window). Use the student's originally scored pattern of linkage level mastery and non-mastery 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[5] for the items of the testlet, conditional on the profile of linkage level mastery or non-mastery for the student.
3. Score the simulated item responses using the operational DLM scoring procedure, estimating linkage level mastery or non-mastery for the simulated student. See Chapter 5 of the *2015–2016 Technical Manual Update—Year-End Model* (DLM Consortium, 2017a) for more information.[6]
4. Compare the estimated linkage level mastery or non-mastery to the known values from Step 2 for all linkage levels at which the student was administered items.

Steps 1 through 4 are then repeated 2,000,000 times to create the full simulated data set. Figure 8.1 shows the steps of the simulation process as a flow chart.

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

[6] All three scoring rules were included when scoring the simulated responses to be consistent with the operational scoring procedure. The scoring rules are described further in Chapter 5 of this manual.

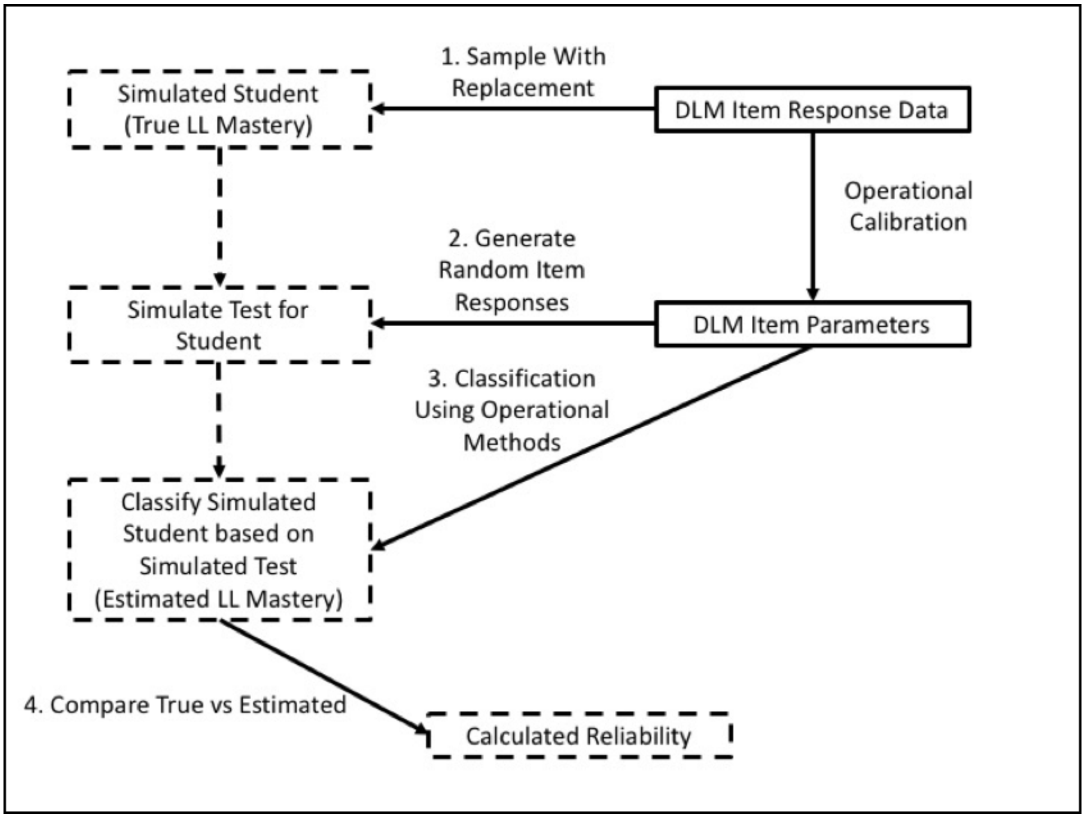

1. Sample With Replacement
Simulated Student (True LL Mastery)
DLM Item Response Data
Operational Calibration
2. Generate Random Item Responses
Simulate Test for Student
DLM Item Parameters
3. Classification Using Operational Methods
Classify Simulated Student based on Simulated Test (Estimated LL Mastery)
4. Compare True vs Estimated
Calculated Reliability

Figure 8.1. Simulation process for creating reliability evidence. *Note.* LL = linkage level.

8.3. 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).

This chapter provides reliability evidence for six levels of data: (a) performance level reliability, (b) subject reliability, (c) conceptual area reliability, (d) EE reliability, (e) linkage level reliability, and (f) conditional reliability by linkage level. With 242 EEs, each comprising five linkage levels, the procedure includes 1,210 analyses to summarize reliability results. Because of the number of analyses, this chapter includes a summary of the reported evidence. An online appendix[7] provides a full report of reliability evidence for all 1,210 linkage levels and 242 EEs. The full set of evidence is furnished in accordance with Standard 2.12.

This chapter provides reliability evidence at six levels, which ensures that the simulation and resulting reliability evidence are aligned with Standard 2.2. Additionally, providing reliability evidence for each of the six levels ensures that these reliability estimation procedures meet Standard 2.5.

8.3.1. Performance Level Reliability Evidence

The DLM Consortium reports results using four performance levels. The scoring procedure sums the linkage levels mastered in each subject, and cut points are applied to distinguish between performance categories.

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

1. the polychoric correlation between the true and estimated performance levels within a grade and subject,
2. the correct classification rate between the true and estimated performance levels within a grade and subject, and
3. the correct classification kappa between the true and estimated performance levels within a grade and subject.

Table 8.1 presents this information across all grades and subjects. Polychoric correlations between true and estimated performance level range from .949 to .991. Correct classification rates range from .854 to .932 and Cohen’s kappa values are between .820 and .960. 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 into performance level categories.

[7] http://dynamiclearningmaps.org/reliabevid

Table 8.1. Summary of Performance Level Reliability Evidence

Grade	Subject	Polychoric correlation	Correct classification rate	Cohen's kappa
3	English language arts	.980	.928	.950
3	Mathematics	.990	.899	.949
4	English language arts	.984	.918	.946
4	Mathematics	.990	.895	.952
5	English language arts	.987	.932	.960
5	Mathematics	.988	.888	.944
6	English language arts	.991	.904	.947
6	Mathematics	.988	.896	.939
7	English language arts	.990	.899	.947
7	Mathematics	.986	.905	.928
8	English language arts	.987	.885	.937
8	Mathematics	.983	.898	.913
9	English language arts	.989	.889	.934
9	Mathematics	.988	.896	.922
10	English language arts	.985	.911	.941
10	Mathematics	.962	.876	.850
11	English language arts	.981	.902	.932
11	Mathematics	.949	.854	.820

8.3.2. *Subject Reliability Evidence*

Subject reliability provides consistency evidence for the number of linkage levels mastered across all EEs for a given subject and grade level. Because students are assessed on multiple linkage levels within a subject, subject reliability evidence is similar to reliability evidence for testing programs that use summative assessments to describe subject performance. That is, the number of linkage levels mastered within a subject is analogous to the number of items answered correctly (i.e., total score) in a different type of testing program.

Subject reliability evidence compares the true and estimated number of linkage levels mastered across all tested levels for a given subject. Reliability is reported with three summary values:

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

Table 8.2 shows the three summary values for each grade and subject. Classification rate information is provided in accordance with Standard 2.16. The two summary statistics included in Table 8.2 also meet Standard 2.19. The correlation between true and estimated number of linkage levels mastered ranges from .959 to .992. Students' average correct classification rates range from .950 to .990 and average Cohen's kappa values range from .844 to .976. These values indicate the DLM scoring procedure of reporting the number of linkage levels mastered provides reliable results of total linkage levels mastered.

Table 8.2. Summary of Subject Reliability Evidence

Grade	Subject	Linkage levels mastered correlation	Average student correct classification	Average student Cohen's kappa
3	English language arts	.990	.963	.889
3	Mathematics	.983	.973	.908
4	English language arts	.991	.959	.876
4	Mathematics	.988	.965	.886
5	English language arts	.992	.964	.895
5	Mathematics	.988	.966	.876
6	English language arts	.990	.962	.888
6	Mathematics	.983	.973	.918
7	English language arts	.990	.958	.878
7	Mathematics	.986	.972	.909
8	English language arts	.988	.950	.844
8	Mathematics	.984	.968	.901
9	English language arts	.987	.955	.865
9	Mathematics	.979	.987	.973
10	English language arts	.989	.955	.857
10	Mathematics	.968	.990	.976
11	English language arts	.986	.955	.867
11	Mathematics	.959	.988	.970

8.3.3. *Conceptual Area Reliability Evidence*

Within each subject, students are assessed on multiple content strands. These strands of related EEs describe the overarching sections of the learning map model that is the foundation of the development of DLM assessments. For more information, see Chapter 2 in the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016). The strands used for reporting are the conceptual areas in ELA and mathematics. Because Individual Student Score Reports summarize the number and percentage of linkage levels students mastered in each conceptual area (see Chapter 7 of this manual for more information), reliability evidence is also provided at these levels in their respective subjects.

Conceptual area reliability provides consistency evidence for the number of linkage levels mastered across all EEs in each conceptual area for each grade and subject. Because conceptual area reporting summarizes the total number of linkage levels a student mastered, the statistics reported for conceptual area reliability are the same as those reported for subject 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, and
3. the correct classification kappa for which linkage levels were mastered as averaged across all simulated students for each conceptual area.

Table 8.3 and Table 8.4 show the three summary values for each conceptual area, by grade, for ELA and mathematics, respectively. Values range from .759 to .999 in ELA and from .632 to .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 8.3. Summary of ELA Conceptual Area Reliability Evidence

Grade	Conceptual area	Linkage levels mastered correlation	Average student correct classification	Average student Cohen's kappa
3	ELA.C1.1	.977	.983	.964
3	ELA.C1.2	.972	.985	.970
3	ELA.C1.3	.905	.995	.994
3	ELA.C2.1	.906	.994	.992
4	ELA.C1.1	.980	.982	.959
4	ELA.C1.2	.972	.974	.937
4	ELA.C1.3	.903	.999	.998
4	ELA.C2.1	.968	.996	.994
5	ELA.C1.1	.960	.995	.992
5	ELA.C1.2	.984	.977	.943
5	ELA.C1.3	.960	.991	.985
5	ELA.C2.1	.953	.997	.997
6	ELA.C1.1	.759	.998	.998
6	ELA.C1.2	.983	.969	.917
6	ELA.C1.3	.958	.994	.991
6	ELA.C2.1	.929	.997	.997
7	ELA.C1.1	.787	.998	.997
7	ELA.C1.2	.983	.977	.942
7	ELA.C1.3	.962	.988	.978
7	ELA.C2.1	.937	.988	.978
8	ELA.C1.2	.981	.960	.881
8	ELA.C1.3	.940	.990	.984
8	ELA.C2.1	.955	.987	.976
9	ELA.C1.2	.982	.971	.924
9	ELA.C1.3	.930	.988	.980
9	ELA.C2.1	.904	.989	.981

Table 8.3. Summary of ELA Conceptual Area Reliability Evidence *(continued)*

Grade	Conceptual area	Linkage levels mastered correlation	Average student correct classification	Average student Cohen's kappa
9	ELA.C2.2	.908	.997	.996
10	ELA.C1.2	.985	.970	.912
10	ELA.C1.3	.928	.988	.980
10	ELA.C2.1	.896	.989	.981
10	ELA.C2.2	.905	.997	.996
11	ELA.C1.2	.977	.977	.946
11	ELA.C1.3	.955	.985	.970
11	ELA.C2.1	.927	.986	.975
11	ELA.C2.2	.834	.995	.994

Table 8.4. Summary of Mathematics Conceptual Area Reliability Evidence

Grade	Conceptual area	Linkage levels mastered correlation	Average student correct classification	Average student Cohen's kappa
3	M.C1.1	.921	.995	.992
3	M.C1.3	.877	.998	.998
3	M.C2.2	.865	.999	.999
3	M.C3.1	.924	.995	.994
3	M.C3.2	.808	.998	.998
3	M.C4.1	.935	.996	.994
3	M.C4.2	.730	.998	.998
4	M.C1.1	.862	.997	.997
4	M.C1.2	.867	.995	.993
4	M.C1.3	.903	.999	.998
4	M.C2.1	.940	.994	.990
4	M.C2.2	.921	.999	.999
4	M.C3.1	.951	.996	.994
4	M.C3.2	.837	.998	.998
4	M.C4.1	.896	.995	.993
4	M.C4.2	.632	.997	.997
5	M.C1.1	.791	.994	.992
5	M.C1.2	.944	.993	.989
5	M.C1.3	.941	.997	.996
5	M.C2.1	.948	.997	.996
5	M.C2.2	.936	.999	.999
5	M.C3.1	.943	.993	.989
5	M.C3.2	.894	.998	.998
5	M.C4.2	.703	.997	.997

Table 8.4. Summary of Mathematics Conceptual Area Reliability Evidence *(continued)*

Grade	Conceptual area	Linkage levels mastered correlation	Average student correct classification	Average student Cohen's kappa
6	M.C1.1	.861	.999	.998
6	M.C1.2	.904	.996	.995
6	M.C1.3	.933	.996	.995
6	M.C2.2	.960	.997	.997
6	M.C3.2	.796	.998	.998
6	M.C4.1	.894	.993	.989
7	M.C1.1	.917	.996	.995
7	M.C1.2	.859	.998	.998
7	M.C1.3	.927	.993	.989
7	M.C2.1	.949	.997	.995
7	M.C2.2	.878	.999	.999
7	M.C3.2	.909	.997	.997
7	M.C4.1	.800	.998	.998
7	M.C4.2	.795	.998	.998
8	M.C1.1	.758	.997	.997
8	M.C1.2	.865	.998	.998
8	M.C1.3	.903	.996	.995
8	M.C2.1	.926	.992	.988
8	M.C2.2	.897	.999	.999
8	M.C3.2	.910	.999	.998
8	M.C4.1	.710	.998	.998
8	M.C4.2	.928	.991	.985
9	M.C1.3	.942	.995	.993
9	M.C2.1	.921	.997	.996
9	M.C2.2	.865	.999	.999
9	M.C4.1	.848	.997	.996
10	M.C1.3	.916	.999	.999
10	M.C2.1	.881	.999	.999
10	M.C3.1	.699	.998	.998
10	M.C3.2	.885	.998	.997
10	M.C4.1	.883	.998	.997
10	M.C4.2	.880	.998	.997
11	M.C1.3	.877	.998	.998
11	M.C2.1	.753	.999	.998
11	M.C3.2	.805	.999	.999
11	M.C4.2	.934	.994	.990

8.3.4. EE 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, EE-level results are reported as the highest linkage level mastered per EE. Considering subject scores as total scores from an entire test, evidence at the EE level is finer grained than reporting at a subject strand level, which is commonly reported by other testing programs. EEs are specific standards within the subject 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, and
3. the correct classification kappa for the number of linkage levels mastered within an EE.

Because there are 242 EEs, the summaries are reported herein according to the number and proportion of EEs that fall within a given range of an index value (results for individual EEs can be found in the online appendix[8]). Results are given in both tabular and graphical forms. Table 8.5 and Figure 8.2 provide the proportions and the number of EEs, respectively, falling within prespecifed ranges of values for the three reliability summary statistics (i.e., correct classification rate, kappa, correlation). In general, the reliability summaries show strong evidence for reliability for the number of linkage levels mastered within EEs.

Table 8.5. Reliability Summaries Across All EEs: Proportion of EEs Falling Within a Specified Index Range

	Index range								
Reliability Index	**< .60**	**0.60-0.64**	**0.65-0.69**	**0.70-0.74**	**0.75-0.79**	**0.80-0.84**	**0.85-0.89**	**0.90-0.94**	**0.95-1.00**
Polychoric correlation	<.001	<.001	.004	.004	.012	.058	.207	.525	.190
Correct classification rate	<.001	<.001	<.001	.029	.136	.347	.397	.087	.004
Cohen's kappa	<.001	.004	.008	.017	.050	.165	.343	.384	.029

[8] http://dynamiclearningmaps.org/reliabevid

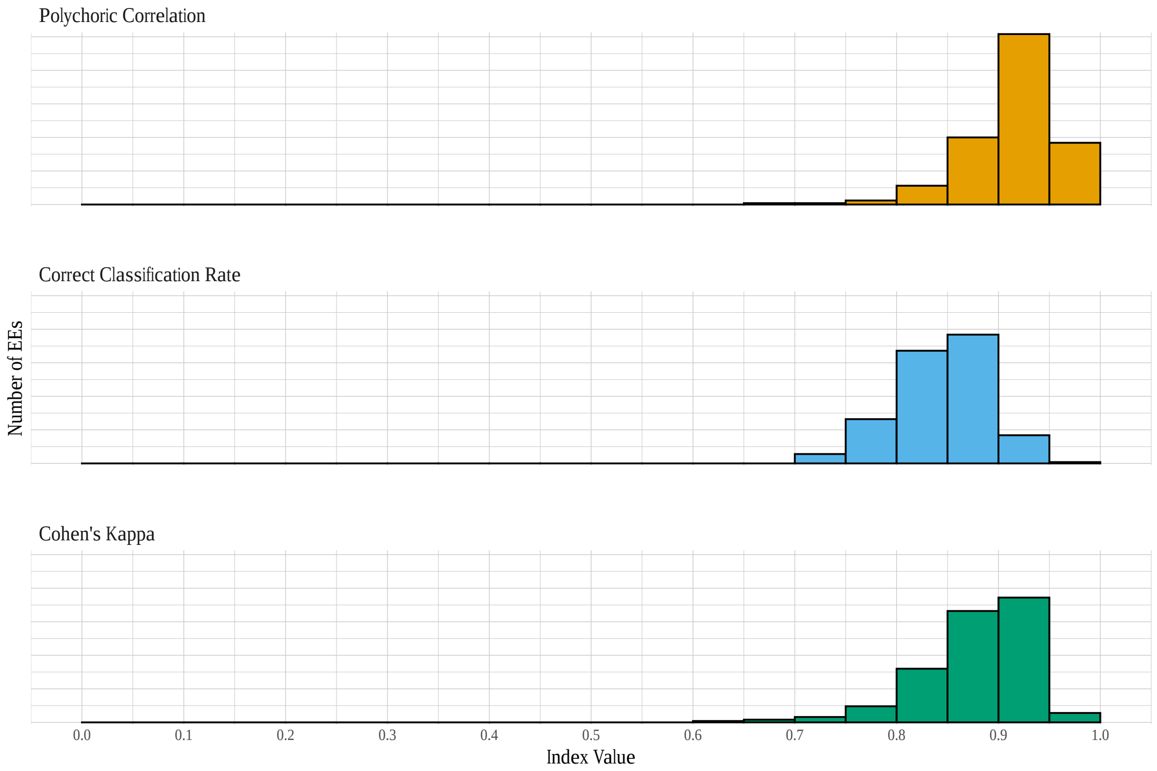


Figure 8.2. Number of linkage levels mastered within EE reliability summaries.

8.3.5. Linkage Level Reliability Evidence

Evidence at the linkage level comes from comparing the true and estimated mastery status for each of the 1,210 linkage levels in the operational DLM assessment.[9] This level of reliability reporting is even finer grained than 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. All reported summary statistics are based on the resulting contingency tables: 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, and
3. the correct classification kappa for the mastery status of each linkage level.

[9] 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 Model* (DLM Consortium, 2016).

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 for individual linkage levels can be found in the online appendix[10]). Results are given in both tabular and graphical forms. Table 8.6 and Figure 8.3 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, correlation). The kappa value and tetrachoric correlation for one linkage level could not be computed because all students were labeled as masters of the linkage level.

The correlations and correct classification rates show reliability evidence for the classification of mastery at the linkage level. Across all linkage levels, three had tetrachoric correlation values below .6, zero had a correct classification rate below .6, and 38 had a kappa value below 0.6.

Table 8.6. Reliability Summaries Across All Linkage Levels: Proportion of Linkage Levels Falling Within a Specified Index Range

	Index range								
Reliability Index	**< .60**	**0.60-0.64**	**0.65-0.69**	**0.70-0.74**	**0.75-0.79**	**0.80-0.84**	**0.85-0.89**	**0.90-0.94**	**0.95-1.00**
Tetrachoric correlation	.003	.002	.001	.001	.003	.017	.045	.162	.766
Correct classification rate	<.001	<.001	<.001	<.001	.002	.021	.093	.374	.510
Cohen’s kappa	.032	.032	.045	.081	.135	.191	.224	.151	.107

[10] http://dynamiclearningmaps.org/reliabevid

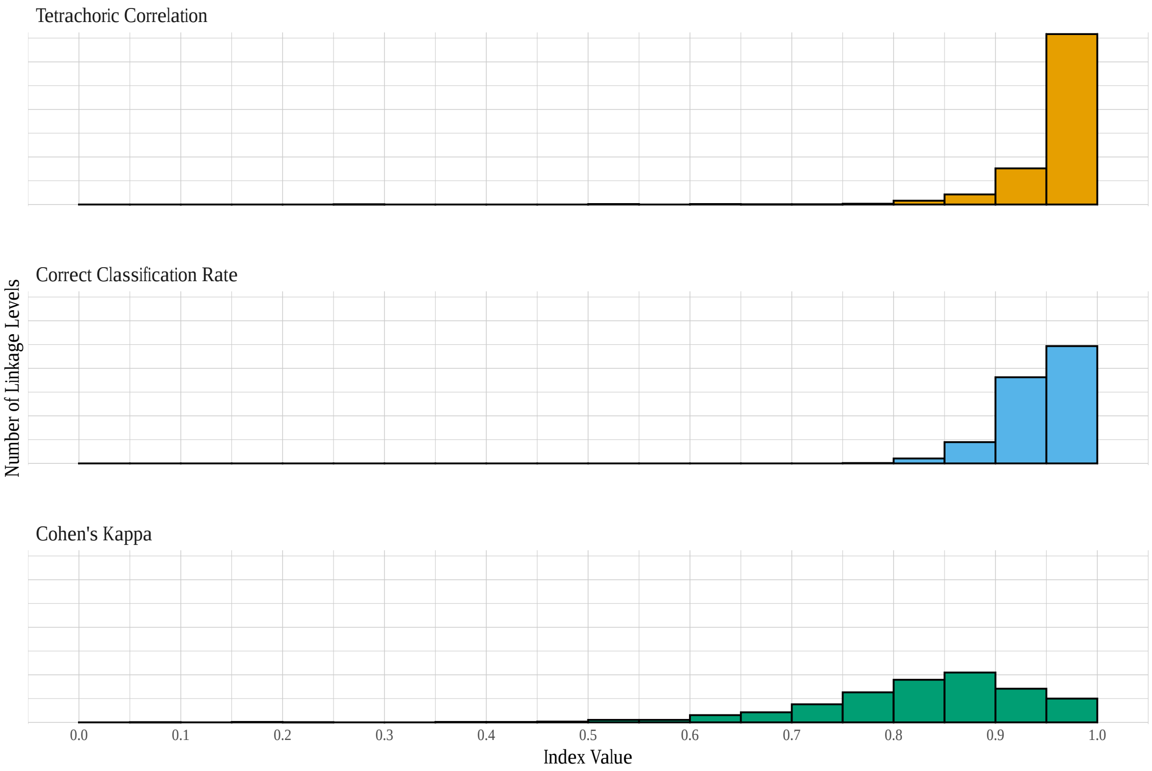


Figure 8.3. Summaries of linkage level reliability.

8.3.6. *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 (tetrachoric correlation, correct classification rate, kappa).

Figure 8.4 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, kappa). The correlations and correct classification rates generally indicate that all five linkage levels provide reliable classifications of student mastery; results are fairly consistent across all linkage levels for each of the three statistics reported.

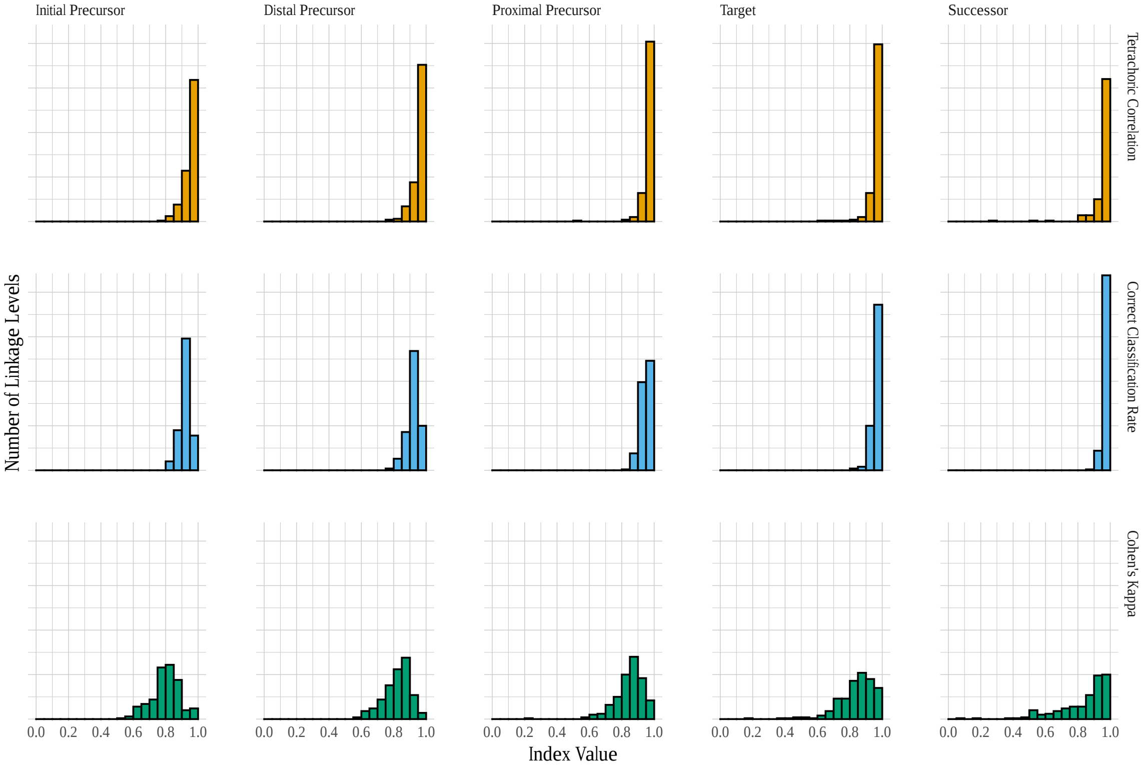


Figure 8.4. Conditional reliability evidence summarized by linkage level.

8.4. Conclusion

In summary, reliability measures for the DLM assessment system address the standards set forth by AERA et al. (2014). The DLM methods are consistent with assumptions of diagnostic classification modeling and yield evidence to support the argument for internal consistency of the program for each level of reporting. Because the reliability results depend upon the model used to calibrate and score the assessment, any changes to the model or evidence obtained when evaluating model fit also affect reliability results. As with any selected methodology for evaluating reliability, the current results assume that the model and model parameters used to score DLM assessments are correct. However, unlike other traditional measures of reliability that often require unattainable assumptions about equivalent test forms, the simulation method described in this chapter provides a replication of the same test items (i.e., perfectly parallel forms), which theoretically reduces the amount of variance that may be found in test scores across administrations. Furthermore, while the reliability measures in general may be higher than those observed for some traditionally scored assessments, research has found that diagnostic classification models have greater reliability with fewer items (e.g., Templin & Bradshaw, 2013), suggesting the results are expected.

9. Validity Studies

The preceding chapters and the Dynamic Learning Maps® (DLM®) Alternate Assessment System *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016) provide evidence in support of the overall validity argument for results produced by the DLM assessment. This chapter presents additional evidence collected during 2018–2019 for four of the five critical sources of evidence described in *Standards for Educational and Psychological Testing* (AERA et al., 2014): evidence based on test content, response process, internal structure, and consequences of testing. Additional evidence can be found in Chapter 9 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016) and the subsequent annual technical manual updates (DLM Consortium, 2017a, 2017b, 2018a).

9.1. 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). This section presents results from data collected during 2018–2019 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 9 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016).

9.1.1. Opportunity to Learn

After administration of the spring 2019 operational assessments, teachers were invited to complete a survey about the assessment (see Chapter 4 of this manual for more information on recruitment and response rates). The survey included three blocks of items. The first and third blocks were fixed forms assigned to all teachers. For the second block, teachers received one randomly assigned section.

The first block of the survey served several purposes.[11] One item provided information about the relationship between students' learning opportunities before testing and the test content (i.e., testlets) they encountered on the assessment. The survey asked teachers to indicate the extent to which they judged test content to align with their instruction across all testlets; Table 9.1 reports the results. Approximately 72% of responses (n = 28,809) reported that most or all reading testlets matched instruction, compared to 46% (n = 18,181) for writing and 59% (n = 23,699) for mathematics. More specific measures of instructional alignment are planned to better understand the extent that content measured by DLM assessments matches students' academic instruction.

Table 9.1. Teacher Ratings of Portion of Testlets That Matched Instruction

	None		Some (< half)		Most (> half)		All		N/A	
Subject	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Reading	2,283	5.7	8,307	20.7	16,236	40.5	12,573	31.4	690	1.7
Writing	4,113	10.4	8,027	20.3	10,533	26.6	7,648	19.3	9,227	23.3
Mathematics	3,513	8.8	11,891	29.8	14,799	37.0	8,900	22.3	856	2.1

[11] Results for other survey items are reported later in this chapter and in Chapter 4 in this manual.

The second block of the survey was randomly spiraled so that teachers received one randomly assigned section. In one of the randomly assigned sections, a subset of teachers were asked to indicate the approximate number of hours they spent instructing students on each of the conceptual areas by subject. Teachers responded using a five-point scale: *0-5 hours*, *6-10 hours*, *11-15 hours*, *16-20 hours*, or *more than 20 hours*. Table 9.2 and Table 9.3 indicate the amount of instructional time spent on conceptual areas, for ELA and mathematics, respectively. Using 11 or more hours per conceptual area as a criterion for instruction, 64% of the teachers provided this amount of instruction to their students in ELA, and 52% did so in mathematics.

Table 9.2. Instructional Time Spent on ELA Conceptual Areas

		Number of hours									
		0-5		**6-10**		**11-15**		**16-20**		**>20**	
Conceptual area	**Median**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Determine critical elements of text	11-15	1,088	23.0	684	14.5	670	14.2	706	14.9	1,577	33.4
Construct understandings of text	16-20	791	16.8	668	14.2	650	13.8	777	16.5	1,828	38.8
Integrate ideas and information from text	16-20	896	19.1	705	15.0	713	15.2	804	17.1	1,577	33.6
Use writing to communicate	11-15	1,130	24.0	709	15.1	675	14.3	680	14.4	1,512	32.1
Integrate ideas and information in writing	11-15	1,257	26.9	723	15.4	685	14.6	716	15.3	1,300	27.8
Use language to communicate with others	>20	485	10.3	459	9.7	574	12.2	748	15.8	2,457	52.0
Clarify and contribute in discussion	16-20	836	17.8	664	14.1	683	14.5	796	16.9	1,718	36.6
Use sources and information	11-15	1,356	28.8	784	16.6	726	15.4	705	15.0	1,140	24.2
Collaborate and present ideas	11-15	1,292	27.4	786	16.7	734	15.6	753	16.0	1,143	24.3

Table 9.3. Instructional Time Spent on Mathematics Conceptual Areas

		Number of hours									
		0-5		6-10		11-15		16-20		>20	
Conceptual area	**Median**	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Understand number structures (counting, place value, fraction)	16-20	739	15.3	632	13.1	573	11.9	728	15.1	2,151	44.6
Compare, compose, and decompose numbers and steps	11-15	1,257	26.2	795	16.6	688	14.3	744	15.5	1,313	27.4
Calculate accurately and efficiently using simple arithmetic operations	16-20	1,115	23.1	574	11.9	597	12.4	726	15.0	1,814	37.6
Understand and use geometric properties of two- and three-dimensional shapes	6-10	1,608	33.4	1,037	21.6	749	15.6	701	14.6	713	14.8
Solve problems involving area, perimeter, and volume	0-5	2,558	53.1	782	16.2	541	11.2	489	10.1	449	9.3
Understand and use measurement principles and units of measure	6-10	1,700	35.4	1,108	23.1	770	16.0	617	12.8	610	12.7
Represent and interpret data displays	6-10	1,670	34.8	979	20.4	802	16.7	650	13.6	696	14.5
Use operations and models to solve problems	11-15	1,416	29.5	754	15.7	724	15.1	798	16.6	1,112	23.1
Understand patterns and functional thinking	11-15	1,113	23.1	952	19.8	900	18.7	818	17.0	1,036	21.5

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

Table 9.4 summarizes the Spearman rank-order correlations between ELA conceptual area instructional time and linkage levels mastered in the conceptual area and between mathematics conceptual area instructional time and linkage levels mastered in the conceptual area. Correlations ranged from .12 to .38, with the strongest correlations observed for writing conceptual areas (ELA.C2.1 and ELA.C2.2) in ELA and measurement, data, and analytic procedures conceptual areas (M.C3.1 and M.C3.2) collectively in mathematics.

Table 9.4. Correlation Between Instuction Time and Linkage Levels Mastered

Statement	Correlation with instruction time
English language arts	
ELA.C1.1: Determine critical elements of text	.17
ELA.C1.2: Construct understandings of text	.26
ELA.C1.3: Integrate ideas and information from text	.27
ELA.C2.1: Use writing to communicate	.32
ELA.C2.2: Integrate ideas and information in writing	.38
Mathematics	
M.C1.1: Understand number structures (counting, place value, fraction)	.12
M.C1.2: Compare, compose, and decompose numbers and steps	.24
M.C1.3: Calculate accurately and efficiently using simple arithmetic operations	.30
M.C2.1: Understand and use geometric properties of two- and three-dimensional shapes	.23
M.C2.2: Solve problems involving area, perimeter, and volume	.25
M.C3.1: Understand and use measurement principles and units of measure	.27
M.C3.2: Represent and interpret data displays	.26
M.C4.1: Use operations and models to solve problems	.31
M.C4.2: Understand patterns and functional thinking	.19

9.2. Evidence Based on Response Processes

The study of test takers' response processes provides evidence about the fit between the test construct and the nature of how students actually experience test content (AERA et al., 2014). The validity studies presented in this section include teacher survey data collected in spring 2019 regarding students' ability to respond to testlets, test administration observation data collected during

2018–2019, and a study of interrater agreement on the scoring of teacher-administered writing testlets during spring 2019. 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 9 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016).

9.2.1. *Evaluation of Test Administration*

After administering spring operational assessments in 2019, teachers provided feedback via a teacher survey. Survey data that inform evaluations of assumptions regarding response processes include teacher perceptions of students' ability to respond as intended, free of barriers, and with necessary supports available.[12]

One of the fixed-form sections of the spring 2019 teacher survey included three items about students' ability to respond. Teachers were asked to use a four-point scale (*strongly disagree, disagree, agree,* or *strongly agree*). Results were combined in the summary presented in Table 9.5. The majority of teachers (85% or greater) agreed or strongly agreed that their students (a) responded to items to the best of their knowledge and ability; (b) were able to respond regardless of disability, behavior, or health concerns; and (c) had access to all supports necessary to participate. These results are similar to those observed in previous years and suggest that students are able to effectively interact with and respond to the assessment content.

Table 9.5. Teacher Perceptions of Student Experience With Testlets

	SD		**D**		**A**		**SA**		**A+SA**	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
This student responded to the items on this assessment to the best of his or her knowledge and ability.	1,370	3.4	2,747	6.9	20,605	51.4	15,341	38.3	35,946	89.7
This student was able to respond to items regardless of his or her disability, behavior, or health concerns.	2,488	6.2	3,559	8.9	20,255	50.5	13,821	34.4	34,076	84.9
This student had access to all necessary supports in order to participate in the assessment.	993	2.5	1,257	3.1	19,388	48.3	18,474	46.1	37,862	94.4

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

9.2.2. *Test Administration Observations*

Test administration observations were conducted in multiple states during 2018–2019 to further understand student response processes. Students' typical test administration process with their

[12] Recruitment and response information for this survey is provided in Chapter 4 of this manual.

actual test administrator was observed. Test administration observations were collected by state and local education agency staff.

Consistent with previous years, the DLM Consortium used a test administration observation protocol to gather information about how educators in the consortium states deliver testlets to students with the most significant cognitive disabilities. This protocol gave observers, regardless of their role or experience with DLM assessments, a standardized way to describe how DLM testlets were administered. The test administration observation protocol captured data about student actions (e.g., navigation, responding), educator assistance, variations from standard administration, engagement, and barriers to engagement. The observation protocol was used only for descriptive purposes; it was not used to evaluate or coach educators or to monitor student performance. Most items on the protocol were a direct report of what was observed, such as how the test administrator prepared for the assessment and what the test administrator and student said and did. One section of the protocol asked observers to make judgments about the student's engagement during the session.

During computer-delivered testlets, students are intended to interact independently with a computer, using special devices such as alternate keyboards, touch screens, or switches as necessary. For teacher-administered testlets, the test administrator was responsible for setting up the assessment, delivering the testlet to the student, and recording responses in the Kite® system. The test administration protocol contained different questions specific to each type of testlet.

During the 2018–2019 academic year, the DLM Consortium added a new option for states to use when collecting test administration observation data. In previous years, the DLM Consortium collected observations using paper forms which were submitted via mail or email, or Qualtrics surveys completed in a web browser. In 2018–2019, the DLM Consortium also collected observations in a new mobile application, Kite Collector. The application allows state and local education agency staff to collect observation data electronically using smart phones and tablets. The Kite Collector mobile application allows observers to collect data offline without internet access in a testing location. Observers can then later upload their observations using the mobile application when they regain internet access.

In 2018–2019, the total number of observations increased to 523 observations collected by 10 states, a 436% increase compared with the 120 total observations collected by 5 states in 2017–2018.

Table 9.6 shows the number of observations collected by state. Of the observations, 341 (65%) were of computer-delivered assessments and 182 (35%) were of teacher-administered testlets. The observations were comprised of 259 (50%) ELA reading testlets, 23 (4%) ELA writing testlets, and 241 (46%) mathematics testlets.

Table 9.6. Teacher Observations by State (*N* = 523)

State	*n*	%
Arkansas	244	46.7
Colorado	3	0.6
Delaware	12	2.3
Iowa	51	9.8
Kansas	84	16.1
Missouri	26	5.0
New York	35	6.7
North Dakota	2	0.4
West Virginia	44	8.4
Wisconsin	22	4.2

To investigate the assumptions that underlie the claims of the validity argument, several parts of the test administration observation protocol were designed to provide information corresponding to the assumptions. One assumption addressed is that educators allow students to engage with the system as independently as they are able. For computer-delivered testlets, related evidence is summarized in Table 9.7; behaviors were identified as supporting, neutral, or nonsupporting. For example, clarifying directions (79% of observations) removes student confusion about the task demands as a source of construct-irrelevant variance and supports the student's meaningful, construct-related engagement with the item. In contrast, using physical prompts (e.g., hand-over-hand guidance) indicates that the teacher directly influenced the student's answer choice. Overall, 58% of observed behaviors were classified as supporting, with 1% of observed behaviors reflecting nonsupporting actions.

Table 9.7. Test Administrator Actions During Computer-Delivered Testlets (*n* = 341)

Action	*n*	%
Supporting		
Read one or more screens aloud to the student	205	74.5
Clarified directions or expectations for the student	196	79.0
Navigated one or more screens for the student	140	57.1
Repeated question(s) before student responded	120	52.2
Neutral		
Used pointing or gestures to direct student attention or engagement	128	57.7
Used verbal prompts to direct the student's attention or engagement (e.g. ”look at this.”)	145	62.2
Asked the student to clarify or confirm one or more responses	60	29.4
Used materials or manipulatives during the administration process	65	32.0
Allowed student to take a break during the testlet	34	17.6
Repeated question(s) after student responded (gave a second trial at the same item)	32	16.6
Nonsupporting		
Physically guided the student to a response	8	4.2
Reduced the number of answer choices available to the student	5	2.7

Note. Respondents could select multiple responses to this question.

For DLM assessments, interaction with the system includes interaction with the assessment content as well as physical access to the testing device and platform. The fact that educators navigated one or more screens in 57% of the observations does not necessarily indicate the student was prevented from engaging with the assessment content as independently as possible. Depending on the student, test administrator navigation may either support or minimize students' independent, physical interaction with the assessment system. While not the same as interfering with students' interaction with the content of assessment, navigating for students who are able to do so independently conflicts with the assumption that students are able to interact with the system as intended. The observation protocol did not capture why the test administrator chose to navigate, and the reason was not always obvious.

A related assumption is that students are able to interact with the system as intended. Evidence for this assumption was gathered by observing students taking computer-delivered testlets, as shown in Table 9.8. Independent response selection was observed in 88% of the cases. Non-independent response selection may include allowable practices, such as test administrators entering responses for the student. The use of materials outside of Kite Student Portal was seen in 15% of the observations. Verbal prompts for navigation and response selection are strategies within the realm of allowable flexibility during test administration. These strategies, which are commonly used during direct instruction for students with the most significant cognitive disabilities, are used to maximize student engagement with the system and promote the type of student-item interaction needed for a construct-relevant response. However, they also indicate that students were not able to sustain independent interaction with the system throughout the entire testlet.

Table 9.8. Student Actions During Computer-Delivered Testlets (*n* = 341)

Action	*n*	%
Selected answers independently	249	87.7
Navigated screens independently	195	73.6
Navigated screens after verbal prompts	109	50.9
Selected answers after verbal prompts	102	47.4
Navigated screens after TA pointed or gestured	92	43.0
Used materials outside of Kite Student Portal to indicate responses to testlet items	30	15.1
Revisited one or more questions after verbal prompt(s)	25	12.9
Independently revisited a question after answering it	22	11.6
Skipped one or more items	9	4.7

Note: Respondents could select multiple responses to this question.

Another assumption in the validity argument is that students are able to respond to tasks irrespective of sensory, mobility, health, communication, or behavioral constraints. This assumption was evaluated by having observers note whether there was difficulty with accessibility supports (including lack of appropriate available supports) during observations of teacher-administered testlets. Of the 182 observations of teacher-administered testlets, observers noted difficulty in 18 cases (10%). For computer-delivered testlets, evidence to evaluate the assumption was collected by noting students indicating responses to items using varied response modes such as eye gaze (1%) and using manipulatives or materials outside of Kite (15%). Additional evidence for this assumption was gathered by observing whether students were able to complete testlets. Of the 523 test administration observations collected, students completed the testlet in 507 cases (97%).[13]

Another assumption underlying the validity argument is that test administrators enter student responses with fidelity. To record student responses with fidelity, test administrators needed to observe multiple modes of communication, such as verbal, gesture, and eye gaze. Table 9.9 summarizes students' response modes for teacher-administered testlets. The most frequently observed behavior was *gestured to indicate response to test administrator who selected answers*.

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

Table 9.9. Primary Response Mode for Teacher-Administered Testlets (n = 182)

Response mode	*n*	%
Gestured to indicate response to TA who selected answers	63	34.6
Verbally indicated response to TA who selected answers	45	24.7
Used computer/device to respond independently	30	16.5
Eye-gaze system indication to TA who selected answers	12	6.6
Used switch system to respond independently	3	1.6
No response	56	30.8

Note. Respondents could select multiple responses to this question.

Computer-delivered testlets provided another opportunity to confirm fidelity of response entry when test administrators entered responses on behalf of students. This support is recorded on the Personal Needs and Preferences Profile and is recommended for a variety of situations (e.g., students who have limited motor skills and cannot interact directly with the testing device even though they can cognitively interact with the onscreen content). Observers recorded whether the response entered by the test administrator matched the student's response. In 67 of 341 (20%) observations of computer-delivered testlets, the test administrator entered responses on the student's behalf. In 64 (96%) of those cases, observers indicated that the entered response matched the student's response, while the remaining three observers left the item blank.

9.2.3. *Interrater Agreement of Writing Sample Scoring*

All students are assessed on writing EEs as part of the ELA blueprint. Teachers administer writing testlets at two levels: emergent and conventional. Emergent testlets measure nodes at the Initial Precursor and Distal Precursor levels, while conventional testlets measure nodes at the Proximal Precursor, Target, and Successor levels. All writing testlets include items that require teachers to evaluate students' writing processes; some testlets also include items that require teachers to evaluate students' writing samples. Evaluation of students' writing samples does not use a high-inference process common in large-scale assessment, such as applying analytic or holistic rubrics. Instead, writing samples are evaluated for text features that are easily perceptible to a fluent reader and require little or no inference on the part of the rater (e.g., correct syntax, orthography). The test administrator is presented with an onscreen selected-response item and is instructed to choose the option(s) that best matches the student's writing sample. Only test administrators rate writing samples, and their item responses are used to determine students' mastery of linkage levels for writing and some language EEs on the ELA blueprint. The purpose of this study was to evaluate how reliably teachers rate students' writing samples. For a complete description of writing testlet design and scoring, including example items, see Chapter 3 of the *2016–2017 Technical Manual Update—Year-End Model* (DLM Consortium, 2017b).

The number of items that evaluate the writing sample per grade-level testlet is summarized in Table 9.10. Testlets included one to six items evaluating the sample, administered as either multiple-choice or multi-select multiple-choice items. Because each answer option could correspond to a unique linkage level and/or EE, writing items are dichotomously scored at the option level. Each item, which included four to nine answer options, was scored as a separate writing item. For this reason, writing items are referred to as writing tasks in the following sections, and the options were scored as individual items. The dichotomous option responses (i.e., each scored as an item) were the basis for the evaluation of interrater agreement.

Table 9.10. Number of Items That Evaluate the Writing Product per Testlet, by Grade

Grade	Emergent testlet	Conventional testlet
3	*	3
4	1	4
5	*	2
6	*	4
7	1	4
8	*	4
9	1	8
10	1	8
11	1	8
12	1	8

Note. Items varied slightly by blueprint model; the maximum number of items per testlet is reported here. * The testlet at this grade included only items evaluating the writing process, with no evaluation of the sample.

9.2.3.1. Recruitment

Recruitment for the evaluation of interrater agreement of writing samples included the submission of student writing samples and direct recruitment of teachers to serve as raters.

9.2.3.1.1. Samples

During the spring 2019 administration, district coordinators were asked to submit student writing samples. Requested submissions included papers that students used during testlet administration, copies of student writing samples, or printed photographs of student writing samples. To allow the sample to be matched with test administrator response data from the spring 2019 administration, each sample was submitted with limited information to enable matching to the observed educator ratings.

A total of 171 student writing samples were submitted from districts in eight states. In several grades, the emergent writing testlet does not include any tasks that evaluate the writing sample (as shown in Table 9.10); therefore, emergent samples submitted for these grades were not included in the interrater reliability analysis (e.g., grade 3 emergent writing samples). Additionally, writing samples that could not be matched with student data were excluded (e.g., student name or identifier was not provided). These exclusion criteria resulted in the assignment of 145 writing samples to raters for evaluation of interrater agreement.

9.2.3.1.2. Raters

Recruited teachers were required to have experience administering and rating DLM writing testlets to ensure they had already completed required training and were familiar with how to score the writing samples. In total 10 were selected to participate.

Raters had a range of teaching experience, as indicated in Table 9.11. Most had taught ELA and/or students with the most significant cognitive disabilities for at least six years. Furthermore, one rater (10%) reported experience as a DLM external reviewer.

Table 9.11. Raters' Teaching Experience (*N* = 10)

	1–5 years		6–10 years		> 10 years	
Teaching experience	*n*	%	*n*	%	*n*	%
English language arts	3	30.0	1	10.0	6	60.0
Students with significant cognitive disabilities	3	30.0	1	10.0	6	60.0

Demographic information was collected as part of the volunteer survey administered in Qualtrics and is summarized in Table 9.12. Participating raters were mostly female (80%), white (90%), and non-Hispanic/Latino (90%). Raters came from a variety of teaching settings.

Table 9.12. Raters' Demographic Information (*N* = 10)

Subgroup	*n*	%
Gender		
Female	8	80
Male	2	20
Race		
White	9	90
Native Hawaiian or Pacific Islander	1	10
Hispanic ethnicity		
Non-Hispanic/Latino	9	90
Hispanic/Latino	1	10
Teaching setting		
Suburb	4	40
Town	3	30
Rural	2	20
City	1	10

9.2.3.2. Sample Ratings

All ratings occurred during an on-site event. Raters were provided with PDF versions of student writing samples on secure jump drives, which they returned following completion of ratings. They were also provided a link to a Qualtrics survey that included the writing tasks corresponding to the grade and level (i.e., emerging or conventional) of the assigned writing sample. Raters submitted all ratings online.

Writing samples were assigned to raters in batches of 13–21, using a partially crossed matrix design to assign each sample to a total of three raters. Thus, teachers rated between 51 and 63 writing samples. Table 9.13 summarizes the number of samples that were rated at each grade and level.

Table 9.13. Student Writing Samples with Ratings, by Grade (N = 145)

	Number of writing samples		
Grade	**Emergent**	**Conventional**	**Total number of samples**
3	*	5	5
4	12	4	16
5	*	5	5
6	*	13	13
7	9	12	21
8	*	18	18
9	4	19	23
10	10	14	24
11	11	8	19
12	1	0	1
Total	*47*	*98*	*145*

* The testlet at this grade included only items evaluating the writing process, with no evaluation of the sample.

Ratings submitted in Qualtrics were combined with the original student data from spring 2019, when the writing sample was rated by the student's teacher, resulting in four ratings for each of the 145 student writing samples.

Because writing tasks included multiple response options, each of which could be associated with a unique node measuring different EE(s) and linkage levels, each answer option was dichotomously scored; therefore, a script was used to transform writing data for scoring purposes. For more details on the scoring procedure, see Chapter 3 of the *2016–2017 Technical Manual Update—Year-End Model* (DLM Consortium, 2017b). The script applied nested scoring rules (in instances where selection of the option reflecting the highest-level skill also indicates the student demonstrated lower-level skills, such as student writes a paragraph also encompasses student writes a sentence), and transformed the options to the level of scoring (i.e., treating each option as a dichotomously scored item). While additional steps occur to report EE mastery for summative reporting, the option-level dichotomous scores represent the finest grain size of scoring and were used to calculate interrater reliability. All options were included in the evaluation of agreement, including options not associated with a node or corresponding EE/linkage level (e.g., "Wrote marks or selected symbols other than letters").

9.2.3.3. Interrater Reliability

Because each writing sample was evaluated by multiple and different raters, interrater reliability was summarized by Fleiss's kappa and intraclass correlation (ICC) values. The purpose of Fleiss's kappa is to provide a measure of absolute agreement across two or more raters. Fleiss's kappa (Fleiss, 1981) is defined as

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{9.1}$$

where the denominator gives the degree of absolute agreement attainable above chance and the numerator gives the degree of absolute agreement actually achieved above chance.

The purpose of the ICC is to provide a means for measuring rater agreement and consistency. For interrater reliability studies, rater agreement is of most interest. For this study a one-way, random-effects model using the average kappa rating was selected because each writing sample was rated by a rater who was randomly selected from the pool of available raters. Using this model, only absolute agreement is measured by the ICC.

Interrater agreement results are presented in Table 9.14. To summarize global agreement across all student writing samples, teachers' original ratings (from spring 2019 operational administration) were compared against the additional three ratings. Results are also provided separately for emergent and conventional testlets.

Based on the guidelines specified by Cicchetti (1994), ICC agreement fell in the *excellent* range ($\geq .75$), and Fleiss's kappa fell in the *good* range ($.60 - .74$). Agreement was slightly higher for emergent testlets.

Table 9.14. Interrater Agreement for Writing Samples (*N* = 145)

Group	*n*	ICC	ICC lower bound	ICC upper bound	Fleiss's κ
Overall	145	.91	.90	.91	.71
EW	47	.91	.90	.93	.73
CW	98	.91	.90	.91	.71

Note. ICC = intraclass correlation; EW = emergent writing; CW = conventional writing.

The results presented here reflect an analysis of interrater agreement for teacher-administered writing testlets. Agreement values were consistent with the results from 2017–2018 overall and for the subset of conventional writing testlets testlets. Agreement values for emergent level writing testlets were slightly higher in 2019 compared to 2018 for emergent level writing testlets. The ICC was .87 in 2018, compared to .91 in 2019. Fleiss's κ was .63 in 2018, compared to .73 in 2019. This suggests an improvement in the agreement for those testlets in 2019.

Teacher-administered testlets measuring reading and mathematics were not included in the study. Also, although student writing samples were evaluated, the student writing process was not. Additional data collection related to teacher fidelity, including fidelity in teacher-administered testlets in each subject, is provided in the Test Administration Observations section of this chapter.

Submitted writing samples were assumed to be representative of the types of student writing samples created by the broader population. However, various factors may have influenced a district coordinator's selection of samples for inclusion and therefore the submitted samples may not be a truly random sampling of all products likely to be observed.

A discussion of next steps for refining the evaluation of interrater agreement for writing samples is included in Chapter 11 of this manual.

9.3. Evidence Based on Internal Structure

Analyses of an assessment's internal structure indicate the degree to which "relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA et al., 2014, p. 16). Given the heterogeneous nature of the DLM student population, statistical analyses can examine whether particular items function differently for specific subgroups (e.g., male versus female). Additional evidence based on internal structure is provided across the linkage levels that form the basis of reporting.

9.3.1. Evaluation of Item-Level Bias

Differential item functioning (DIF) addresses the challenge 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). DIF analyses can uncover internal inconsistency if particular items function differently in a systematic way for identifiable subgroups of students (AERA et al., 2014). While identification of DIF does not always indicate weakness in a test item, it can point to construct-irrelevant variance or unexpected multidimensionality, posing considerations for validity and fairness.

9.3.1.1. Method

DIF analyses for 2019 followed the same procedure used in previous years, including data from 2015–2016 through 2017–2018[14] to flag items for evidence of DIF. Items were selected for inclusion in the DIF analyses based on minimum sample-size requirements for the two gender subgroups: male and female. 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 previous years 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 include non-independent response options. See Chapter 3 of the *2016–2017 Technical Manual Update—Year-End Model* (DLM Consortium, 2017b) for more information on the process of scoring writing items.

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

Using the above criteria for inclusion, 3,090 (77%) items on single-EE testlets were selected. The number of items evaluated by grade level and subject ranged from 109 items in grade 8 ELA to 257 items in grade 8 ELA. Item sample sizes ranged from 253 to 17,720.

Of the 928 items that were not included in the DIF analysis, 768 (83%) had a focal group sample size of less than 100, 159 (17%) had an item p-value greater than .95, and 1 (0%) had a subgroup p-value greater than .97. Table 9.15 shows the number and percent of items that did not meet each inclusion

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

criteria, by subject and the linkage level the items assess. The majority of non-included items are from mathematics (*n* = 752; 81%), and fall in the Distal Precursor to Target linkage level. In ELA, items not included due to sample size generally come from the Distal Precursor and Proximal Precursor linkage levels, whereas items not included due to *p*-values tend to come from the Target and Successor linkage levels. In mathematics, most items not include come from the Distal Precursor, Proximal Precursor, Target linkage levels, across the inclusion criteria.

Table 9.15. Items Not Included in DIF Analysis, by Subject and Linkage Level

	Sample Size		Item Proportion Correct		Subgroup Proportion Correct	
Subject and Linkage Level	*n*	%	*n*	%	*n*	%
English language arts						
Initial Precursor	0	0.0	0	0.0	0	0.0
Distal Precursor	55	80.9	0	0.0	0	0.0
Proximal Precursor	10	14.7	13	12.0	0	0.0
Target	3	4.4	57	52.8	0	0.0
Successor	0	0.0	38	35.2	0	0.0
Mathematics						
Initial Precursor	3	0.4	0	0.0	0	0.0
Distal Precursor	176	25.1	6	11.8	0	0.0
Proximal Precursor	197	28.1	23	45.1	1	100.0
Target	276	39.4	13	25.5	0	0.0
Successor	48	6.9	9	17.6	0	0.0

For each item, logistic regression was used to predict the probability of a correct response, given group membership and performance in the subject. Specifically, the logistic regression equation for each item included a matching variable comprised of the student's total linkage levels mastered in the subject of the item and a group membership variable, with 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); the presence of non-uniform DIF indicates that the item functions differently because of the interaction between total linkage levels mastered and gender. When non-uniform DIF is present, the gender group with the highest probability of a correct response to the item differs along the range of total linkage levels mastered, thus one group is favored at the low end of the spectrum and the other group is favored at the high end.

Three logistic regression models were fitted for each item:

$$\mathrm{M}_0\text{: } \mathrm{logit}(\pi_i) = \beta_0 + \beta_1\mathrm{X} \tag{9.2}$$

$$\mathrm{M}_1\text{: } \mathrm{logit}(\pi_i) = \beta_0 + \beta_1\mathrm{X} + \beta_2 G \tag{9.3}$$

$$\mathrm{M}_2\text{: } \mathrm{logit}(\pi_i) = \beta_0 + \beta_1\mathrm{X} + \beta_2 G + \beta_3\mathrm{X}G; \tag{9.4}$$

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

Because of the number of items evaluated for DIF, Type I error rates were susceptible to inflation. The incorporation of an effect-size measure can be used to distinguish practical significance from statistical significance by providing a metric of the magnitude of the effect of adding 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_0 to M_1 or M_2, to account for the effect of the addition of the group and interaction terms to the equation. All effect-size values were reported using both the Zumbo and Thomas (1997) and Jodoin and Gierl (2001) indices for reflecting a negligible, moderate, or large effect. The Zumbo and Thomas thresholds for classifying DIF effect size are based on Cohen's (1992) guidelines for identifying a small, medium, or large effect. The thresholds for each level are .13 and .26; values less than .13 have a negligible effect, values between .13 and .26 have a moderate effect, and values of .26 or greater have a large effect. The Jodoin and Gierl thresholds are more stringent, with lower threshold values of .035 and .07 to distinguish between negligible, moderate, and large effects.

9.3.1.2. Results

9.3.1.2.1. Uniform DIF Model

A total of 516 items were flagged for evidence of uniform DIF when comparing M_0 to M_1. Table 9.16 summarizes the total number of items flagged for evidence of uniform DIF by subject and grade for each model. The percentage of items flagged for uniform DIF ranged from 8% to 22%.

Table 9.16. Items Flagged for Evidence of Uniform DIF

Grade	Items flagged (n)	Total items (N)	Items flagged (%)	Items with moderate or large effect size (n)
English language arts				
3	26	149	17.4	1
4	25	164	15.2	0
5	19	159	11.9	0
6	22	143	15.4	0
7	15	118	12.7	0
8	24	109	22.0	0
9	20	130	15.4	0
10	14	154	9.1	0
11	31	142	21.8	0
Mathematics				
3	34	165	20.6	0
4	43	198	21.7	0
5	42	209	20.1	0
6	48	228	21.1	0
7	34	192	17.7	1
8	39	201	19.4	0
9	32	257	12.5	0
10	15	195	7.7	0
11	33	177	18.6	0

Using the Zumbo and Thomas (1997) effect-size classification criteria, all but one item were found to have a negligible effect-size change after the gender term was added to the regression equation.

Using the Jodoin and Gierl (2001) effect-size classification criteria, all but two items were found to have a negligible effect-size change after the gender term was added to the regression equation.

Table 9.17 provides information about the flagged items with a non-negligible effect-size change after the addition of the gender term, as represented by a value of B (moderate) or C (large). The $\beta_2 G$ values in Table 9.17 indicate which group was favored on the item after accounting for total linkage levels mastered, with positive values indicating that the focal group (females) had a higher probability of success on the item. Females were favored on one item.

Table 9.17. Items Flagged for Uniform DIF With Moderate or Large Effect Size

Item ID	Grade	EE	χ^2	p-value	β_2G	R^2	Z&T*	J&G*
ELA								
64895	3	RI.3.2	15.69	<.01	1.20	.04	A	B
Math								
24493	7	7.NS.2.b	50.90	<.01	-0.36	.84	C	C

Note. EE = Essential Element; Z&T = Zumbo & Thomas; J&G = Jodoin & Gierl; ELA = English language arts. * Effect-size measure.

9.3.1.2.2. Combined Model

A total of 600 items were flagged for evidence of DIF when both the gender and interaction terms were included in the regression equation, as shown in equation (9.4). Table 9.18 summarizes the number of items flagged by subject and grade. The percentage of items flagged for each grade and subject ranged from 5% to 28%.

Table 9.18. Items Flagged for Evidence of DIF for the Combined Model

Grade	Items flagged (n)	Total items (N)	Items flagged (%)	Items with moderate or large effect size (n)
English language arts				
3	29	149	19.5	1
4	33	164	20.1	0
5	30	159	18.9	0
6	27	143	18.9	0
7	13	118	11.0	0
8	26	109	23.9	0
9	19	130	14.6	0
10	8	154	5.2	0
11	30	142	21.1	0
Mathematics				
3	47	165	28.5	0
4	55	198	27.8	0
5	48	209	23.0	0
6	51	228	22.4	1
7	44	192	22.9	1
8	47	201	23.4	1
9	37	257	14.4	0
10	18	195	9.2	1
11	38	177	21.5	0

Using the Zumbo and Thomas (1997) effect-size classification criteria, all but three item had 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, 2 items had a moderate change in effect size, 3 had a large change in effect size, and the remaining 595 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 non-negligible change in effect size is summarized in Table 9.19. There was one ELA item and one mathematics items that had a moderate change in effect-size values, as represented by a value of B. In addition, there were three mathematics items that had a large change in effect-size values, as represented by a value of C. A total of one item favored the female group at higher levels of ability and males at lower levels of ability (as indicated by a positive $\beta_3 XG$).

Table 9.19. Items Flagged for DIF With Moderate or Large Effect Size for the Combined Model

Item ID	Grade	EE	χ^2	p-value	$\beta_2 G$	R^2	$\beta_3 XG$	Z&T*	J&G*
ELA									
64895	3	RI.3.2	15.71	<.01	1.36	.04	-0.01	A	B
Math									
24167	6	6.SP.5	14.18	<.01	0.12	.79	0.02	C	C
24493	7	7.NS.2.b	51.31	<.01	-0.25	.84	-0.01	C	C
26909	8	8.F.1-3	10.75	<.01	0.65	.81	-0.03	C	C
27610	10	A-CED.1	10.47	.01	2.59	.04	-0.15	A	B

Note. EE = Essential Element; Z&T = Zumbo & Thomas; J&G = Jodoin & Gierl; ELA = English language arts. * Effect-size measure.

Appendix A includes plots labeled by the item ID, which display the best-fitting regression line for each gender group, with jitter plots representing the total linkage levels mastered for individuals in each gender group. Plots are included for the 2 items with non-negligible effects-size changes in the uniform DIF model (Table 9.17), as well as the 5 items with non-negligible effect-size changes in the combined model (Table 9.19).

9.3.1.3. Test Development Team Review of Flagged Items

The test development teams for each subject were provided with data files that listed all items flagged with a moderate or large effect size. To avoid biasing the review of items, these files did not indicate which group was favored.

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 Clark, Beitling, et al., 2016) were provided for the test development teams to consider as they reviewed the items. After reviewing a flagged item and considering its context in the testlet, including the ELA text or the engagement activity in mathematics, test development teams were asked to provide one of three decision codes for each item.

1. Accept: There is no evidence of bias favoring one group or the other. Leave item as is.
2. Minor revision: There is a 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 no allowable edit to correct the issue. The item is slated for retirement.

After review, all ELA items flagged with a moderate or large effect size were given a decision code of 1 by the test development teams. One mathematics item was given a decision code of 3 and retired, while the remaining mathematics items flagged with a moderate or large effect size were given a decision code of 1. No evidence could be found in any of the items with a decision code of 1 indicating the content favored one gender group over the other.

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.

9.3.2. *Internal Structure Within Linkage Levels*

Internal structure traditionally indicates the relationships among items measuring the construct of interest. However, for DLM assessments, the level of scoring is each linkage level, and all items measuring the linkage level are assumed to be fungible. Therefore, DLM assessments instead present evidence of internal structure across linkage levels, rather than across items. Further, traditional evidence, such as item-total correlations, are not presented because DLM assessment results consist of the set of mastered linkage levels, rather than a scaled score or raw total score.

Chapter 5 of this manual includes a summary of the parameters used to score the assessment, which includes the probability of a master providing a correct response to items measuring the linkage level and the probability of a non-master providing a correct response to items measuring the linkage level. Because a fungible model is used for scoring, these parameters are the same for all items measuring the linkage level. Chapter 5 also provides a description of the linkage level discrimination (i.e., the ability to differentiate between masters and non-masters).

When linkage levels perform as expected, masters should have a high probability of providing a correct response, and non-masters should have a low probability of providing a correct response. As indicated in Chapter 5 of this manual, for 1,192 (99%) linkage levels, masters had a greater than .5 chance of providing a correct response to items. Additionally, for 1,161 (96%) linkage levels, masters had a greater than .5 chance of providing a correct response, compared to only 4 (<1%) linkage levels where masters had a less than .4 chance of providing a correct response. Similarly, for 897 (74%) linkage levels, non-masters had a less than .5 chance of providing a correct response to items. For most linkage levels (n = 673; 56%) non-masters had a less than .4 chance of providing a correct response; however, for 131 (11%) linkage levels, non-masters had a greater than .6 chance of providing a correct response. Finally, 868 (72%) linkage levels had discrimination index of greater than .4, indicating that linkage levels are largely able to discriminate between masters and non-masters.

Chapter 3 of this manual includes additional evidence of internal consistency in the form of standardized difference figures. Standardized difference values are calculated to indicate how far from the linkage level mean each item's p-value falls. Across all linkage levels, 4,449 (97%) of items fell within two standard deviations of the mean for the linkage level.

These sources, combined with procedural evidence for developing fungible testlets at the linkage level, provide evidence of the consistency of measurement at the linkage levels. For more information on the development of fungible testlets, see the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016). In instances where linkage levels and the items measuring them do not perform as expected, test development teams review flags to ensure the content measures the construct as expected.

9.4. Evidence Based on Consequences of Testing

Validity evidence must include the evaluation of the overall "soundness of these proposed interpretations of test scores 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.

Consistent with previous years, evidence was collected in spring 2019 via teacher survey responses regarding teacher perceptions of assessment content.

9.4.1. *Teacher Perception of Assessment Content*

On the spring 2019 survey,[15] teachers were asked two questions about their perceptions of assessment content: whether the content measured important academic skills and knowledge and whether the content reflected high expectations. Table 9.20 summarizes teachers' responses. Teachers generally agreed or strongly agreed that content reflected high expectations for their students (86%) and measured important academic skills (75%).

While the majority of teachers agreed with these statements, 14-25% disagreed. DLM assessments represent a departure from the breadth of academic skills assessed by many states' previous alternate assessments. Given the short history of general curriculum access for this population and the tendency to prioritize the instruction of functional academic skills (Karvonen et al., 2011), teachers' responses may reflect awareness that DLM assessments contain challenging content. However, teachers were divided on its importance in the educational programs of students with the most significant cognitive disabilities.

Table 9.20. Teacher Perceptions of Assessment Content

	Strongly Disagree		**Disagree**		**Agree**		**Strongly Agree**		**Agree + Strongly Agree**	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
The content of the assessments measured important academic skills and knowledge for this student.	3,739	9.3	6,189	15.4	22,960	57.0	7,373	18.3	30,333	75.3
The content of the assessments reflected high expectations for this student.	1,879	4.7	3,864	9.6	23,361	58.3	10,965	27.4	34,326	85.7

9.5. Conclusion

This chapter presents additional studies as evidence for the overall validity argument for the DLM Alternate Assessment System. The studies are organized into categories where available (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 this manual, Chapter 11, references evidence presented through the technical

[15] Recruitment and sampling are described in Chapter 4 of this manual.

manual, including Chapter 9, and expands the discussion of the overall validity argument. Chapter 11 also provides areas for further inquiry and ongoing evaluation of the DLM Alternate Assessment System, building on the evidence presented in the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016) and the subsequent annual technical manual updates (DLM Consortium, 2017a, 2017b, 2018a), in support of the assessment's validity argument.

10. Training and Professional Development

Chapter 10 of the Dynamic Learning Maps® (DLM®) Alternate Assessment System *2015–2016 Technical Manual—Year-End Model* (DLM Consortium, 2017a) 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. This chapter presents the participation rates and evaluation results from 2018–2019 instructional professional development. This chapter also describes the professional development webinars held for teachers and staff and the updates made to the professional development system during 2018–2019. There were no updates to training in 2018–2019.

For a complete description of training and professional development for 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 10 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016).

10.1. Instructional Professional Development

The DLM professional development system includes approximately 50 modules, including 20 focused on English language arts (ELA) instruction, 28 focused on mathematics instruction, and 5 others that address individual education programs, the DLM claims and conceptual areas, Universal Design for Learning, DLM Essential Elements (EEs), and the Common Core State Standards. The complete list of module titles is included in Table 10.2. The modules are available in two formats, self-directed and facilitated, which are accessed at the DLM professional development website[16]. The professional development website was updated for the 2018–2019 administration to allow teachers to easily navigate the website. The redesigned website contains the following tabs: Exemplar Text Supports, Instructional Resources, Professional Development, FAQs, and Blog. Teachers are encouraged to explore the modules in the professional development section and explore the other resources available.

The self-directed modules were designed to meet the needs of all educators, especially those in rural and remote areas, offering educators just-in-time, on-demand training. The self-directed modules are available online via an open-access, interactive portal that combines videos, text, student work samples, and online learning activities to engage educators with a range of content, strategies, and supports. It also gives educators 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 to be used 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 be subject-matter experts in content related to academic instruction and about the population of students with the most significant cognitive disabilities, the facilitated training is delivered via video recorded by subject-matter experts instead. Facilitators are provided with an agenda, a detailed guide, handouts, and other supports required to enable a meaningful, face-to-face training. By definition, they are facilitating training developed and provided by members of the DLM professional development team.

[16] http://dlmpd.com

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 about the faculty who developed and delivered the training.

10.1.1. Professional Development Participation and Evaluation

As reported in Table 10.1, a total of 9,115 modules were completed in the self-directed format from September 1, 2018, to August 31, 2019. Since the first module was launched in the fall of 2012, a total of 120,840 modules have been completed. Data are not available for the number of educators who have completed the modules in the facilitated format, but several states (e.g., Iowa, Missouri, and West Virginia) use the facilitated modules extensively.

Table 10.1. Number of Self-Directed Modules Completed in 2018–2019 by Educators in DLM States and Other Localities (*N* = 9,115)

State	Self-directed modules completed
Kansas	1,788
Colorado	1,462
Wisconsin	1,121
Arkansas	976
Iowa	294
Missouri	271
Illinois	248
Utah	208
Oklahoma	207
Rhode Island	179
New York	158
New Jersey	132
Maryland	101
Delaware	33
West Virginia	16
Alaska	15
New Hampshire	14
North Dakota	9
Non-DLM states and other locations	1,883

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. Three questions asked about importance of content, whether new concepts were presented, and the utility of the module. Educators responded using a four-point scale ranging from *stongly disagree* to *strongly agree*. A fourth question asked whether educators planned to use what they learned, with the same response options. During the 2018–2019 year, educators completed the evaluation questions 87% of the time. The responses were consistently positive, as illustrated in Table 10.2. Across all modules approximately 81% of respondents either agreed or strongly agreed with each statement.

Table 10.2. Response Rates and Rate of *Agree* or *Strongly Agree* on 2018–2019 Self-Directed Module Evaluation Questions

Module	**Total modules completed (*n*)**	**Response rate**	**The module addressed content that is important for professionals working with SWSCDs. (%)**	**The module presented me with new ideas to improve my work with SWSCDs. (%)**	**Completing this module was worth my time and effort. (%)**	**I intend to apply what I learned in the module to my professional practice. (%)**
Algebraic Thinking	130	93.1	88.5	90.0	85.4	90.0
Basic Geometric Shapes	155	80.0	76.1	75.5	76.8	77.4
Beginning Communicators	540	84.8	83.0	80.7	81.1	81.7
Calculating Accurately with Addition	154	87.0	78.6	78.6	77.9	78.6
Calculating Accurately With Division	68	89.7	83.8	80.9	80.9	82.4
Calculating Accurately With Multiplication	79	88.6	84.8	84.8	84.8	83.5
Calculating Accurately With Subtraction	75	85.3	78.7	76.0	74.7	80.0
Common Core Overview	191	92.1	83.8	79.6	83.8	85.9
Composing and Decomposing Shapes and Area	86	93.0	87.2	88.4	90.7	88.4

Table 10.2. Response Rates and Rate of *Agree* or *Strongly Agree* on 2018–2019 Self-Directed Module Evaluation Questions *(continued)*

Module	**Total modules completed (*n*)**	**Response rate**	**The module addressed content that is important for professionals working with SWSCDs. (%)**	**The module presented me with new ideas to improve my work with SWSCDs. (%)**	**Completing this module was worth my time and effort. (%)**	**I intend to apply what I learned in the module to my professional practice. (%)**
Composing, Decomposing, and Comparing Numbers	217	92.6	88.5	87.6	88.5	90.3
Core Vocabulary and Communication	288	84.0	79.2	78.1	77.1	77.8
Counting and Cardinality	270	90.4	87.4	85.6	84.1	85.2
DLM Claims and Conceptual Areas	144	93.8	88.2	86.1	86.8	87.5
DLM Essential Elements Overview	593	87.4	81.1	77.9	78.1	79.4
DR-TA and Other Text Comprehension Approaches	135	89.6	82.2	82.2	80.7	82.2
Effective Instruction in Mathematics	218	93.6	89.9	86.2	86.7	89.0
Emergent Writing	445	84.5	80.4	78.7	78.9	78.2

Table 10.2. Response Rates and Rate of *Agree* or *Strongly Agree* on 2018–2019 Self-Directed Module Evaluation Questions *(continued)*

Module	**Total modules completed (*n*)**	**Response rate**	**The module addressed content that is important for professionals working with SWSCDs. (%)**	**The module presented me with new ideas to improve my work with SWSCDs. (%)**	**Completing this module was worth my time and effort. (%)**	**I intend to apply what I learned in the module to my professional practice. (%)**
Exponents and Probability	22	100.0	90.9	95.5	86.4	95.5
Forms of Number	105	69.5	61.0	57.1	55.2	60.0
Fraction Concepts and Models Part I	37	83.8	75.7	73.0	73.0	75.7
Fraction Concepts and Models Part II	22	95.5	86.4	90.9	86.4	95.5
Functions and Rate	17	100.0	88.2	88.2	88.2	88.2
Generating Purposes for Reading	143	86.7	81.8	81.1	77.6	79.7
IEPs Linked to DLM Essential Elements	272	84.6	73.2	72.1	70.6	73.5
Measuring and Comparing Lengths	136	92.6	87.5	89.0	89.0	90.4
Organizing and Using Data to Answer Questions	51	96.1	90.2	90.2	90.2	84.3

Table 10.2. Response Rates and Rate of *Agree* or *Strongly Agree* on 2018–2019 Self-Directed Module Evaluation Questions *(continued)*

Module	**Total modules completed (*n*)**	**Response rate**	**The module addressed content that is important for professionals working with SWSCDs. (%)**	**The module presented me with new ideas to improve my work with SWSCDs. (%)**	**Completing this module was worth my time and effort. (%)**	**I intend to apply what I learned in the module to my professional practice. (%)**
Patterns and Sequences	39	92.3	87.2	84.6	87.2	87.2
Perimeter, Volume, and Mass	59	91.5	86.4	88.1	86.4	88.1
Place Value	72	75.0	68.1	68.1	66.7	68.1
Predictable Chart Writing	105	83.8	80.0	81.0	81.0	81.0
Principles of Effective Instruction ELA	237	91.1	82.3	82.7	81.4	83.5
Properties of Lines and Angles	26	88.5	80.8	80.8	84.6	84.6
Shared Reading	666	87.8	83.6	81.7	80.3	83.0
Speaking and Listening	177	78.5	78.0	77.4	76.8	77.4
Standards of Mathematical Practice	3	66.7	66.7	66.7	66.7	66.7
Strategies and Formats for Presenting Ideas	141	78.7	75.9	75.9	75.2	76.6

Table 10.2. Response Rates and Rate of *Agree* or *Strongly Agree* on 2018–2019 Self-Directed Module Evaluation Questions *(continued)*

Module	**Total modules completed (*n*)**	**Response rate**	**The module addressed content that is important for professionals working with SWSCDs. (%)**	**The module presented me with new ideas to improve my work with SWSCDs. (%)**	**Completing this module was worth my time and effort. (%)**	**I intend to apply what I learned in the module to my professional practice. (%)**
Supporting Participation in Discussions	118	74.6	71.2	70.3	70.3	68.6
Symbols	144	85.4	85.4	84.0	84.0	85.4
Teaching Text Comprehension: Anchor-Read-Apply	295	81.4	75.3	74.9	74.2	74.2
The Power of Ten-Frames	92	92.4	88.0	87.0	88.0	89.1
Time and Money	80	96.2	91.2	92.5	91.2	91.2
Unitizing	35	91.4	74.3	71.4	68.6	80.0
Units and Operations	19	89.5	89.5	84.2	84.2	84.2
Universal Design for Learning	340	94.1	92.1	88.8	87.4	90.0
Who are Students with Significant Cognitive Disabilities?	1,115	91.9	89.3	83.3	84.6	87.4
Writing Information Texts	51	82.4	80.4	80.4	80.4	78.4

Table 10.2. Response Rates and Rate of *Agree* or *Strongly Agree* on 2018–2019 Self-Directed Module Evaluation Questions *(continued)*

Module	**Total modules completed (*n*)**	**Response rate**	**The module addressed content that is important for professionals working with SWSCDs. (%)**	**The module presented me with new ideas to improve my work with SWSCDs. (%)**	**Completing this module was worth my time and effort. (%)**	**I intend to apply what I learned in the module to my professional practice. (%)**
Writing With Alternate Pencils	292	84.2	82.2	80.8	80.5	80.8
Writing: Getting Started in Narrative Writing	59	81.4	76.3	74.6	76.3	74.6
Writing: Getting Started Writing Arguments	35	82.9	68.6	71.4	71.4	71.4
Writing: Production and Distribution	41	85.4	80.5	80.5	78.0	80.5
Writing: Research and Range of Writing	110	87.3	85.5	86.4	85.5	85.5
Writing: Text Types and Purposes	171	73.1	69.6	69.0	69.0	68.4
Total	*9,115*	*87.3*	*82.8*	*80.9*	*80.6*	*82.1*

Note. SWSCDs = students with significant cognitive disabilities.

In addition to the modules, the DLM instructional professional development system has a variety of other resources and supports. These include DLM EE unpacking documents; extended descriptions of the Initial Precursor and Distal Precursor linkage levels and how they relate to grade-level EEs; 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.

Finally, the DLM instructional professional development system includes webinars for teachers to get a review of modules and have discussions about instructional practices around featured modules. During the 2018–2019 year, The Center for Literacy and Disability Studies at the University of North Carolina-Chapel Hill, an ATLAS partner, held webinars for teachers and staff who work with students with significant cognitive disabilities. A total of seven webinars were held. Webinar topics include Beginning Communicator Symbols, Writing with Alternate Pencils, DLM Core Vocabulary, Writing, DLM Familiar Texts, Measurement, Counting and Cardinality and the Power of Tens Frames, Composing and Decomposing and Comparing Numbers, and Composition and Decomposition of shapes and area. Teachers were given the professional development module to review prior to the webinar, in order to drive the conversation and have any questions answered about teaching each of the topic. Each webinar was recorded and posted on the DLM site under professional development. Additionally, there is a DLM Instructional Support Facebook page where teachers can post questions and ideas related to instruction. The DLM professional development team at the University of North Carolina at Chapel Hill continues to work to seed and support the development of this online community and is working to identify new ways to attract more active users.

11. 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. 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 after a full year of instruction.

The DLM system completed its fifth operational administration year in 2018–2019. This technical manual update provides updated evidence from the 2018–2019 year intended to evaluate 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 11.1. Evidence summarized in this manual builds on the original evidence included in the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016) and in subsequent years (DLM Consortium, 2017a, 2017b, 2018a). Together, the documents summarize the validity evidence collected to date.

Table 11.1. Review of Technical Manual Update Contents

Chapter	Contents
1	Provides an overview of information updated for the 2018–2019 year
2	Not updated for 2018–2019
3, 4, 10	Provides evidence collected during 2018–2019 of test content development and administration, including field-test information, teacher-survey results, and professional development module use
5	Describes the statistical model used to produce results based on student responses, along with a summary of item parameters
6	Not updated for 2018–2019
7, 8	Describes results and analyses from the fifth operational administration, evaluating how students performed on the assessment, the distributions of those results, including aggregated and disaggregated results, and analysis of the consistency of student responses
9	Provides additional studies from 2018–2019 focused on specific topics related to validity

This chapter reviews the evidence provided in this technical manual update and discusses future research studies as part of ongoing and iterative processes of program responsiveness, validation, and evaluation.

11.1. Validity Evidence Summary

The accumulated evidence available by the end of the 2018–2019 year provides additional support for the validity argument. Four interpretation and use claims are summarized in Table 11.2. Each claim is addressed by evidence in one or more of the sources of validity evidence defined in the *Standards for Educational and Psychological Testing* (AERA et al., 2014). While many sources of evidence contribute to multiple propositions, Table 11.2 lists the primary associations. For example, Proposition 4 is indirectly supported by content-related evidence described for Propositions 1 through 3. Table 11.3 shows the titles and sections for the chapters cited in Table 11.2.

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

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

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

Table 11.3. Evidence Sources Cited in Table 11.2

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

11.2. Continuous Improvement

11.2.1. Operational Assessment

As noted previously in this manual, 2018–2019 was the fifth year the DLM Alternate Assessment System was operational. While the 2018–2019 assessments were carried out in a manner that supports the validity of inferences made from results for the intended purposes, the DLM 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 2019–2020. This section describes significant changes from the fourth to fifth year of operational administration, as well as examples of improvements to be made during the 2019–2020 year.

Overall, there were no significant changes to the learning map models, item-writing procedures, item flagging outcomes, the modeling procedure used to calibrate and score assessments, or the method for quantifying the reliability of results from previous years to 2018–2019.

Based on an ongoing effort to improve Kite® system functionality, several changes were implemented during 2018–2019. Educator Portal was updated to enhance the usability of the online platform. A teacher cadre provided feedback on the Instructional Tools Interface to support future design work and system updates. Additionally, a new system was implemented for the collection of test administration observations, resulting in a more robust sample of observations for evaluating the administration of testlets. Writing sample collection was also updated to accept sample uploads virtually rather than requiring participants to mail in hard copies.

The validity evidence collected in 2018–2019 expands upon the data compiled in the first four operational years for four of the critical sources of evidence as described in *Standards for Educational and Psychological Testing* (AERA et al., 2014): evidence based on test content, internal structure, response process, and consequences of testing. Specifically, analysis of opportunity to learn contributed to the evidence collected based on test content. Teacher-survey responses on test administration further contributed to the body of evidence collected based on response process, in addition to test-administration observations and evaluation of interrater agreement on the scoring of student writing samples. Evaluation of item-level bias via differential item functioning analysis, along with item-pool statistics and model parameters, provided additional evidence collected based on internal structure. Teacher-sruvey responses also provided evidence based on consequences of testing. Studies planned for 2019–2020 to provide additional validity evidence are summarized in the following section.

11.2.2. Future Research

The continuous improvement process also leads to future directions for research to inform and improve the DLM Alternate Assessment System in 2019–2020 and beyond. The manual identifies some areas for further investigation.

Additional Kite enhancements will be implemented for the 2019–2020 year, including updating the current Instructional Tools Interface used to create instructional plans and assign instructionally embedded testlets.

DLM staff members are planning several studies for spring 2020 to collect data from teachers in the DLM Consortium states. Additional updates to the writing sample collection process are planned for 2019–2020 to further streamline the upload process with the intention of expanding the number of writing samples collected. The teacher survey will include a new spiraled block to collection additional information on relation to other variables, whereby teacher ratings of student mastery will be correlated with model-derived mastery. Teacher-survey data collection will also continue during spring 2020 to obtain the fourth year of data for longitudinal survey items as further validity evidence. State partners will continue to collaborate with additional data collection as needed. Additionally, teacher feedback will be solicited to identify any remaining accessibility gaps, based on the present teacher survey findings that a small percentage of students may not be able to fully access assessment content.

In addition to data collected from students and teachers in the DLM Consortium, a research trajectory is underway to improve the model used to score DLM assessments. This includes the evaluation of a Bayesian estimation approach to improve on the current linkage-level scoring model and evaluation of item-level model misfit. Furthermore, research is underway to potentially support making inferences over tested linkage levels, with the ultimate goal of supporting node-based estimation. This research agenda is being guided by a modeling subcommittee of DLM Technical Advisory Committee (TAC) members. Additional research will also be conducted to further evaluate the calculation of reliability.

Other ongoing operational research is also anticipated to grow as more data become available. For example, differential item functioning analyses will be expanded to include evaluating items across ethnicity subgroups.

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.

12. References

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.

Camilli, G., & Shepard, L. A. (1994). *Method for Identifying Biased Test Items* (4th). Thousand Oaks, CA, Sage.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290. https://doi.org/10.1037/1040-3590.6.4.284[17]

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

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* (tech. rep. No. 15-04). University of Kansas, Center for Educational Testing and Evaluation. Lawrence, KS.

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

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

Dynamic Learning Maps Consortium. (2017a). *2015–2016 Technical Manual Update—Year-End Model* (tech. rep.). University of Kansas, Center for Educational Testing and Evaluation. Lawrence, KS.

Dynamic Learning Maps Consortium. (2017b). *2016–2017 Technical Manual Update—Year-End Model* (tech. rep.). University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS). Lawrence, KS.

Dynamic Learning Maps Consortium. (2017c). *Accessibility Manual for the Dynamic Learning Maps Alternate Assessment, 2017–2018* (tech. rep.). University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS). Lawrence, KS.

Dynamic Learning Maps Consortium. (2018a). *2017–2018 Technical Manual Update—Year-End Model* (tech. rep.). University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS). Lawrence, KS.

Dynamic Learning Maps Consortium. (2018b). *Educator Portal User Guide* (tech. rep.). University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS). Lawrence, KS.

Dynamic Learning Maps Consortium. (2018c). *Test Administration Manual 2018–2019* (tech. rep.). University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS). Lawrence, KS.

Dynamic Learning Maps Consortium. (2019). *2018–2019 Technical Manual Update—Science* (tech. rep.). University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS). Lawrence, KS.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions* (2nd). New York, NY, John Wiley.

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, United Kingdom, Cambridge University Press.

[17] https://doi.org/10.1037/1040-3590.6.4.284

Gelman, A., Meng, X., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733–807.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power raters 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 [Retrieved from ERIC database].

Nash, B., Clark, A., & Karvonen, M. (2015). *First contact: A census report on the characteristics of students eligible to take alternate assessments* (tech. rep. No. 16-01). University of Kansas, Center for Educational Testing and Evaluation. Lawrence, KS.

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. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*(2), 251–275. https://doi.org/10.1007/s00357-013-9129-4[18]

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4[19]

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

[18] https://doi.org/10.1007/s00357-013-9129-4

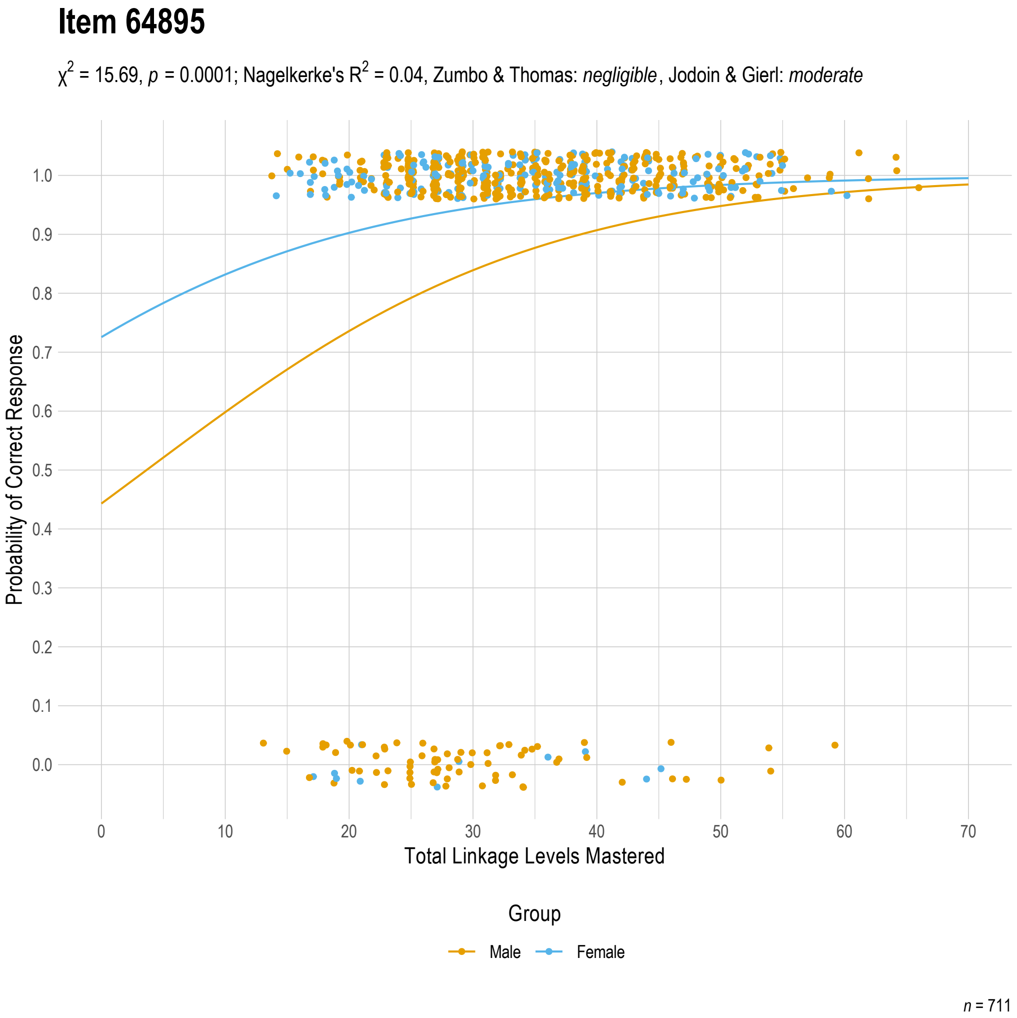
[19] https://doi.org/10.1007/s11222-016-9696-4

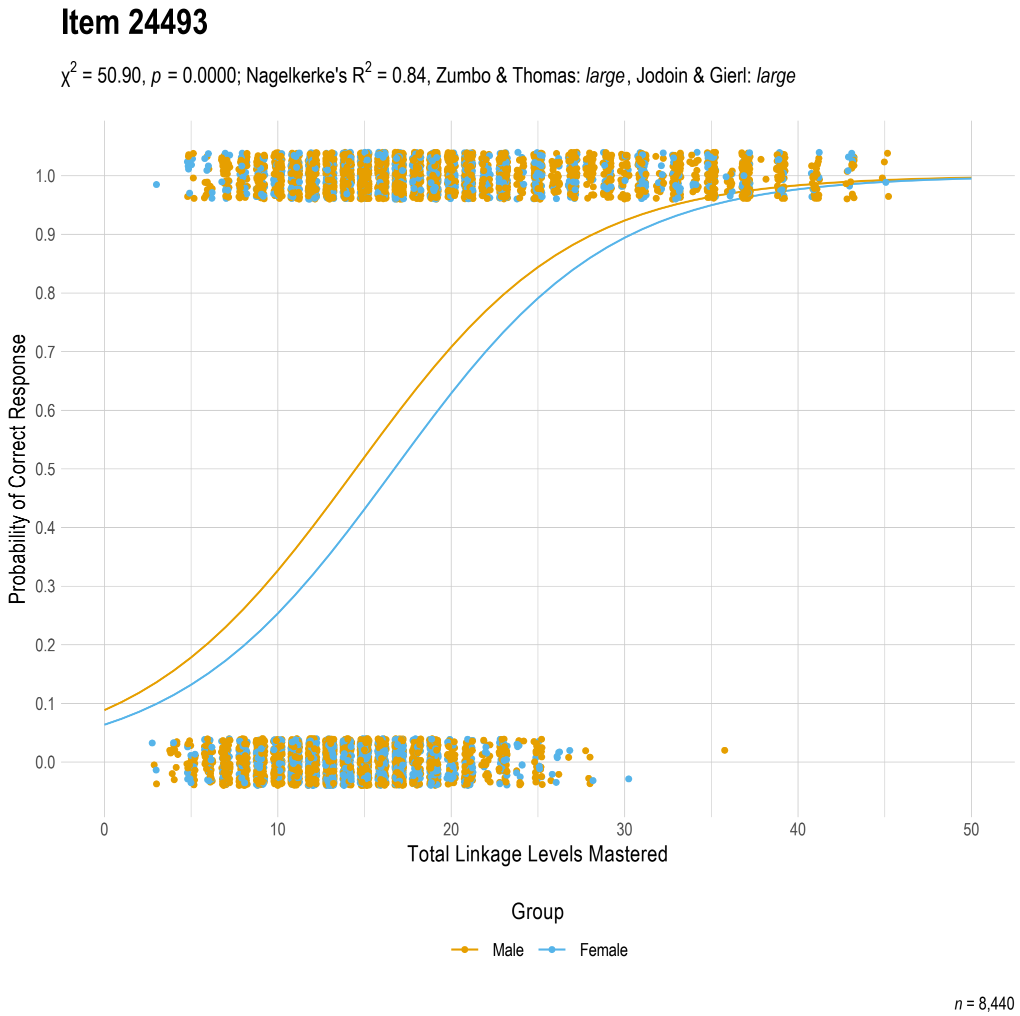
A. Differential Item Functioning Plots

The plots in this section display the best-fitting regression line for each gender group, with jittered plots representing the total linkage levels mastered for individuals in each gender group. Plots are labeled with the item ID, and only items with non-negligible effect-size changes are included. The results from the uniform and combined logistic regression models are presented separately. For a full description of the analysis, see the Evaluation of Item-Level Bias section.

A.1. Uniform Model

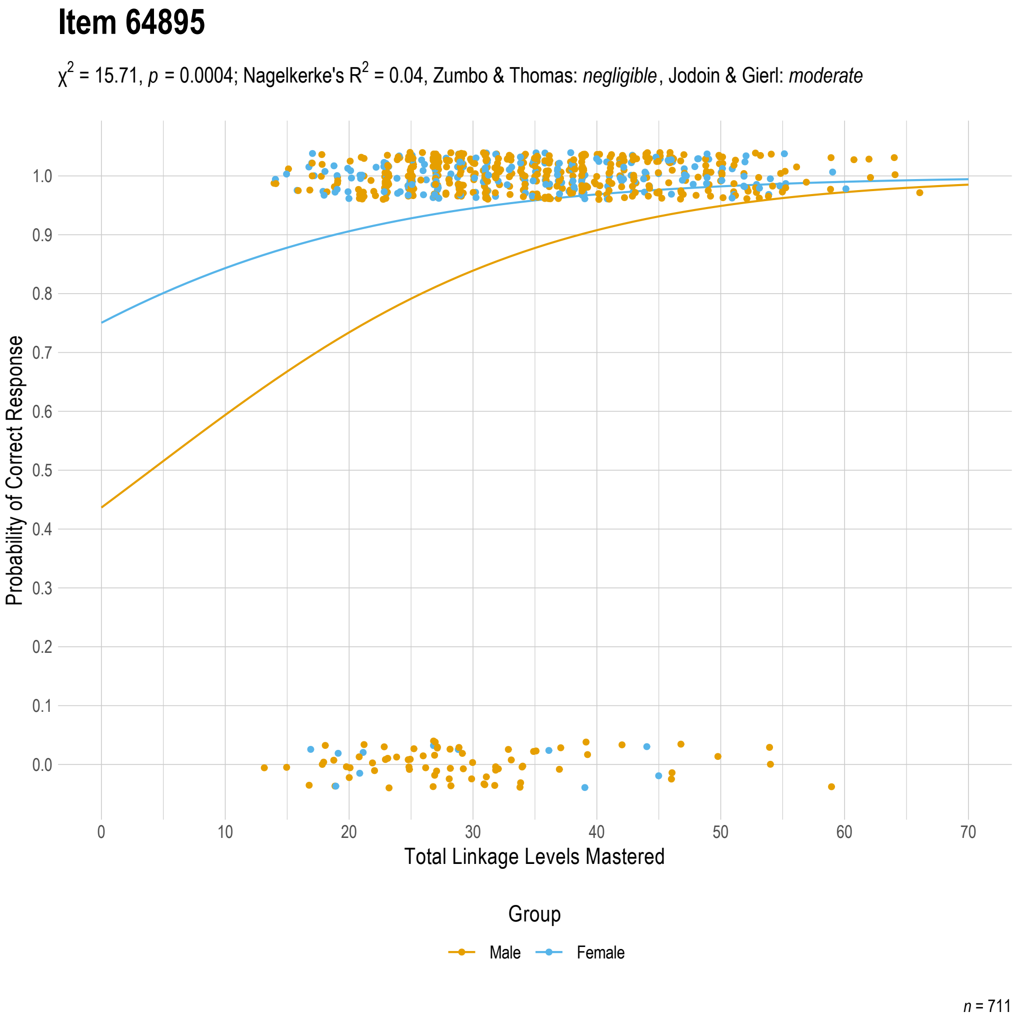
These plots show items that had a non-negligible effect-size change when comparing equation (9.3) to equation (9.2). In this model, the probability of a correct response was modeled as a function of ability and gender.





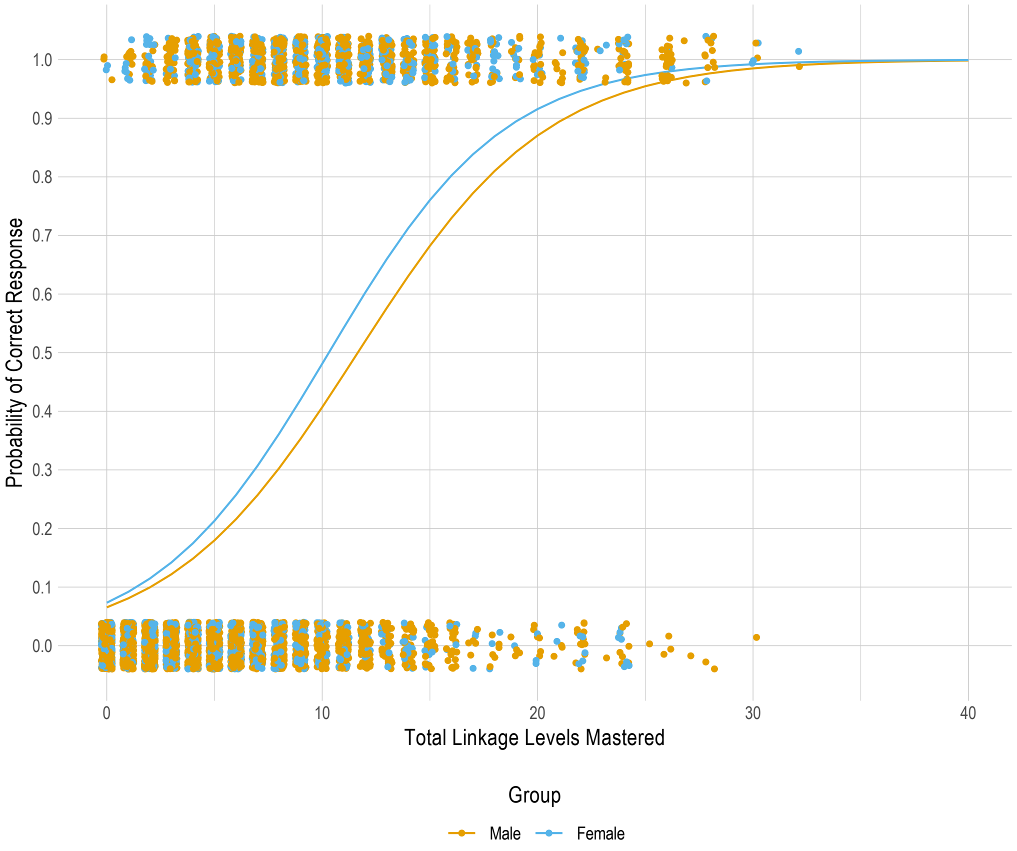
A.2. Combined Model

These plots show items that had a non-negligible effect-size change when comparing equation (9.4) to equation (9.2). In this model, the probability of a correct response was modeled as a function of ability, gender, and their interaction.


Item 64895
χ^2 = 15.71, *p* = 0.0004; Nagelkerke's R^2 = 0.04, Zumbo & Thomas: *negligible*, Jodoin & Gierl: *moderate*
Probability of Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
70
Total Linkage Levels Mastered
Group
Male
Female
n = 711

Item 24167

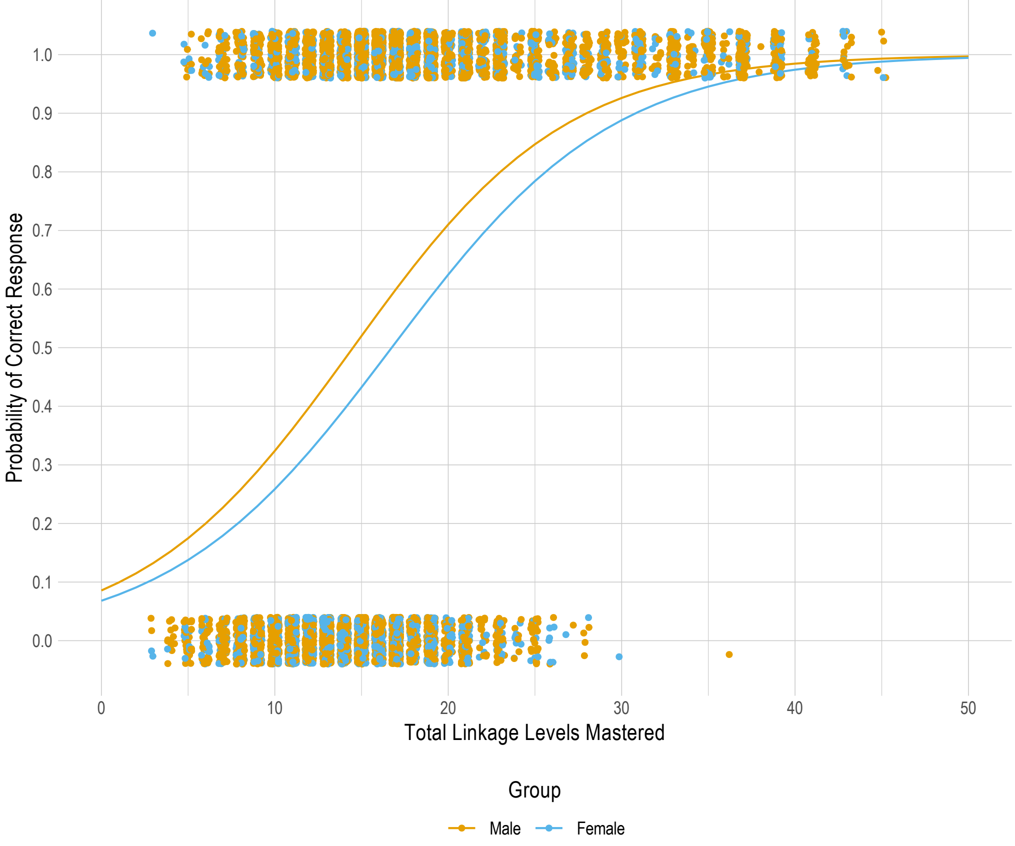
χ^2 = 14.18, *p* = 0.0008; Nagelkerke's R^2 = 0.79, Zumbo & Thomas: *large*, Jodoin & Gierl: *large*



n = 6,103

Item 24493

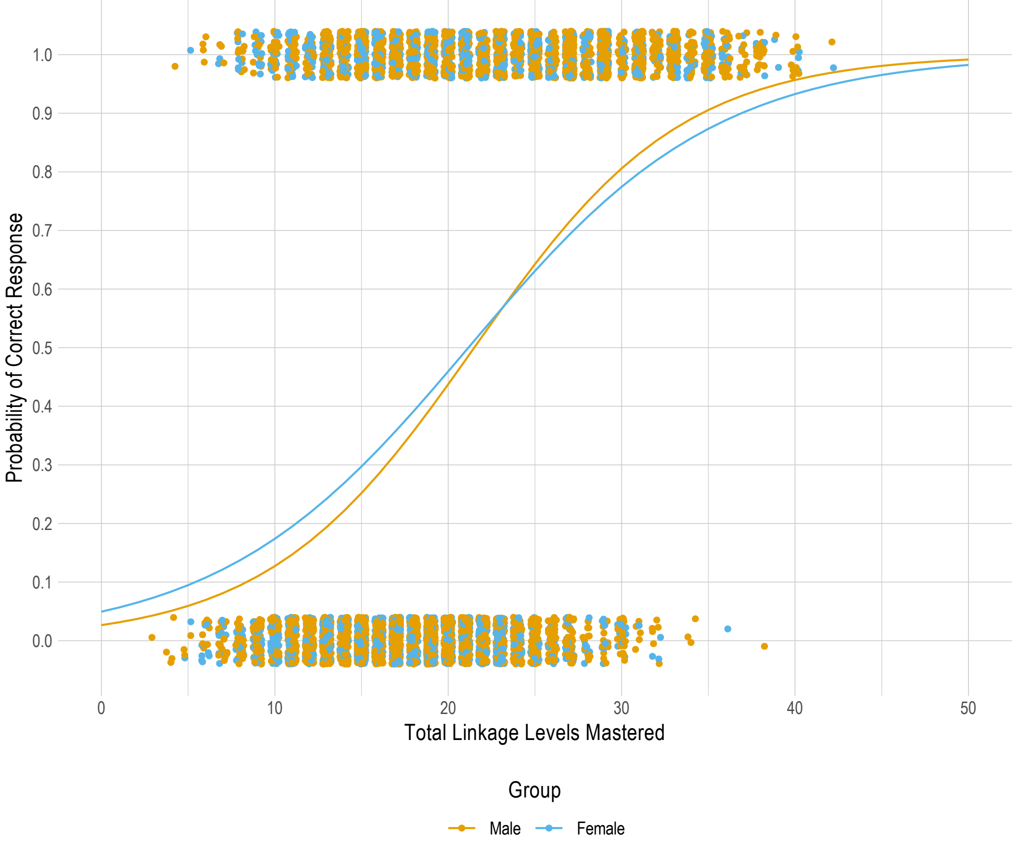
χ^2 = 51.31, *p* = 0.0000; Nagelkerke's R^2 = 0.84, Zumbo & Thomas: *large*, Jodoin & Gierl: *large*



n = 8,440

Item 26909

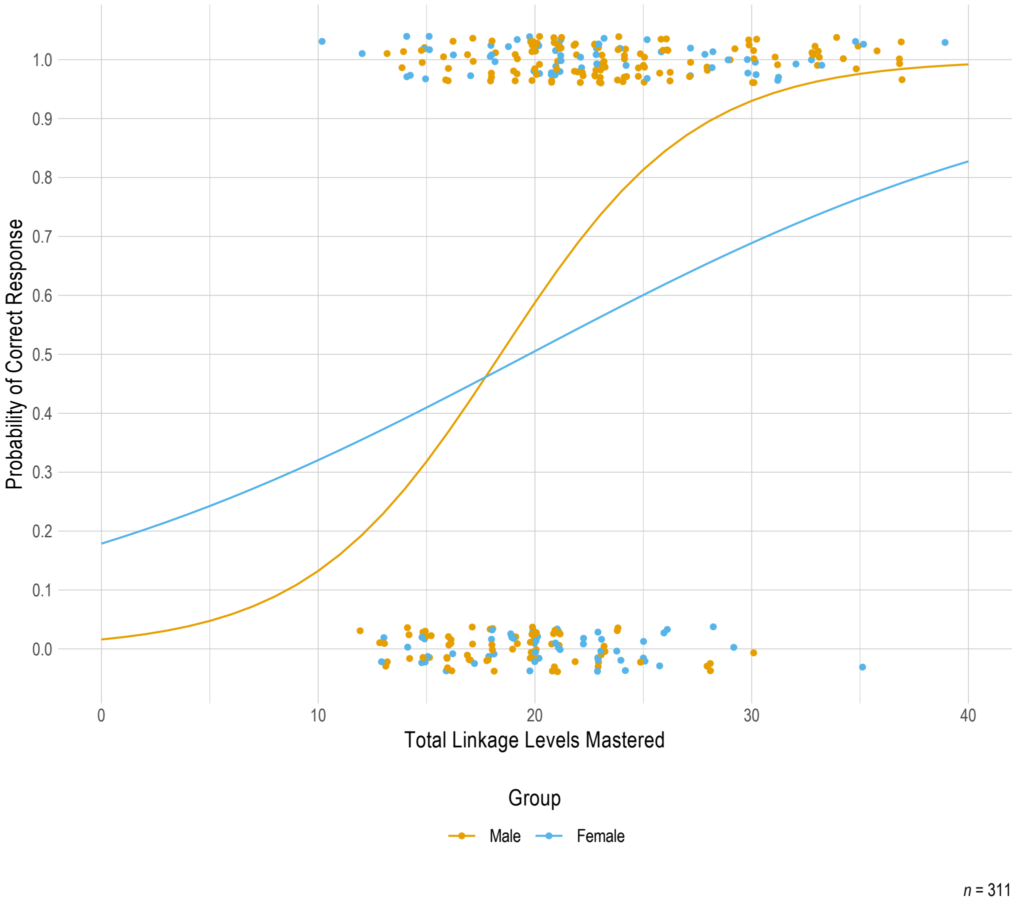
χ^2 = 10.75, *p* = 0.0046; Nagelkerke's R^2 = 0.81, Zumbo & Thomas: *large*, Jodoin & Gierl: *large*



n = 7,082

Item 27610

χ^2 = 10.47, *p* = 0.0053; Nagelkerke's R^2 = 0.04, Zumbo & Thomas: *negligible*, Jodoin & Gierl: *moderate*


Probability of Correct Response
Total Linkage Levels Mastered
Group
Male
Female
n = 311