DYNAMIC®
LEARNING MAPS

2017–2018 Technical Manual Update

Year-End Model

December 2018

Dynamic Learning Maps Consortium. (2018, December). *2017–2018 Technical Manual Update—Year-End Model*. Lawrence, KS: University of Kansas, Center for 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.

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

The authors wish to acknowledge Zachary Hopper, Teri Millar, Chelsea Nehler, Noelle Pablo, Courtney Rodgers, 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 2017–2018 operational year include:

Russell Almond, Ph.D., *Florida State University*
Greg Camilli, Ph.D., *Rutgers University*
Karla Egan, Ph.D., *Independent Consultant*
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 2017–2018 academic year was the fourth 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 2017–2018 (see Dynamic Learning Maps Consortium [DLM Consortium], 2018).

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 2017–2018 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 2017–2018, DLM assessments were administered to students in 17 states and one Bureau of Indian Education school: Alaska, Colorado, Delaware, Illinois, Iowa, Kansas, Maryland, Miccosukee Indian School, Missouri, New Hampshire, New Jersey, New York, North Dakota, Oklahoma, Rhode Island, Utah, West Virginia, and Wisconsin.

One DLM Consortium partner, Maryland, did not administer operational assessments in ELA or mathematics in 2017–2018.

In 2017–2018, the Center for 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

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

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 (TAC).

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® Client. 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 2017–2018, the states participating in the integrated model included 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 2017–2018, 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 2017–2018 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 2017–2018 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 2017–2018; 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 2017–2018 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 2017–2018 year. The chapter provides 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 inclues a summary of 2017–2018 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 2017–2018; 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 2017–2018 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 2017–2018 administration.

Chapter 8 summarizes reliability evidence for the 2017–2018 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 2017–2018 administration not covered in previous chapters. The chapter provides study results for the five critical sources of evidence: test content, internal structure, response process, relation to other variables, and consequences of testing.

Chapter 10 describes the training and professional development offered across the DLM Consortium in 2017–2018, including participation rates and evaluation results.

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 2017–2018 academic year. The first portion of the chapter provides an overview of 2017–2018 item writers' characteristics. The next portion of the chapter describes the pool of operational and field test testlets administered during spring 2018.

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 2017–2018 year. For a complete summary of item and testlet development procedures that began in 2014–2015 and were implemented in 2015–2016, see Chapter 3 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016).

For the 2017–2018 year, only a limited number of items were written to replenish the pool. The item writing process for 2017–2018 began with an on-site event in January 2018. 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 10 multi-EE testlets were written for English language arts (ELA), and 2 were written for mathematics.

3.1.1. 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, 28 item writers contributed to writing testlets for the 2017–2018 year, including 13 for mathematics and 15 for ELA. The median and range of years of teaching experience in four areas the item writers had is shown in Table 3.1. Item writers for ELA testlets had a higher median years of experience than item writers for mathematics testlets in all areas except for special education. The median years of experience was at least 10 years for item writers of both ELA and mathematics testlets in pre-K–12, as well as the ELA and mathematics subject areas.

Table 3.1. Item Writers' Years of Teaching Experience

	English language arts		Mathematics	
Area	**Median**	**Range**	**Median**	**Range**
Pre-K–12	19	5-29	10	5-32
ELA	19	5-27	10	5-32
Mathematics	18	4-22	11	5-32
Special Education	7	2-20	9	5-32

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

All item writers held at least a Bachelor's degree. The distribution and types of degrees held by item writers are shown in Table 3.2 and Table 3.3. All item writers held at least a Bachelor's degree, with the most common field of study being education (*n* = 14; 50%), followed by special education (*n* = 8; 29%). A majority (*n* = 22; 79%) also held a Master's degree, for which the most common field of study was special education (*n* = 13; 59%).

Table 3.2. Item Writers' Level of Degree

	English language arts		Mathematics	
Degree	***n***	**%**	***n***	**%**
Bachelor's	15	100	13	100
Master's	13	87	9	62
Other	1	7	0	0

Table 3.3. Item Writers' Degree Type

	English language arts	Mathematics
Degree	*n*	*n*
Bachelor's Degree		
Education	10	4
Content Specific	0	0
Special Education	3	5
Other	2	4
Master's Degree		
Education	0	0
Content Specific	2	0
Special Education	7	6
Other	4	3

Most item writers had experience working with students with disabilities, as summarized in Table 3.4. Teachers collectively had the most experience working with students with a severe cognitive disability, other health impairment, or emotional disability.

Table 3.4. Item Writers' Experience with Disability Categories

	English language arts		Mathematics	
Diability Category	*n*	%	*n*	%
Blind/Low Vision	6	40	5	38
Deaf/Hard of Hearing	6	40	6	46
Emotional Disability	10	67	10	77
Mild Cognitive Disability	8	53	7	54
Multiple Disabilities	8	53	10	77
Orthopedic Impairment	5	33	6	46
Other Health Impairment	12	80	10	77
Severe Cognitive Disability	12	80	10	77
Specific Learning Disability	7	47	8	62
Speech Impairment	8	53	9	69
Traumatic Brain Injury	2	13	5	38
None of the above	2	13	2	15

Of the items writers, 79% had experience administering an Alternate Assessment of Alternate Achievement Standards (AA-AAS) prior to their work on the DLM project, and 64% reported working with students eligible for AA-AAS at the time of the survey.

3.2. External Reviews

Due to the implementation of a new external review timeline, there were limited external review activities during the 2017–2018 year. Because of this, external review activities for recently developed testlets were scheduled for an on-site external review event during summer of 2018 and will be documented in the *2018–2019 Technical Manual Update—Year-End Model*.

3.3. Operational Assessment Items for Spring 2018

A total of 1,121,216 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 sprint 2018 based on the 2016–2017 operational pool and the promotion of testlets field-tested during 2016–2017 to the operational pool following their review. Table 3.5 summarizes the total number of operational testlets for spring 2018 for ELA and mathematics. There were 752 operational testlets available across grades and subjects. This total included 475 (135 mathematics, 340 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 3.5. 2018 Operational Testlets, by Subject ($N = 752$)

Grade	English language arts (*n*)	Mathematics (*n*)
3	53	37
4	53	46
5	49	41
6	41	41
7	39	38
8	31	41
9	36	48
10	39	44
11	30	45

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.

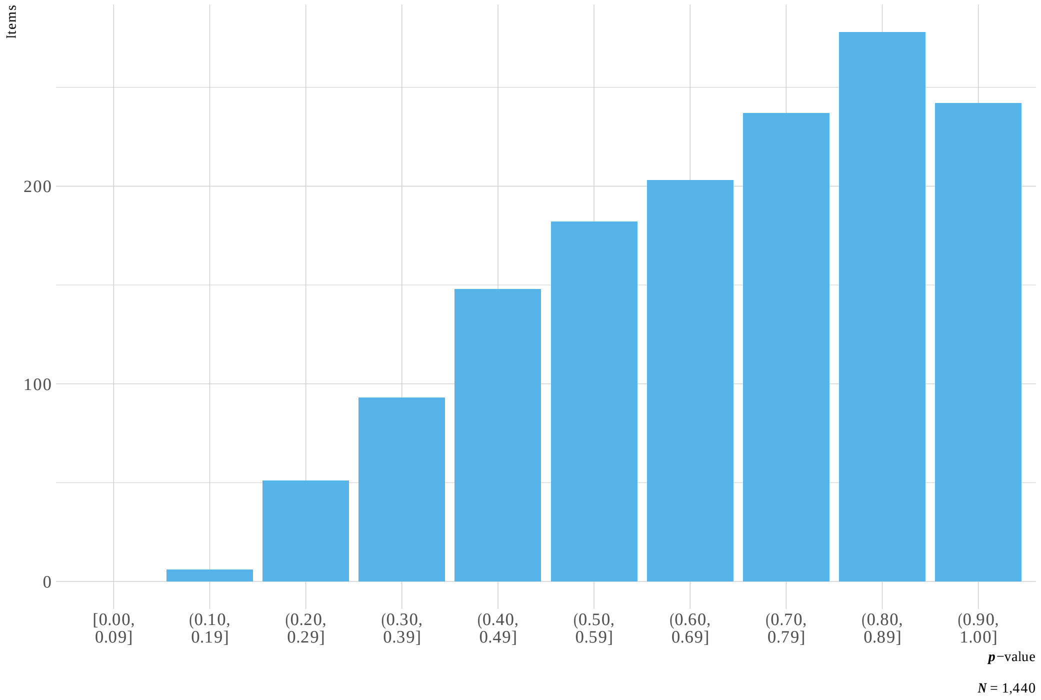


Figure 3.1. *p*-values for ELA 2018 operational items. *Note*. Items with a sample size of less than 20 were omitted.

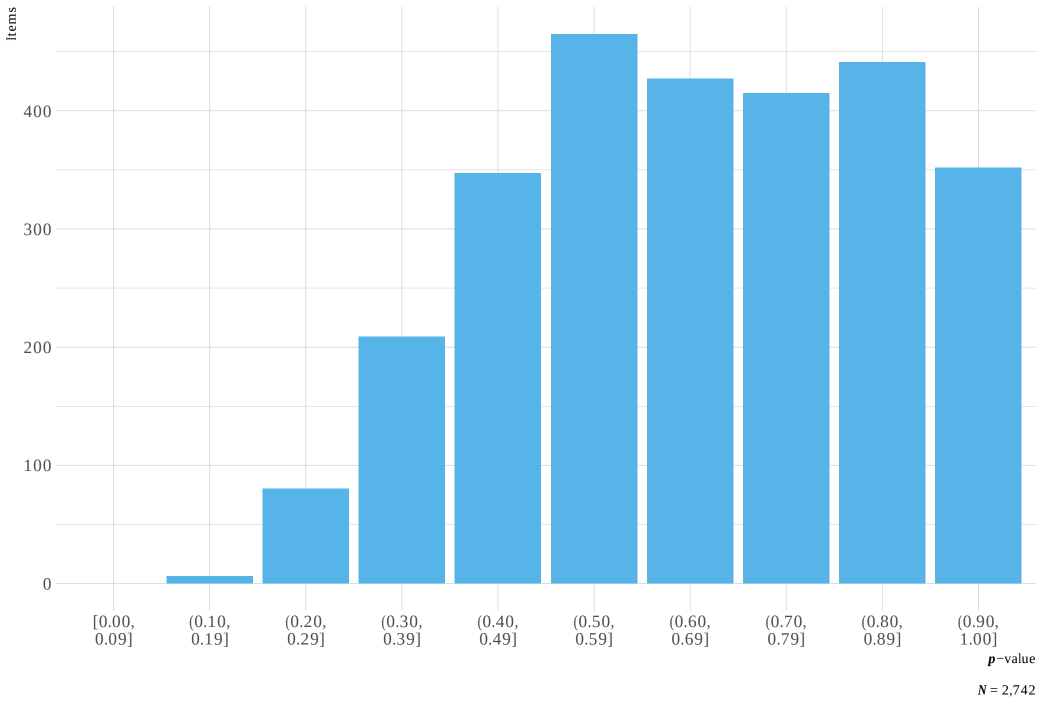


Figure 3.2. *p*-values for mathematics 2018 operational items. *Note*. Items with a sample size of less than 20 were omitted.

Standardized difference values were also calculated for all operational items, with a student sample size of at least 20 to 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 respective 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.

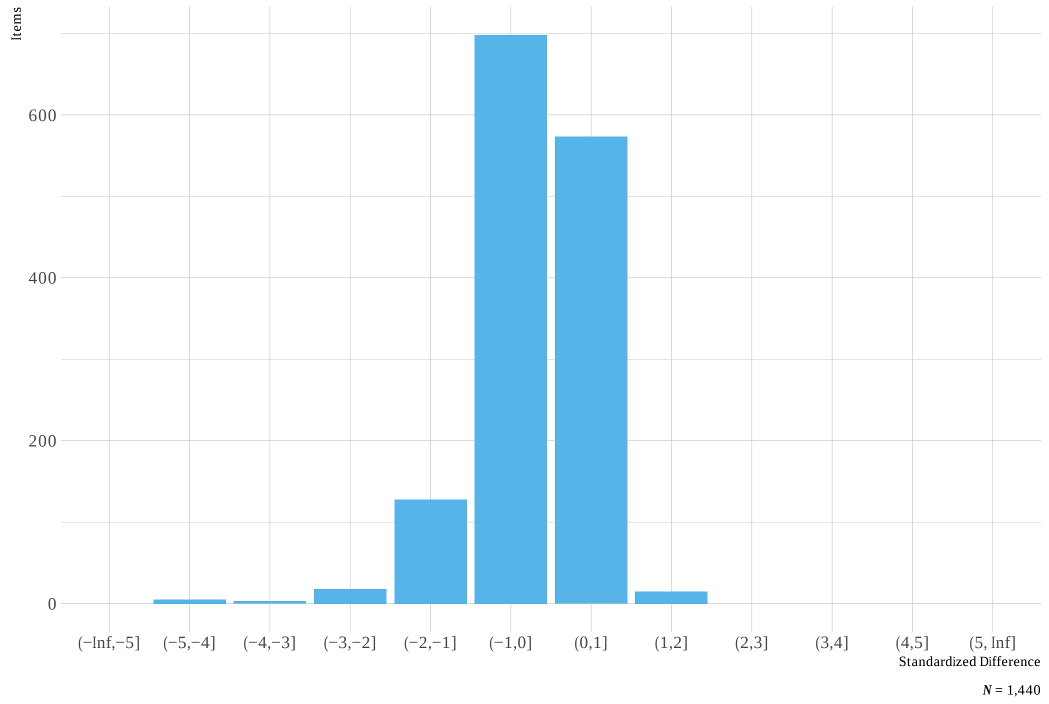


Figure 3.3. Standardized difference *z*-scores for ELA 2018 operational items. *Note.* Items with a sample size of less than 20 were omitted.

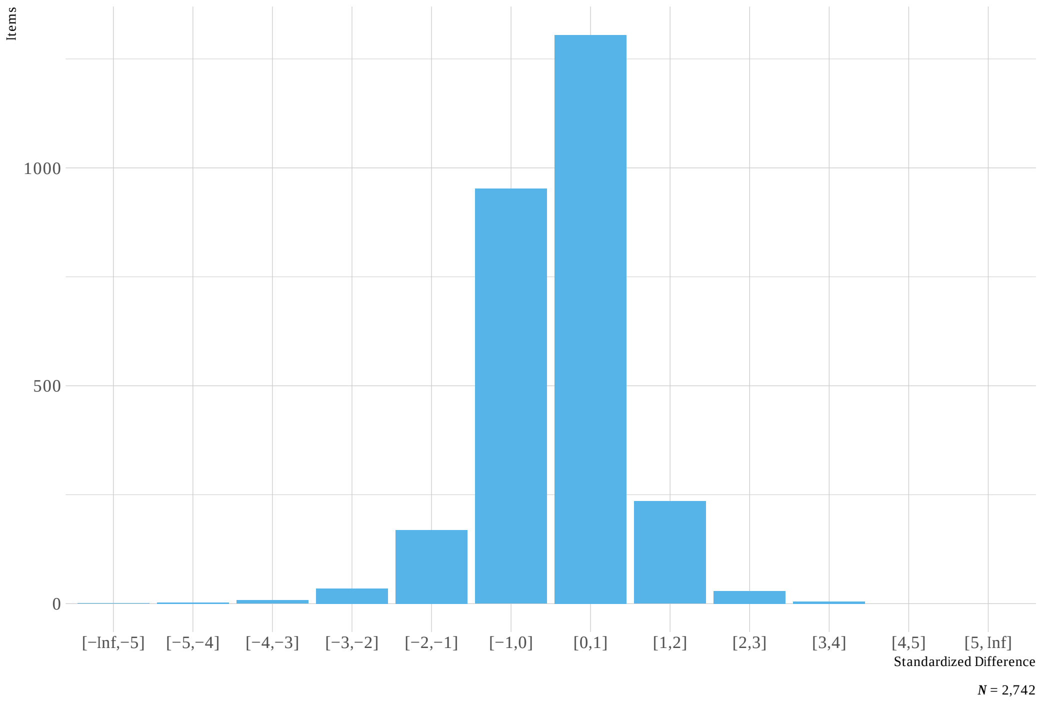


Figure 3.4. Standardized difference *z*-scores for mathematics 2018 operational items. *Note.* Items with a sample size of less than 20 were omitted.

3.4. Field Testing

During the spring 2018 administration, DLM field tests were administered to evaluate item quality for EEs assessed at each grade level for ELA and mathematics. Field testing is conducted to deepen operational pools so that multiple testlets are available in each window. By deepening the operational pools, testlets can also be evaluated for retirement in instances where other testlets perform better.

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, & Wells-Moreaux, 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. During the spring administration, all 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 operational 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 operational pool.

For the spring field test, 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 the field test content was drawn, ELA EEs are banded in grades 9 and 10. Therefore, one EE was selected from the grade band, that was administered to both grade 9 and grade 10 students in ELA. This resulted in a total of 26 EEs being selected for the spring field test. Although two mathematics EEs were selected for field testing, both EEs were administered on a single form. Table 3.6 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. There was one linkage level and grade 3 and grade 4 that had two field test testlets available in ELA.

Table 3.6. Spring 2018 Field Test Testlets Available

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

Note:
In mathematics, two testlets were administered on a single form. ELA is grade banded in grades 9–10.

Participation in spring field testing was not required in any state, but teachers were encouraged to administer all available testlets to their students. Participation rates for ELA and mathematics in spring 2018 are shown in Table 3.7. In total, 89% of students in ELA and 88% 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). The purpose of the spring field test was to collect additional cross-linkage-level data, and thus the design utilized the pool of currently available operational testlets; therefore, test development team review of items included in the field test was not necessary.

Table 3.7. Students Who Completed a Field Test Testlet, by Subject

Subject	***n***	**%**
English language arts	63,209	89.4
Mathematics	61,842	87.5

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 2017–2018, including a summary of 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 administration time, 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 2017–2018. For a complete description of key administration features, including information on assessment delivery, the KITE® system, 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 2017–2018* (DLM Consortium, 2017e) and the *Educator Portal User Guide* (DLM Consortium, 2017d).

4.1.1. Test Windows

Instructionally embedded assessments were available for teachers to optionally administer between September 20 and December 20, 2017, and between January 2 and February 28, 2018. During the consortium-wide spring testing window, which occurred between March 12 and June 8, 2018, 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 2018 operational administration of the DLM alternate assessment. The categories of evidence include data relating to the adaptive delivery of testlets in the spring window, user experience, and accessibility.

4.2.1. Adaptive Delivery

During the spring 2018 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 4.1.

Table 4.1. 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 2018 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 4.2 and Table 4.3 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 95.1% 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. Further investigation is needed to evaluate reasons for these different patterns.

The 2017–2018 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 4.2. Adaptation of Linkage Levels Between First and Second English Language Arts Testlets (*N* = 70,723)

	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	20.5	79.5	32.5	39.6	27.9	69.8	14.8	15.3	85.4	9.6	5.1
Grade 4	33.2	66.8	18.0	42.4	39.5	33.3	42.3	24.4	55.5	43.0	1.6
Grade 5	21.3	78.7	26.8	31.1	42.2	62.0	26.1	11.9	65.1	26.1	8.9
Grade 6	14.4	85.6	23.0	10.1	66.9	40.0	22.5	37.5	37.0	21.1	41.9
Grade 7	18.7	81.3	20.4	31.2	48.5	32.1	34.9	33.0	41.1	32.2	26.7
Grade 8	34.2	65.8	30.3	41.8	27.9	51.1	39.2	9.7	84.8	12.2	3.1
Grade 9	10.6	89.4	18.7	9.7	71.6	33.4	14.0	52.6	43.2	10.5	46.3
Grade 10	7.1	92.9	14.5	36.5	49.1	25.5	45.2	29.3	45.0	46.2	8.9
Grade 11	12.2	87.8	4.6	26.5	68.9	26.5	40.8	32.7	40.3	43.8	15.9

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

Table 4.3. Adaptation of Linkage Levels Between First and Second Mathematics Testlets (N = 70,694)

Grade	Foundational		Band 1			Band 2			Band 3		
	Adapted Up (%)	Did Not Adapt (%)	Adapted Up (%)	Did Not Adapt (%)	Adapted Down (%)	Adapted Up (%)	Did Not Adapt (%)	Adapted Down (%)	Adapted Up (%)	Did Not Adapt (%)	Adapted Down (%)
Grade 3	4.9	95.1	7.0	30.2	62.8	14.4	26.9	58.7	9.3	48.4	42.2
Grade 4	15.2	84.8	51.3	13.9	34.8	62.1	18.4	19.6	51.7	23.0	25.3
Grade 5	22.8	77.2	10.8	18.8	70.5	16.1	9.1	74.8	57.1	6.4	36.5
Grade 6	11.7	88.3	12.6	27.0	60.4	17.8	32.6	49.6	40.9	28.4	30.7
Grade 7	11.5	88.5	8.1	20.6	71.3	33.7	34.1	32.3	38.9	8.9	52.1
Grade 8	16.5	83.5	14.3	6.5	79.2	3.7	11.8	84.5	14.1	17.1	68.9
Grade 9	18.3	81.7	7.7	34.6	57.7	8.5	41.5	50.0	12.5	46.4	41.0
Grade 10	9.7	90.3	1.1	21.0	77.9	2.0	21.6	76.4	18.5	53.1	28.4
Grade 11	12.0	88.0	2.7	27.2	70.1	3.2	25.0	71.8	9.1	57.2	33.7

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

4.2.2. *Administration Incidents*

As in all previous operational years, testlet assignment during the spring 2018 assessment window was monitored to ensure students were correctly assigned to testlets. Administration incidents that have the potential to affect scoring are reported to states in a supplemental Incident File. Improving on the previous operational years, no incidents were observed during the spring 2018 administration. 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 2018 operational implementation of the DLM alternate assessment. The categories of evidence include data relating to user experience and accessibility.

4.3.1. *User Experience With the DLM System*

User experience with the spring 2018 assessments was evaluated through the spring 2018 survey, which was disseminated to teachers who had administered a DLM assessment during the spring window. In 2018, the survey was distributed to teachers in KITE Client, 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 Client feedback, the relationship of assessment content to instruction by subject, and teacher experience with the system.

A total of 14,922 teachers in year-end model states responded to the survey (with a response rate of 77.7%) for 48,249 students.

Participating teachers responded to surveys for between one and 74 students. Teachers most frequently reported having 0 to 5 years of experience in ELA, mathematics, and with students with significant cognitive disabilities. The median response to the number of years of experience in each of these areas was 6 to 10 years. Approximately 36% indicated they had experience administering the DLM assessment in all four 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, 2016; DLM Consortium, 2017a; DLM Consortium, 2017b).

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 4.4.

Nearly all teachers (97.1%) agreed or strongly agreed that they were confident administering DLM testlets. Most respondents (91.4%) agreed or strongly agreed that the required test administrator training prepared them for their responsibilities as test administrators. Most teachers also responded that manuals and the Educator Resources page helped them understand how to use the system (91.2%); that they knew how to use accessibility supports, allowable supports, and options for flexibility (94.8%); and that the Testlet Information Pages helped them deliver the testlets (90.1%).

Table 4.4. Teacher Responses Regarding Test Administration

	SD		**D**		**A**		**SA**		**A+SA**	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Confidence in ability to deliver DLM testlets	111	1.0	205	1.9	4,626	43.2	5,776	53.9	10,402	97.1
Test administrator training prepared respondent for responsibilities of test administrator	229	2.1	688	6.4	5,476	51.3	4,278	40.1	9,754	91.4
Manuals and DLM Educator Resources Page materials helped respondent understand how to use assessment system	198	1.9	731	6.9	5,876	55.1	3,850	36.1	9,726	91.2
Respondent knew how to use accessibility features, allowable supports, and options for flexibility	133	1.2	424	4.0	5,823	54.6	4,288	40.2	10,111	94.8
Testlet Information Pages helped respondent to deliver the testlets	239	2.2	811	7.6	5,727	53.6	3,903	36.5	9,630	90.1

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 Client and Educator Portal.

The software used for the administration of DLM testlets is KITE Client. Teachers were asked to consider their experiences with KITE Client and respond to each question on a five-point scale: *very hard, somewhat hard, neither hard nor easy, somewhat easy*, or *very easy*. Table 4.5 summarizes teacher responses to these questions.

Respondents found it to be either *somewhat easy* or *very easy* to log in to the system (79%), to navigate within a testlet (82.9%), to record a response (85.3%), to submit a completed testlet (86.1%), and to administer testlets on various devices (76.3%). Open-ended survey response feedback indicated testlets were easy to administer and that technology had improved compared to previous years.

Table 4.5. Ease of Using KITE Client

	VH		**SH**		**N**		**SE**		**VE**		**SE+VE**	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Enter the site	104	1.0	493	4.6	1,643	15.4	3,182	29.9	5,225	49.1	8,407	79.0
Navigate within a testlet	86	0.8	307	2.9	1,425	13.4	3,029	28.5	5,791	54.4	8,820	82.9
Record a response	71	0.7	207	2.0	1,281	12.1	2,774	26.2	6,260	59.1	9,034	85.3
Submit a completed testlet	67	0.6	165	1.6	1,245	11.8	2,627	24.9	6,466	61.2	9,093	86.1
Administer testlets on various devices	137	1.3	391	3.7	1,983	18.8	3,055	28.9	5,009	47.4	8,064	76.3

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

Educator Portal is an area of the KITE system used to store and manage student data and enter PNP and First Contact information. Teachers were asked to assess the ease of navigating and using Educator Portal for its intended purposes. The data are summarized in Table 4.6 using the same scale used to rate experiences with KITE Client. Overall, respondents' feedback was mixed to favorable: a majority of teachers found it to be either *somewhat easy* or *very easy* to navigate the site (69.4%), enter PNP and First Contact information (73.9%), manage student data (67.9%), manage their accounts (70.2%), or manage tests (69.8%).

Open-ended survey responses indicated that teachers want less wait time between testlet generation. They also want to be able to generate Testlet Information Pages for the entire class at one time.

Table 4.6. Ease of Using Educator Portal

	VH		SH		N		SE		VE		SE+VE	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Navigate the site	226	2.1	1,106	10.4	1,927	18.1	3,582	33.6	3,814	35.8	7,396	69.4
Enter Access Profile and First Contact information	118	1.1	689	6.5	1,963	18.5	3,726	35.1	4,114	38.8	7,840	73.9
Manage student data	225	2.1	1,022	9.6	2,171	20.4	3,769	35.5	3,443	32.4	7,212	67.9
Manage my account	167	1.6	793	7.5	2,209	20.8	3,816	35.9	3,649	34.3	7,465	70.2
Manage tests	253	2.4	959	9.0	1,999	18.8	3,609	34.0	3,808	35.8	7,417	69.8

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

Finally, respondents were asked to rate their overall experience with KITE Client and Educator Portal on a four-point scale: *poor, fair, good,* and *excellent*. Results are summarized in Table 4.7. The majority of respondents reported a positive experience with KITE Client. A total of 84.6% of respondents rated their KITE Client experience as *good* or *excellent*, while 77.7% rated their overall experience with Educator Portal as *good* or *excellent.*

Table 4.7. Overall Experience With KITE Client and Educator Portal

	Poor		Fair		Good		Excellent	
Statement	*n*	%	*n*	%	*n*	%	*n*	%
KITE Client	225	2.1	1,416	13.3	5,297	49.7	3,724	34.9
Educator Portal	397	3.7	1,990	18.6	5,499	51.5	2,799	26.2

Overall, feedback from teachers indicated that KITE Client was easy to navigate and user friendly. Teachers also provided useful feedback about how to improve the Educator Portal user experience, which will be considered for technology development for 2018–2019 and beyond.

4.3.2. *Accessibility*

Accessibility supports provided in 2017–2018 were the same as those available in previous years. DLM accessibility guidance, in accordance with DLM Consortium (2017c), distinguishes among accessibility supports that are provided in KITE Client via the Access Profile[2], require additional tools or materials, and are provided by the test administrator outside the system.

[2] The Access Profile includes both the PNP profile and the First Contact Survey.

Table 4.8 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 4.8. Accessibility Supports Selected for Students (*N* = 68,392)

Support	*n*	%
Supports provided in KITE Client via Access Profile		
Spoken audio	12,535	18.3
Magnification	7,624	11.1
Color contrast	5,836	8.5
Overlay color	3,857	5.6
Invert color choice	2,625	3.8
Supports requiring additional tools/materials		
Individualized manipulatives	31,180	45.6
Calculator	20,167	29.5
Single-switch system	2,005	2.9
Alternate form - visual impairment	1,614	2.4
Two-switch system	936	1.4
Uncontracted braille	37	0.1
Supports provided outside the system		
Human read aloud	60,230	88.1
Test administrator enters responses for student	36,469	53.3
Partner assisted scanning	6,091	8.9
Language translation of text	1,368	2.0
Sign interpretation of text	1,121	1.6

Table 4.9 describes teacher responses to survey items about the accessibility supports used during administration. Teachers were asked to respond to two items using a four-point Likert-type scale (*strongly disagree*, *disagree*, *agree*, or *strongly agree*) or indicate if the item did not apply to the student. The majority of teachers agreed that students were able to effectively use accessibility supports (81.6%), and that accessibility supports were similar to ones students used for instruction (82.4%). These data support the conclusions that the accessibility supports of the DLM alternate assessment were effectively used by students, emulated accessibility supports used during instruction, and met student needs for test administration. Additional data will be collected during the spring 2019 survey to determine whether results improve over time.

Table 4.9. Teacher Report of Student Accessibility Experience

	SD		D		A		SA		A+SA		N/A	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Student was able to effectively use accessibility features.	301	2.7	377	3.4	4928	44.2	4166	37.4	9094	81.6	1365	12.3
Accessibility features were similar to ones student uses for instruction.	280	2.5	421	3.8	4921	44.3	4237	38.1	9158	82.4	1253	11.3

Note: SD = strongly disagree; D = disagree; A = agree; SA = strongly agree; A+SA = agree and strongly agree. N/A = not applicable.

4.4. Conclusion

During the 2017–2018 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 Access Profile selections. Results from the teacher survey indicated that teachers felt confident administering testlets in the system, that KITE Client was easy to use, and that Educator Portal posed some challenges but 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 2017–2018 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 2017–2018 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. A summary of the operational parameters used to score the 2017–2018 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 2018 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, 98.5%) performed as expected.

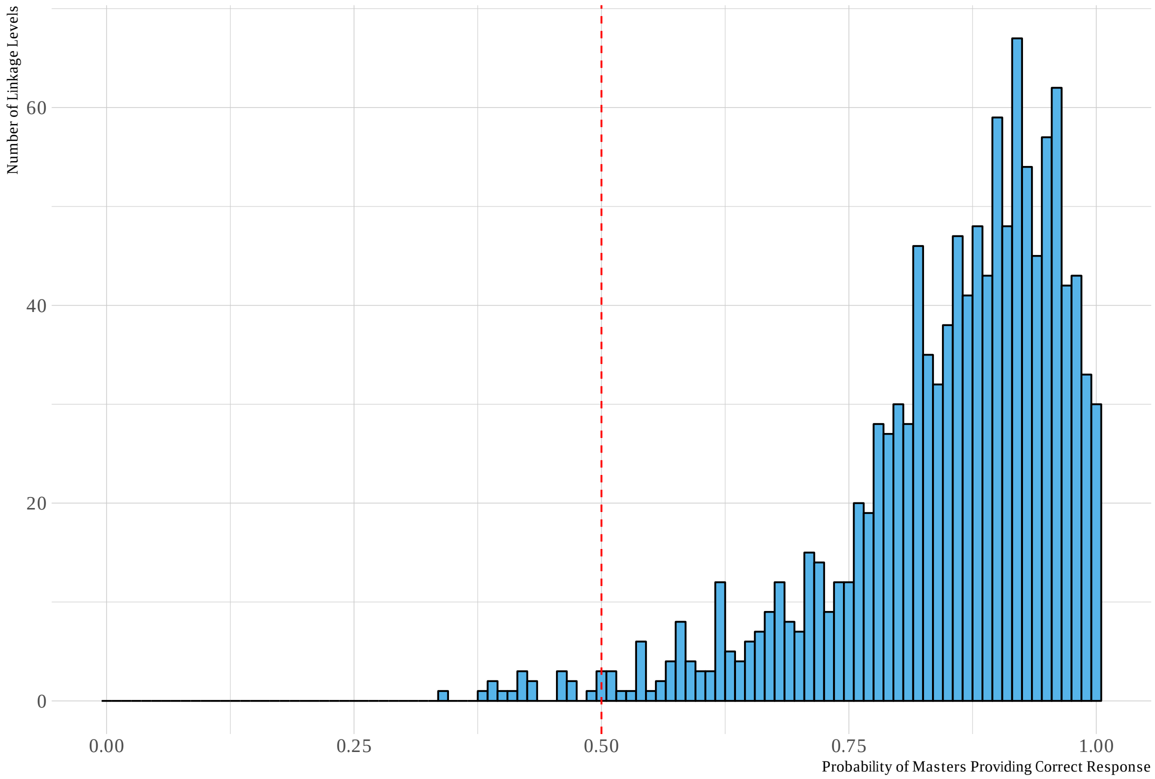


Figure 5.1. Probability of masters providing a correct response to items measuring each linkage level. *Note*: Histogram bins are shown in increments of .01. Reference line indicates .5.

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 = 892, 73.7%) performed as expected, non-masters sometimes had a greater than chance (> .5) likelihood of providing a correct response to items measuring the linkage level. 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.

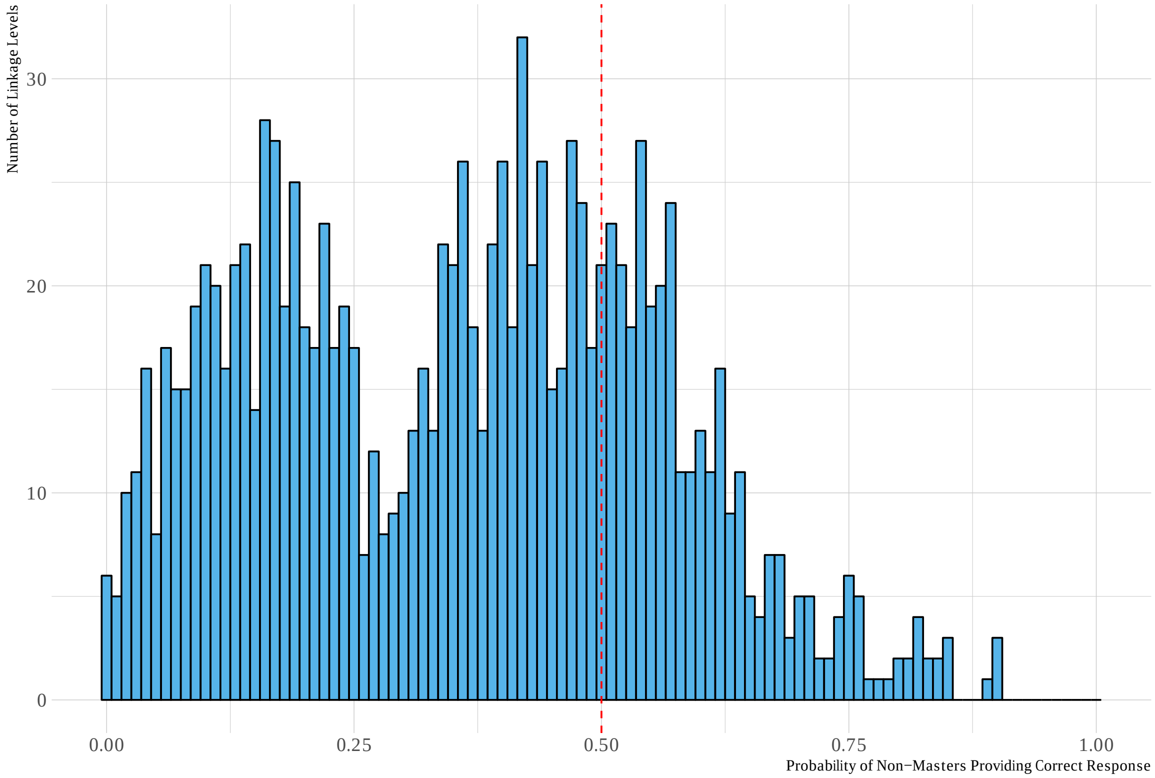


Figure 5.2. Probability of non-masters providing a correct response to items measuring each linkage level. *Note*: Histogram bins are in increments of .01. Reference line indicates .5.

5.3. 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). 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 2017–2018 administration of DLM assessments, the percentage of mastery statuses obtained by each scoring rule was calculated, as shown in Figure 5.3. 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 66% to 78% of mastered linkage levels were derived from the posterior probability obtained from the modeling procedure. The next approximately 2% to 19% of linkage levels were assigned mastery status by the percentage correct rule. The remaining approximately 10% to 30% 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 2017–2018 operational year, the rate of agreement between the two methods was 83%. 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).

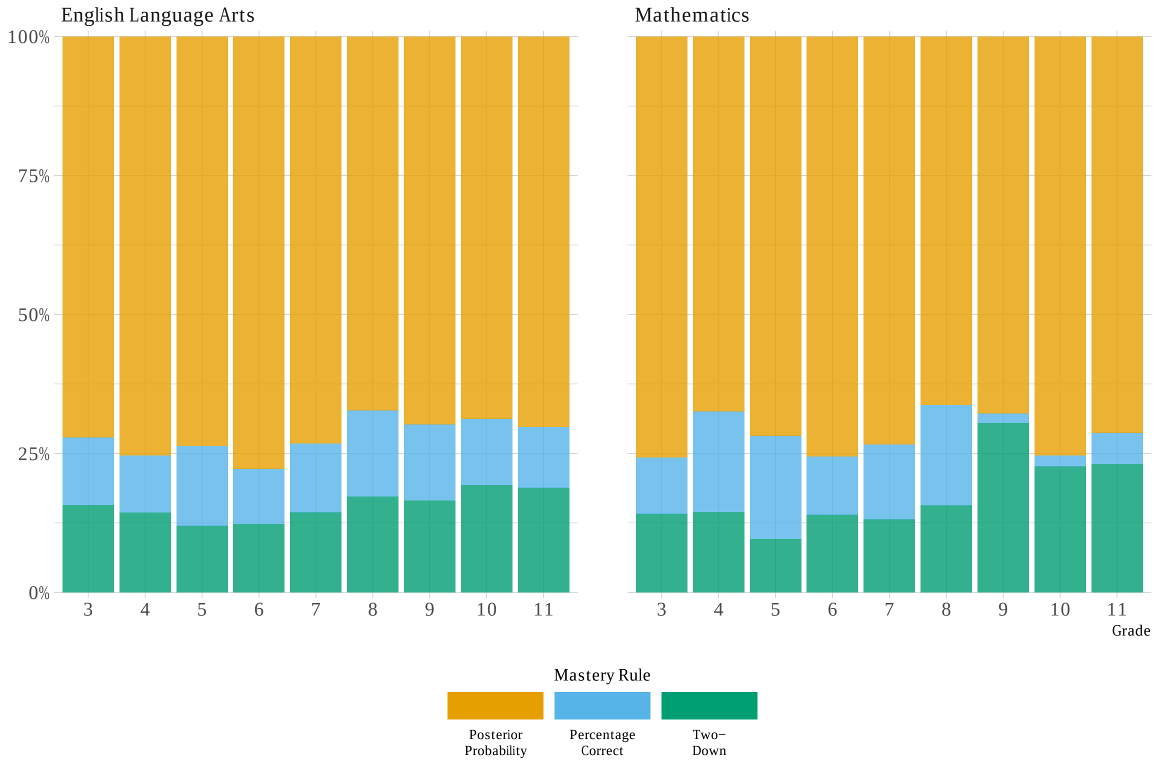


Figure 5.3. Linkage level mastery assignment by mastery rule for each subject and grade.

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 operational year, 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 2017–2018 year, the modeling subcommittee reviewed research related to Bayesian methods for assessing modeling fit using posterior predictive model checks (Gelman & Hill, 2006; Gelman, Meng, & Stern, 1996) and a newly defined model with partial equivalency of model parameters.

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 2017–2018 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 2018. Finally, this chapter describes updates made to score reports and data files during spring 2018. 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 2018, assessments were administered to 70,968 students in 12 statesand 1 Bureau of Indian Education (BIE) school. Counts of students tested in each state and BIE are displayed in Table 7.1. The assessments were administered by 18,683 educators in 9,836 schools and 3,146 school districts.

Table 7.1. Student Participation by State (*N* = 70,968)

State	Students (*n*)
Alaska	542
Colorado	5,224
Delaware	1,114
Illinois	11,524
Miccosukee Indian School	9
New Hampshire	857
New Jersey	11,409
New York	21,599
Oklahoma	5,842
Rhode Island	1,000
Utah	4,258
West Virginia	1,707
Wisconsin	5,883

Table 7.2 summarizes the number of students tested in each grade. In grades 3 through 8, over 9,200 students participated in each grade. In high school, the largest number of students participated in grade 11, and the smallest number participated in grade 10. 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* = 70,968)

Grade	Students (*n*)
3	9,278
4	9,567
5	9,522
6	9,642
7	9,642
8	9,886
9	5,250
10	1,810
11	6,371

Table 7.3 summarizes the demographic characteristics of the students who participated in the spring 2018 administration. The majority of participants were male (67%) and white (60%). About 6% of students were monitored or eligible for EL services.

Table 7.3. Demographic Characteristics of Participants (*N* = 70,968)

Subgroup	*n*	%
Gender		
Male	47,560	67.02
Female	23,407	32.98
Missing	1	<0.01
Race		
White	42,422	59.78
African American	14,315	20.17
Two or more races	7,750	10.92
Asian	3,611	5.09
American Indian	2,247	3.17
Native Hawaiian or Pacific Islander	386	0.54
Alaska Native	233	0.33
Missing	4	0.01
Hispanic ethnicity		
No	54,031	76.13
Yes	16,930	23.86
Missing	7	0.01
English learner (EL) participation		
Not EL eligible or monitored	66,616	93.87
EL eligible or monitored	4,352	6.13

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 268 students took at least one instructionally embedded testlet during the 2017–2018 academic year.

Table 7.4. Students Completing Instructionally Embedded Testlets by State (*N* = 268)

State	*n*
Colorado	22
Delaware	26
Illinois	6
New Hampshire	1
New York	43
Oklahoma	128
Utah	26
West Virginia	16

Table 7.5 summarizes the number of instructionally embedded test sessions taken in ELA and mathematics. Across all states, students took 1,416 ELA testlets and 1,476 mathematics testlets.

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

Grade	English language arts	Mathematics
3	143	120
4	205	190
5	290	249
6	159	164
7	232	258
8	199	199
9	26	45
10	24	26
11	136	224
Total	*1,414*	*1,475*

7.2. Student Performance

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

For the spring 2018 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 2018 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 24% to 38%. In mathematics, the percentage of students meeting or exceeding Target expectations ranged from approximately 8% to 32%.

Table 7.6. Percentage of Students by Grade and Performance Level

Grade	**Emerging (%)**	**Approaching (%)**	**Target (%)**	**Advanced (%)**	**Target+ Advanced (%)**
English language arts					
3 (*n* = 9,257)	60.1	15.7	21.8	2.4	24.2
4 (*n* = 9,548)	51.4	20.6	23.6	4.4	28.0
5 (*n* = 9,499)	49.0	19.2	27.4	4.4	31.8
6 (*n* = 9,619)	49.5	24.2	16.6	9.6	26.2
7 (*n* = 9,615)	36.1	26.2	25.0	12.7	37.7
8 (*n* = 9,863)	38.0	25.8	26.6	9.7	36.2
9 (*n* = 5,228)	34.2	30.5	25.8	9.5	35.3
10 (*n* = 1,807)	31.6	31.9	30.6	5.9	36.5
11 (*n* = 6,307)	35.5	32.4	26.8	5.4	32.2
Mathematics					
3 (*n* = 9,245)	59.5	15.5	17.3	7.7	25.0
4 (*n* = 9,539)	51.6	16.8	22.1	9.5	31.6
5 (*n* = 9,489)	59.7	20.3	10.5	9.5	20.0
6 (*n* = 9,611)	57.2	25.0	10.1	7.7	17.8
7 (*n* = 9,604)	63.1	24.5	7.4	5.0	12.4
8 (*n* = 9,854)	54.7	31.6	10.7	3.0	13.7
9 (*n* = 5,240)	48.8	32.3	15.0	3.9	18.9
10 (*n* = 1,805)	48.8	39.2	11.4	0.7	12.1
11 (*n* = 6,327)	61.0	30.8	8.0	0.3	8.2

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. Rows labeled Missing indicate the student's demographic data were not entered into the system.

Table 7.7. Students at Each ELA Performance Level, by Demographic Subgroup (*N* = 70,743)

	Emerging		**Approaching**		**Target**		**Advanced**	
Subgroup	*n*	%	*n*	%	*n*	%	*n*	%
Gender								
Male	21,273	44.9	11,208	23.6	11,484	24.2	3,434	7.2
Female	10,420	44.6	5,643	24.2	5,615	24.1	1,665	7.1
Missing	0	0.0	0	0.0	0	0.0	1	100.0
Race								
White	18,892	44.7	9,971	23.6	10,313	24.4	3,107	7.3
African American	5,897	41.3	3,568	25.0	3,676	25.8	1,133	7.9
Two or more races	3,820	49.5	1,791	23.2	1,668	21.6	445	5.8
Asian	1,994	55.4	806	22.4	620	17.2	178	4.9
American Indian	793	35.3	557	24.8	708	31.6	186	8.3
Native Hawaiian or Pacific Islander	157	40.9	106	27.6	77	20.1	44	11.5
Alaska Native	140	60.3	50	21.6	36	15.5	6	2.6
Missing	0	0.0	2	50.0	1	25.0	1	25.0
Hispanic ethnicity								
No	24,105	44.7	12,860	23.9	13,034	24.2	3,877	7.2
Yes	7,585	45.0	3,989	23.7	4,064	24.1	1,222	7.2
Missing	3	42.9	2	28.6	1	14.3	1	14.3
English learner (EL) participation								
Not EL eligible or monitored	29,891	45.0	15,690	23.6	16,042	24.2	4,782	7.2
EL eligible or monitored	1,802	41.5	1,161	26.8	1,057	24.4	318	7.3

Table 7.8. Students at Each Mathematics Performance Level, by Demographic Subgroup (*N* = 70,714)

	Emerging		Approaching		Target		Advanced	
Subgroup	*n*	%	*n*	%	*n*	%	*n*	%
Gender								
Male	26,532	56.0	11,358	24.0	6,264	13.2	3,216	6.8
Female	13,799	59.1	5,829	25.0	2,662	11.4	1,053	4.5
Missing	0	0.0	1	100.0	0	0.0	0	0.0
Race								
White	24,297	57.5	10,439	24.7	5,172	12.2	2,359	5.6
African American	7,634	53.5	3,533	24.8	2,035	14.3	1,066	7.5
Two or more races	4,684	60.7	1,780	23.1	855	11.1	403	5.2
Asian	2,318	64.6	643	17.9	416	11.6	212	5.9
American Indian	1,033	46.0	641	28.6	373	16.6	198	8.8
Native Hawaiian or Pacific Islander	203	52.6	96	24.9	56	14.5	31	8.0
Alaska Native	161	69.1	55	23.6	17	7.3	0	0.0
Missing	1	25.0	1	25.0	2	50.0	0	0.0
Hispanic ethnicity								
No	31,012	57.6	13,206	24.5	6,624	12.3	3,001	5.6
Yes	9,314	55.2	3,982	23.6	2,301	13.6	1,267	7.5
Missing	5	71.4	0	0.0	1	14.3	1	14.3
English learner (EL) participation								
Not EL eligible or monitored	38,131	57.4	16,111	24.3	8,231	12.4	3,902	5.9
EL eligible or monitored	2,200	50.7	1,077	24.8	695	16.0	367	8.5

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 7% 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 44% in grade 6 to 56% in grade 7. For mathematics, the average percentage of students who mastered the Target or Successor linkage level across all EEs ranged from approximately 14% in grade 11 to 31% 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* = 9,257)	2.6	7.1	23.6	21.8	16.7	28.2
4 (*n* = 9,548)	3.3	6.6	23.8	10.6	14.5	41.2
5 (*n* = 9,499)	2.4	6.0	25.2	12.1	10.1	44.3
6 (*n* = 9,619)	2.4	6.2	28.3	18.7	7.7	36.8
7 (*n* = 9,615)	2.9	4.3	22.3	14.5	13.2	42.8
8 (*n* = 9,863)	3.2	4.6	23.7	15.0	15.5	38.1
9 (*n* = 5,228)	3.6	8.1	17.3	16.7	17.4	36.8
10 (*n* = 1,807)	3.8	8.6	19.9	15.9	14.9	36.9
11 (*n* = 6,307)	3.9	5.7	27.1	17.5	12.5	33.3

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* = 9,245)	4.7	26.2	29.7	15.3	11.9	12.2
4 (*n* = 9,539)	2.8	15.6	23.4	27.5	16.0	14.7
5 (*n* = 9,489)	4.3	20.7	36.7	17.2	10.1	10.9
6 (*n* = 9,611)	6.7	19.0	22.8	28.1	11.0	12.5
7 (*n* = 9,604)	3.9	19.5	20.2	28.7	18.5	9.3
8 (*n* = 9,854)	3.7	10.8	23.1	34.0	15.9	12.5
9 (*n* = 5,240)	6.9	21.6	19.6	22.6	13.5	15.9
10 (*n* = 1,805)	7.8	22.2	15.7	34.3	15.1	4.9
11 (*n* = 6,327)	9.2	29.8	36.2	11.0	10.4	3.4

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 2018 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 was updated in spring 2018 to include all students who exited at any point during the academic year, rather than only including students who had exited and did not later re-entered the system. Additional demographic fields were also added to this file in order to assist in the matching of students across the multiple return files. In the event of observed incidents during assessment delivery, state partners are provided with an Incident File describing students impacted. Because no incidents were observed during the spring 2018 administration, these files were not delivered.

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 back to DLM staff. The final GRF was uploaded to Educator Portal and used to generate score reports.

In addition to the GRF and its supplemental files, states were provided with a de-identified teacher survey data file. The file provided state-specific teacher survey responses, with all identifying information about the student and educator removed. For more information regarding survey content and response rates, see Chapter 4 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 2018; however, classroom and school reports were generated by the system in Educator Portal following final GRF upload (as the district and state reports were beginning in 2016–2017), rather than being generated outside the system by the score report program. 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).

7.5. Quality Control Procedures for Data Files and Score Reports

No changes were made to the manual or automated quality control procedures for spring 2018. 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 2018 administration, four data files were delivered to state partners. Overall, between 8% and 38% of students achieved at the At Target or Advanced levels across all grades and subjects, which is consistent with prior years. No incidents were observed during the spring 2018 administration, so an incident file was not needed. Future years will consider any additional updates needed for score reports.

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 reporting levels. 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 2017–2018 administration year for six levels, consistent with the levels of reporting.

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, American Psychological Association, & National Council on Measurement in Education [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 testing data to generate simulated examinees. Using this process guarantees that the simulation takes on characteristics of the DLM operational test data that are likely to affect reliability results. For one simulated examinee, the process was as follows:

1. Draw with replacement the student record of one student from the operational testing data (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[3] 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.[4]
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.

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

[4] 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.

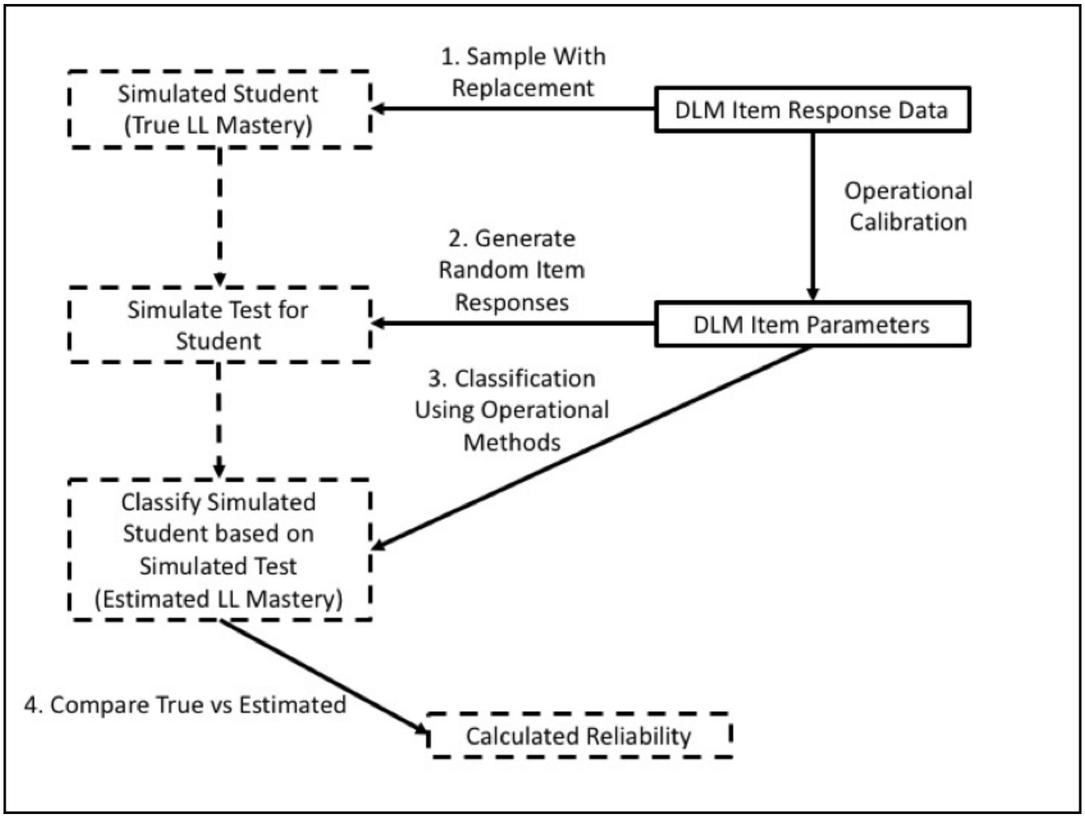

1. Sample With Replacement
Simulated Student (True LL Mastery)
DLM Item Response Data
Operational Calibration
2. Generate Random Item Responses
Simulate Test for Student
DLM Item Parameters
3. Classification Using Operational Methods
Classify Simulated Student based on Simulated Test (Estimated LL Mastery)
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[5] 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 .961 to .991. Correct classification rates range from .856 to .925 and Cohen's kappa values are between .849 and .958. 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.

[5] 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	.981	.922	.951
3	Mathematics	.990	.896	.948
4	English language arts	.984	.911	.948
4	Mathematics	.990	.893	.952
5	English language arts	.985	.925	.958
5	Mathematics	.987	.886	.943
6	English language arts	.991	.901	.950
6	Mathematics	.989	.895	.941
7	English language arts	.988	.881	.944
7	Mathematics	.986	.906	.931
8	English language arts	.988	.883	.941
8	Mathematics	.983	.895	.914
9	English language arts	.988	.884	.936
9	Mathematics	.986	.882	.915
10	English language arts	.983	.897	.935
10	Mathematics	.962	.856	.852
11	English language arts	.983	.900	.936
11	Mathematics	.961	.871	.849

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 .962 to .993. Students' average correct classification rates range from .947 to .988 and average Cohen's kappa values range from .832 to .973. These values indicate the DLM scoring procedure of reporting the number of linkage levels mastered provides reliable results of student performance.

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	.991	.963	.888
3	Mathematics	.984	.976	.922
4	English language arts	.992	.958	.868
4	Mathematics	.988	.965	.885
5	English language arts	.993	.963	.889
5	Mathematics	.988	.966	.878
6	English language arts	.990	.960	.881
6	Mathematics	.983	.972	.913
7	English language arts	.990	.953	.859
7	Mathematics	.986	.972	.910
8	English language arts	.989	.947	.832
8	Mathematics	.984	.967	.897
9	English language arts	.988	.952	.854
9	Mathematics	.979	.986	.971
10	English language arts	.989	.951	.851
10	Mathematics	.971	.988	.973
11	English language arts	.987	.955	.867
11	Mathematics	.962	.988	.971

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 4 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 .769 to .999 in ELA and from .617 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	.976	.983	.965
3	ELA.C1.2	.974	.986	.971
3	ELA.C1.3	.905	.995	.994
3	ELA.C2.1	.898	.994	.992
4	ELA.C1.1	.981	.981	.957
4	ELA.C1.2	.973	.974	.935
4	ELA.C1.3	.909	.999	.998
4	ELA.C2.1	.966	.996	.994
5	ELA.C1.1	.960	.995	.992
5	ELA.C1.2	.985	.978	.946
5	ELA.C1.3	.961	.990	.983
5	ELA.C2.1	.933	.997	.996
6	ELA.C1.1	.769	.998	.998
6	ELA.C1.2	.983	.968	.912
6	ELA.C1.3	.958	.994	.991
6	ELA.C2.1	.916	.997	.996
7	ELA.C1.1	.784	.998	.997
7	ELA.C1.2	.983	.977	.942
7	ELA.C1.3	.962	.988	.977
7	ELA.C2.1	.917	.982	.967
8	ELA.C1.2	.981	.959	.877
8	ELA.C1.3	.940	.990	.983
8	ELA.C2.1	.945	.984	.969
9	ELA.C1.2	.982	.970	.922
9	ELA.C1.3	.932	.988	.980
9	ELA.C2.1	.878	.984	.972
9	ELA.C2.2	.888	.996	.994
10	ELA.C1.2	.985	.968	.909
10	ELA.C1.3	.928	.988	.980

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
10	ELA.C2.1	.891	.987	.978
10	ELA.C2.2	.898	.996	.995
11	ELA.C1.2	.975	.974	.938
11	ELA.C1.3	.957	.985	.971
11	ELA.C2.1	.935	.987	.977
11	ELA.C2.2	.848	.996	.995

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	.924	.996	.994
3	M.C1.3	.876	.998	.998
3	M.C2.2	.862	.999	.999
3	M.C3.1	.924	.995	.994
3	M.C3.2	.836	.998	.998
3	M.C4.1	.936	.996	.994
3	M.C4.2	.732	.998	.998
4	M.C1.1	.862	.997	.996
4	M.C1.2	.843	.994	.992
4	M.C1.3	.900	.998	.998
4	M.C2.1	.938	.994	.990
4	M.C2.2	.914	.999	.999
4	M.C3.1	.950	.996	.994
4	M.C3.2	.823	.998	.998
4	M.C4.1	.893	.995	.993
4	M.C4.2	.617	.997	.997
5	M.C1.1	.791	.995	.993
5	M.C1.2	.941	.993	.989
5	M.C1.3	.942	.997	.996
5	M.C2.1	.957	.997	.997
5	M.C2.2	.934	.999	.999
5	M.C3.1	.946	.993	.989
5	M.C3.2	.893	.998	.998
5	M.C4.2	.704	.997	.997
6	M.C1.1	.865	.999	.998
6	M.C1.2	.895	.995	.993
6	M.C1.3	.934	.996	.995
6	M.C2.2	.957	.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.C3.2	.813	.998	.998
6	M.C4.1	.889	.992	.988
7	M.C1.1	.916	.996	.995
7	M.C1.2	.843	.998	.998
7	M.C1.3	.925	.993	.989
7	M.C2.1	.951	.996	.995
7	M.C2.2	.885	.999	.999
7	M.C3.2	.912	.997	.997
7	M.C4.1	.808	.998	.998
7	M.C4.2	.799	.998	.998
8	M.C1.1	.764	.997	.997
8	M.C1.2	.872	.998	.998
8	M.C1.3	.906	.996	.995
8	M.C2.1	.922	.992	.986
8	M.C2.2	.895	.999	.999
8	M.C3.2	.906	.998	.998
8	M.C4.1	.705	.998	.998
8	M.C4.2	.927	.991	.985
9	M.C1.3	.941	.995	.993
9	M.C2.1	.919	.996	.995
9	M.C2.2	.850	.999	.999
9	M.C4.1	.820	.997	.996
10	M.C1.3	.890	.999	.999
10	M.C2.1	.862	.999	.999
10	M.C3.1	.678	.998	.998
10	M.C3.2	.900	.998	.997
10	M.C4.1	.892	.998	.997
10	M.C4.2	.787	.996	.995
11	M.C1.3	.887	.998	.998
11	M.C2.1	.741	.998	.998
11	M.C3.2	.805	.999	.999
11	M.C4.2	.938	.994	.989

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 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 for number of linkage levels mastered within EEs show strong evidence of reliability.

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	<.001	.012	.062	.223	.492	.207
Correct classification rate	<.001	<.001	<.001	.041	.140	.335	.397	.079	.008
Cohen's kappa	<.001	.004	.008	.021	.041	.178	.318	.397	.033

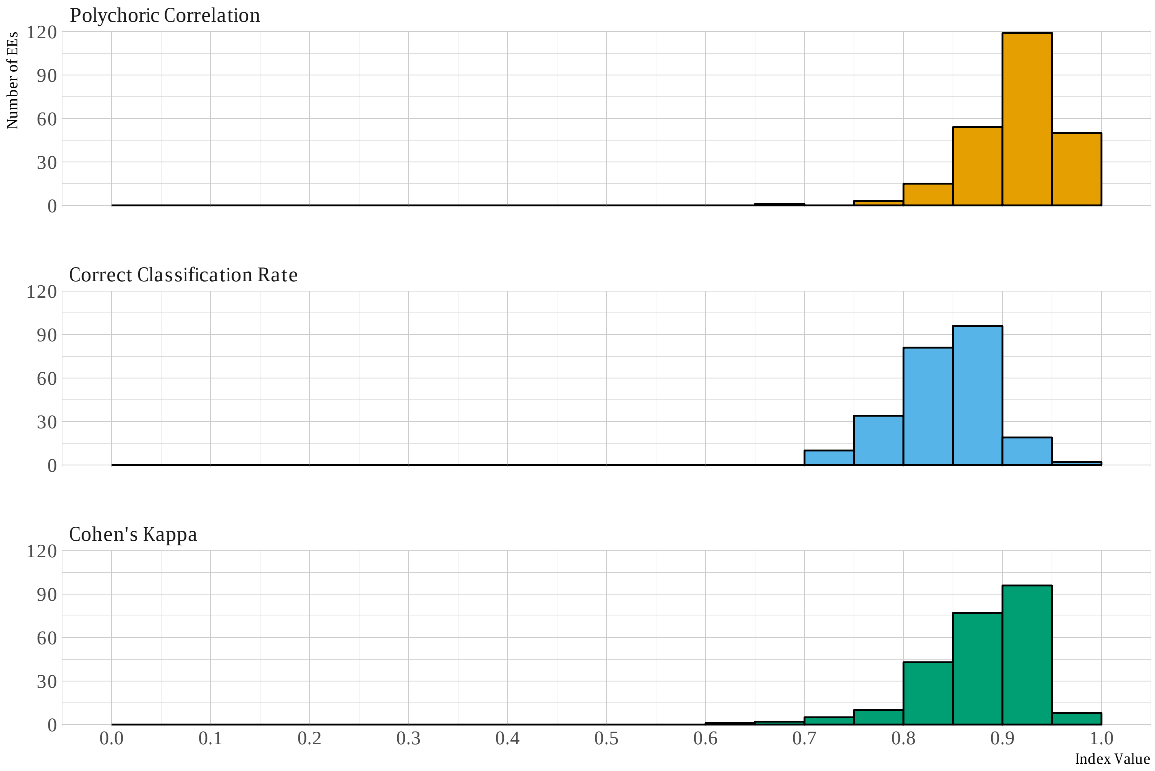


Figure 8.2. Number of linkage levels mastered within EE reliability summaries.

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.[6] 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.

[6] 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 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 42 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	.005	.017	.054	.152	.767
Correct classification rate	<.001	<.001	<.001	<.001	.002	.020	.107	.382	.490
Cohen’s kappa	.036	.032	.050	.070	.144	.186	.225	.151	.105

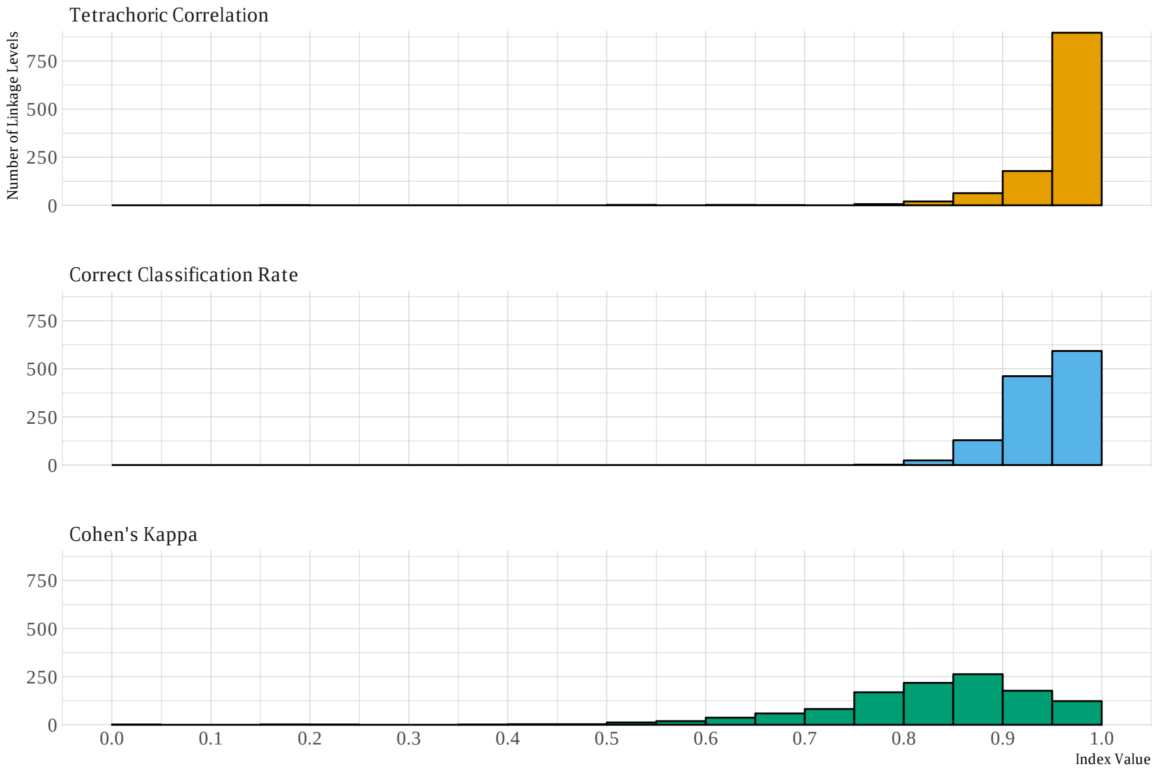


Figure 8.3. Summaries of linkage level reliability.

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.

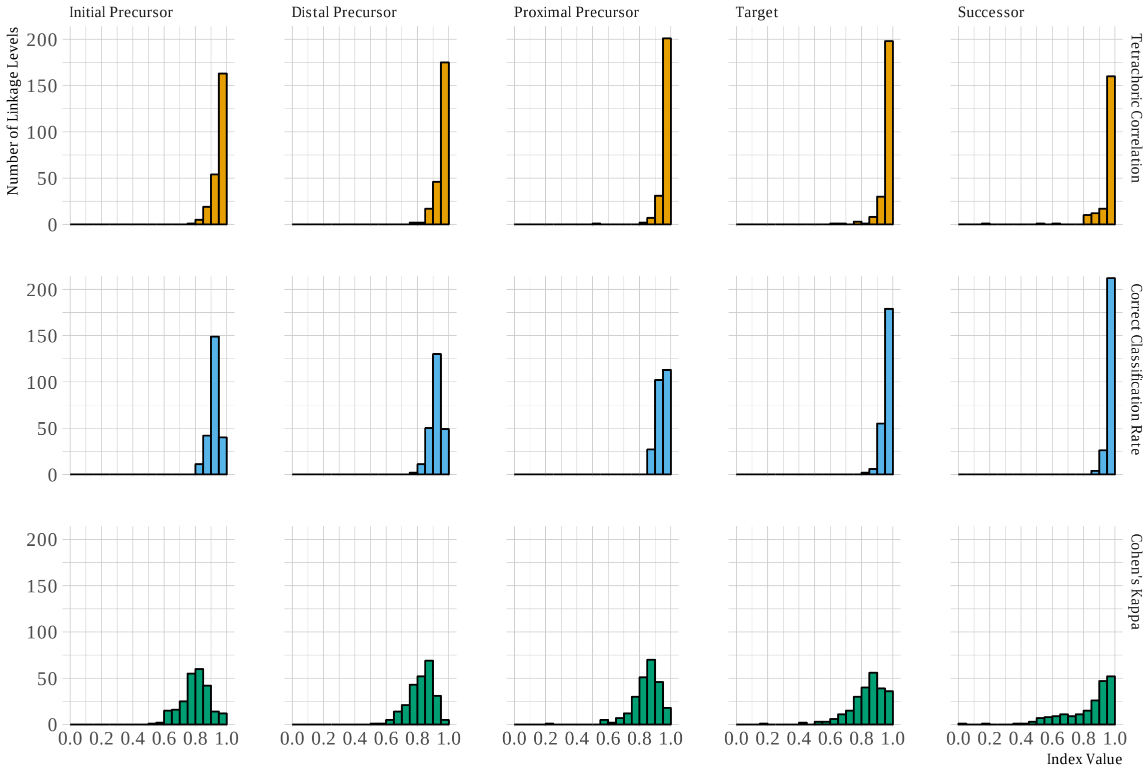


Figure 8.4. Conditional reliability evidence summarized by linkage level.

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. Chapter 9 presents additional evidence collected during 2017–2018 for 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, relation to other variables, 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; DLM Consortium, 2017b).

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 2017–2018 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 completing administration of the spring 2018 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.[7] 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 68% of responses (n = 30,658) reported that most or all reading testlets matched instruction, compared to 43% (n = 19,104) for writing and 56% (n = 25,064) 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.

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

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	2,936	6.5	10,334	23.0	18,484	41.1	12,174	27.1	1,062	2.4
Writing	5,255	11.9	9,450	21.4	11,809	26.7	7,295	16.5	10,410	23.5
Mathematics	4,261	9.5	14,059	31.4	16,376	36.6	8,688	19.4	1,357	3.0

The survey also asked teachers 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 hours	1,095	22.7	770	16.0	614	12.7	720	14.9	1,626	33.7
Construct understandings of text	16-20 hours	784	16.3	739	15.4	657	13.7	744	15.5	1,887	39.2
Integrate ideas and information from text	16-20 hours	878	18.4	804	16.8	676	14.2	821	17.2	1,595	33.4
Use writing to communicate	11-15 hours	1,117	23.3	750	15.7	642	13.4	726	15.2	1,557	32.5
Integrate ideas and information in writing	11-15 hours	1,241	25.9	770	16.1	693	14.5	734	15.3	1,351	28.2
Use language to communicate with others	>20 hours	459	9.6	539	11.2	534	11.1	720	15.0	2,554	53.1
Clarify and contribute in discussion	16-20 hours	822	17.2	709	14.8	700	14.6	817	17.1	1,739	36.3
Use sources and information	11-15 hours	1,305	27.2	859	17.9	754	15.7	739	15.4	1,149	23.9
Collaborate and present ideas	11-15 hours	1,231	25.6	894	18.6	771	16.1	772	16.1	1,133	23.6

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 hours	715	15.0	668	14.0	546	11.4	710	14.9	2,133	44.7
Compare, compose, and decompose numbers and steps	11-15 hours	1,256	26.5	802	16.9	649	13.7	716	15.1	1,312	27.7
Calculate accurately and efficiently using simple arithmetic operations	16-20 hours	1,037	21.9	613	13.0	531	11.2	685	14.5	1,867	39.4
Understand and use geometric properties of two- and three-dimensional shapes	6-10 hours	1,561	32.9	996	21.0	771	16.3	713	15.0	699	14.7
Solve problems involving area, perimeter, and volume	0-5 hours	2,485	52.4	736	15.5	561	11.8	490	10.3	472	9.9
Understand and use measurement principles and units of measure	6-10 hours	1,658	34.9	1,085	22.9	783	16.5	608	12.8	613	12.9
Represent and interpret data displays	6-10 hours	1,630	34.4	944	19.9	841	17.7	657	13.9	671	14.1
Use operations and models to solve problems	11-15 hours	1,330	28.0	782	16.5	707	14.9	763	16.1	1,162	24.5
Understand patterns and functional thinking	11-15 hours	1,052	22.1	969	20.4	900	18.9	799	16.8	1,040	21.8

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 .33, 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	.18
ELA.C1.2: Construct understandings of text	.28
ELA.C1.3: Integrate ideas and information from text	.26
ELA.C2.1: Use writing to communicate	.33
ELA.C2.2: Integrate ideas and information in writing	.29
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	.29
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	.22
M.C3.1: Understand and use measurement principles and units of measure	.28
M.C3.2: Represent and interpret data displays	.25
M.C4.1: Use operations and models to solve problems	.29
M.C4.2: Understand patterns and functional thinking	.17

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 2018 regarding students' ability to respond to testlets, test administration observation data collected during 2017–2018, and a study of interrater agreement on the scoring of teacher-administered writing testlets. For additional evidence based on response process, including studies on student and teacher behaviors during testlet administration and evidence of fidelity of administration, see Chapter 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 2018, 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.[8]

One of the fixed-form sections of the spring 2018 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 (more than 85%) 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*	%
The student responded to items to the best of their knowledge and ability.	1,618	3.6	2,966	6.6	24,265	54.1	16,034	35.7	40,299	89.8
The student was able to respond regardless of disability, behavior, or health concerns.	2,885	6.4	3,976	8.8	24,271	53.9	13,861	30.8	38,132	84.7
The student had access to all supports necessary to participate.	1,075	2.4	1,384	3.1	23,379	51.9	19,180	42.6	42,559	94.5

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

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

9.2.2. *Test Administration Observations*

Test administration observations were conducted in multiple states during 2017–2018 to further understand student response processes. Students' typical test administration process with their actual test administrator was observed. Administrations were observed for the range of students eligible for DLM assessments (i.e., students with the most significant cognitive disabilities). 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.

While all consortium states are encouraged to submit test administration observations, these observations are optional. Test administration observations were received from five states during the 2017–2018 academic year. A total of 120 test administration observations were collected. Of those, 40 (33.3%) were of computer-delivered assessments and 80 (66.7%) were of teacher-administered testlets. The observations were comprised of 60 (50%) ELA reading testlets, 16 (13%) ELA writing testlets, and 44 (37%) mathematics testlets.

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 9.6; behaviors were identified as supporting, neutral, or nonsupporting. For example, clarifying directions (48% 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, 63% of observed behaviors were classified as supporting, with no observed behaviors reflecting nonsupporting actions.

Table 9.6. Test Administrator Actions During Computer-Delivered Testlets (n = 40)

Action	*n*	%
Supporting		
Read one or more screens aloud to the student	20	50.0
Clarified directions or expectations for the student	19	47.5
Navigated one or more screens for the student	19	47.5
Repeated question(s) before student responded	14	35.0
Neutral		
Used pointing or gestures to direct student attention or engagement	21	52.5
Used verbal prompts to direct the student's attention or engagement (e.g., "look at this")	19	47.5
Asked the student to clarify or confirm one or more responses	2	5.0
Used materials or manipulatives during the administration process	1	2.5
Allowed student to take a break during the testlet	0	0.0
Repeated question(s) after student responded (i.e., gave a second trial at the same item)	0	0.0
Nonsupporting		
Physically guided the student's hand to an answer choice	0	0.0
Reduced the number of answer choices available to the student	0	0.0

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 48% 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 9.7. Independent response selection was observed in 72% 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 Client was seen in 8% 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 9.7. Student Actions During Computer-Delivered Testlets (*n* = 40)

Action	*n*	%
Selected answers independently	29	72.5
Navigated screens independently	22	55.0
Selected answers after verbal prompts	10	25.0
Navigated screens after verbal prompts	8	20.0
Navigated screens after test administrator pointed or gestured	7	17.5
Used materials outside of KITE student portal to indicate responses to testlet items	3	7.5
Independently revisited a question after answering it	2	5.0
Skipped one or more items	1	2.5
Revisited one or more questions after verbal prompt(s)	0	0.0

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 80 observations of teacher-administered testlets, observers noted difficulty in three cases (4%). 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 (2%) and using manipulatives or materials outside of KITE (8%). Additional evidence for this assumption was gathered by observing whether students were able to complete testlets. Of the 120 test administration observations collected, students completed the testlet in 113 cases (94%).[9]

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 9.8 summarizes students' response modes for teacher-administered testlets. The most frequently observed behavior was *gestured to indicate response to test administrator who selected answers*.

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

Table 9.8. Primary Response Mode for Teacher-Administered Testlets (n = 80)

Response mode	***n***	**%**
Gestured to indicate response to test administrator who selected answers	34	42.5
Used computer/device to respond independently	30	37.5
Verbally indicated response to test administrator who selected answers	17	21.2
Eye-gaze system indication to test administrator who selected answers	4	5.0
Used switch system to respond independently	0	0.0
No response	8	10.0

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 six of 40 (15%) observations of computer-delivered testlets, the test administrator entered responses on the student's behalf. In four (67%) of those cases, observers indicated that the entered response matched the student's response, while two 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 language and writing 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 9.9. 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 9.9. Number of Items That Evaluate the Writing Product per Testlet, by Grade

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

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

9.2.3.1. Recruitment

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

9.2.3.1.1. Samples

During the spring 2018 administration, state partners were asked to recruit district coordinators 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 2018 administration, each sample was submitted with a cover sheet that indicated the state, district, school, teacher, student identifier, and the testlet information page (TIP).

A total of 147 student writing samples were submitted from districts in six states. In several grades, the emergent writing testlet does not include any tasks that evaluate the writing sample (as shown in Table 9.9); therefore, 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 109 writing samples to raters for evaluation of interrater agreement.

9.2.3.1.2. Raters

The process for rating writing samples was adjusted in 2018 based on findings from the prior year. High attrition rates and incomplete rating assignments were observed during the remote writing sample rating process used to evaluate 2017 agreement. To remedy these challenges, during 2018 raters were recruited for an on-site rating event. As part of the recruitment process for the summer 2018 external review event, educators were also invited to participate in the rating of the submitted writing samples. 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 9 were selected to participate.

Raters had a range of teaching experience, as indicated in Table 9.10. Most had taught ELA and/or students with the most significant cognitive disabilities for at least six years. Furthermore, two raters (22%) reported experience as DLM external reviewers.

Table 9.10. Raters' Teaching Experience (*N* = 9)

	1–5 years		6–10 years		> 10 years	
Teaching experience	*n*	%	*n*	%	*n*	%
English language arts	4	44.4	2	22.2	3	33.3
Students with significant cognitive disabilities	3	33.3	3	33.3	3	33.3

Demographic information was collected as part of the volunteer survey administered in Qualtrics and is summarized in Table 9.11. Participating raters were mostly female (77.8%), white (88.9%), and non-Hispanic/Latino (100.0%), which was representative of the full sample who responded to the survey. Roughly one-third of raters taught in each of three settings: urban, suburban, and rural.

Table 9.11. Raters' Demographic Information (*N* = 9)

Subgroup	*n*	%
Gender		
Female	7	77.8
Male	2	22.2
Race		
White	8	88.9
Black/African-American	1	11.1
Hispanic ethnicity		
Non-Hispanic/Latino	9	100.0
Teaching setting		
Urban	4	44.4
Suburban	3	33.3
Rural	2	22.2

9.2.3.2. Sample Ratings

All ratings occured during the 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 or 14, using a partially crossed matrix design to assign each sample to a total of three raters. Thus, teachers rated between 39 and 42 writing samples. Table 9.12 summarizes the number of samples that were rated at each grade and level.

Table 9.12. Student Writing Samples with Ratings, by Grade (*N* = 109)

	Number of writing samples		
Grade	**Emergent**	**Conventional**	**Total number of samples**
3	*	8	8
4	3	10	13
5	*	5	5
6	10	12	22
7	3	10	13
8	*	8	8
9	3	18	21
10	3	3	6
11	7	6	13
Total	*29*	*80*	*109*

* 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 2018, when the writing sample was rated by the student's teacher, resulting in four ratings for each of the 109 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 to transform 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 9.13. To summarize global agreement across all student writing samples, teachers' original ratings (from spring 2018 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 conventional testlets, likely due to the testlets tending to have more tasks and having more samples collected at that level.

Table 9.13. Interrater Agreement for Writing Samples (*N* = 109)

Group	*n*	ICC	ICC lower bound	ICC upper bound	Fleiss's *κ*
Overall	109	.91	.90	.91	.71
EW	29	.87	.84	.90	.63
CW	80	.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 slightly higher in 2018 compared to 2017. The ICCs ranged from .63–.88 in 2017, and from .87–.91 in 2018. Fleiss's κ ranged from .47–.71 in 2017, and to .63–.71 in 2018. In both years, the lowest Fleiss's κ was associated with emergent level writing testlets, suggesting an improvement in the agreement for those testlets in 2018. The agreement ratings in 2018 likely provide a more accurate representation of rater agreement over the prior year due to the use of an on-site event and a consistent number of raters for each writing samples, allowing for more ratings per sample overall.

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 2018 followed the same procedure used in previous years, including data from 2015–2016 through 2016–2017 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, 3,006 (75%) items on single-EE testlets were selected. The number of items evaluated by grade level and subject ranged from 107 items in grade 7 ELA to 249 items in grade 7 ELA. Item sample sizes ranged from 232 to 11,776.

Of the 1,023 items that were not included in the DIF analysis, 867 (84.8%) had a focal group sample size of less than 100 and 156 (15.2%) had an item *p*-value greater than .95. Table 9.14 shows the number and percent of items that failed each inclusion criteria, broken down by subject and the linkage level the items assess. The majority of non-included items are from mathematics (*n* = 800; 78%), and fall in the Distal Precursor to Target linkage level. In ELA, items not included due to sample size generally come from the Initial Precursor and Distal Precursor linkage levels, whereas items not included due to *p*-values tend to come from the Proximal Precursor, Target, and Successor linkage levels.

Table 9.14. 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	42	35.3	0	0.0	0	0.0
Distal Precursor	55	46.2	1	1.0	0	0.0
Proximal Precursor	10	8.4	12	11.5	0	0.0
Target	3	2.5	60	57.7	0	0.0
Successor	9	7.6	31	29.8	0	0.0
Mathematics						
Initial Precursor	3	0.4	0	0.0	0	0.0
Distal Precursor	181	24.2	6	11.5	0	0.0
Proximal Precursor	211	28.2	25	48.1	0	0.0
Target	297	39.7	12	23.1	0	0.0
Successor	56	7.5	9	17.3	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 approach expanded on the Zumbo and Thomas effect-size classification by basing the effect-size thresholds for the simultaneous item bias test procedure (Li & Stout, 1996), which, like logistic regression, also allows for the detection of both uniform and nonuniform DIF and uses classification guidelines based on the widely accepted ETS Mantel-Haenszel classification guidelines. The Jodoin and Gierl threshold values for distinguishing negligible, moderate, and large DIF are more stringent than those of the Zumbo and Thomas approach, with lower threshold values of .035 and .07 to distinguish between negligible, moderate, and large effects. Similar to the ETS Mantel-Haenszel method, negligible effect is denoted with an A, moderate effect with a B, and large effect with a C.

Jodoin and Gierl (2001) also investigated Type I error and power rates in a simulation study examining DIF detection using the logistic regression approach. Under two of their conditions, the sample size ratio between the focal and reference groups was 1:2. The authors found that power increased and Type I error rates decreased as sample size increased for unequal sample size groups. Decreased power to detect DIF items was observed when sample size discrepancies reached a ratio of 1:4. For DLM assessments, a ratio of 1:2 is typical for items included in the analysis.

9.3.1.2. Results

9.3.1.2.1. Uniform DIF Model

A total of 399 items were flagged for evidence of uniform DIF when comparing M_0 to M_1. Table 9.15 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 19%.

Table 9.15. 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	21	132	15.9	0
4	26	157	16.6	0
5	18	155	11.6	0
6	22	134	16.4	0
7	14	107	13.1	0
8	20	109	18.3	0
9	17	131	13.0	0
10	15	154	9.7	0
11	17	142	12.0	0
Mathematics				
3	30	155	19.4	2
4	37	210	17.6	0
5	28	202	13.9	0
6	25	226	11.1	0
7	26	186	14.0	0
8	29	201	14.4	0
9	21	249	8.4	0
10	14	179	7.8	0
11	19	177	10.7	0

Using the Zumbo and Thomas (1997) 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.

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 9.16 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 9.16 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 only one item.

Table 9.16. Items Flagged for Uniform DIF With Moderate or Large Effect Size

Item ID	Grade	EE	χ^2	*p*-value	$\beta_2 G$	R^2	Z&T*	J&G*
Math								
34149	3	3.NBT.2	14.85	<.01	-0.25	.92	C	C
34150	3	3.NBT.2	58.26	<.01	-0.49	.92	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 473 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 9.17 summarizes the number of items flagged by subject and grade. The percentage of items flagged for each grade and subject ranged from 8% to 22%.

Table 9.17. 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	20	132	15.2	0
4	25	157	15.9	0
5	30	155	19.4	0
6	21	134	15.7	1
7	10	107	9.3	0
8	17	109	15.6	0
9	17	131	13.0	0
10	16	154	10.4	1
11	24	142	16.9	0
Mathematics				
3	34	155	21.9	2
4	44	210	21.0	0
5	34	202	16.8	0
6	31	226	13.7	0
7	35	186	18.8	1
8	44	201	21.9	0
9	32	249	12.9	0
10	15	179	8.4	0
11	24	177	13.6	0

Using the Zumbo and Thomas (1997) effect-size classification criteria, all but three items 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, two items had a moderate change in effect size, three had a large change in effect size, and the remaining 468 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 9.18. There were two ELA items and zero 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 two items 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 9.18. 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									
39680	6	RI.6.3	9.26	<.01	1.59	-0.19	.04	A	B
28628	10	RI.9-10.4	8.63	<.01	-5.20	0.19	.04	A	B
Math									
34149	3	3.NBT.2	14.91	<.01	-0.21	-0.00	.92	C	C
34150	3	3.NBT.2	59.71	<.01	-0.34	-0.01	.92	C	C
30547	7	7.EE.1	11.97	<.01	-0.38	0.05	.93	C	C

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 9.16), as well as the 5 items with non-negligible effect-size changes in the combined model (Table 9.18).

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, Bell, & Karvonen, 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 and mathematics items flagged with a moderate or large effect size were given a decision code of 1 by the test development teams. No evidence could be found in any of the items 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.

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 (98.5%) linkage levels, masters had a greater than .5 chance of providing a correct response to items. Similarly, for 892 (73.7%) linkage levels, non-masters had a less than .5 chance of providing a correct response to items.

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, 4,609 (96.8%) 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, one source of evidence was collected in spring 2018 via teacher survey responses regarding teacher perceptions of assessment content. An additional study was conducted to evaluate teachers' use of report contents for instructional planning and decision making.

9.4.1. Teacher Perception of Assessment Content

On the spring 2018 survey,[10] teachers were asked three questions about their perceptions of assessment content: whether the content measured important academic skills and knowledge, whether the content reflected high expectations, and whether the testlet activities were similar to instructional activities in the classroom. Table 9.19 summarizes their responses. Teachers generally agreed or strongly agreed that content reflected high expectations for their students (84%), measured important academic skills (74%), and was similar to instructional activities used in the classroom (71%).

While the majority of teachers agreed with these statements, 16-29% 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, Wakeman, Browder, Rogers, & Flowers, 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.

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

Table 9.19. Teacher Perceptions of Assessment Content

	Strongly Disagree		Disagree		Agree		Strongly Agree		Agree + Strongly Agree	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Content measured important academic skills and knowledge for this student.	4,365	9.7	7,154	15.8	26,792	59.3	6,833	15.1	33,625	74.4
Content reflected high expectations for this student.	2,166	4.8	4,956	11.0	27,302	60.8	10,516	23.4	37,818	84.2
Activities in testlets were similar to instructional activities used in the classroom.	3,851	8.6	9,163	20.4	25,400	56.6	6,472	14.4	31,872	71.0

9.4.2. *Use of Reports for Instruction*

Consequential validity evidence is collected to evaluate the extent that assessment results are used as intended. Results from DLM assessments are intended for inclusion in state accountability models; reporting results to districts, teachers, and parents; and use in instructional planning and decision making. Because summative results are delivered after the end of the school year, teacher use of results occurs in the subsequent academic year. To evaluate use of DLM summative score reports[11] for instructional planning and decision making, a series of teacher focus groups were conducted during spring 2018.

Consortium state partners recruited teachers to participate in small, virtual focus groups. Because the study focused on use of reports in the subsequent academic year, several eligibility criteria were included. To participate, teachers must have indicated they

1. currently taught one or more students who would take DLM assessments in 2017–2018,
2. received DLM 2017 summative score reports for their 2017–2018 students, and
3. used the DLM 2017 reports during the 2017–2018 academic year.

Interested teachers were asked to complete a Qualtrics survey listing their background information and responding to the three eligibility questions. A total of 135 teachers responded to the survey. Of those, 40 responded *yes* to all three eligibility questions and were contacted to set up a time to participate. Of those contacted, 17 participated in the virtual meetings. Because of attrition challenges between scheduling and conducting phone calls, the number of participants per call ranged from one to five. This resulted in several focus groups being conducted as one-on-one interviews; they are collectively referred to as focus groups throughout this section. While the views described in this section are based on a limited sample of teachers, their feedback provides some

[11] Individual student score reports include a Performance Profile, which summarizes overall performance in the subject.

evidence as to how reports are used in the subsequent academic year. Additional consequential evidence is planned (see Chapter 11 of this manual).

The 17 participating teachers represented three states and mostly self-reported as white (n = 13) and female (n = 13). Teachers taught in a range of settings, including rural (n = 2), suburban (n = 9), and urban (n = 5). Teachers reported a range of teaching experience by subject and for students with significant cognitive disabilities (SCD), with most teaching more than one subject and spanning all tested grades 3–12. Teachers indicated they taught between 1 (n = 3) and 15 or more (n = 2) students currently taking DLM assessments, with most indicating they had between 2–5 students taking DLM assessments (n = 8).

Focus groups were conducted virtually using Zoom video conferencing software. Participants were asked to describe how they used summative results from the 2016–2017 administration during the subsequent 2017–2018 academic year.

9.4.2.1. Receiving Reports

Individual student score reports are made available at the state level and to district test coordinators in Educator Portal. States and districts have differing policies regarding distribution of reports to schools, teachers, and parents at the local level. Despite responding affirmatively to the eligibility questions around score report use, several teachers indicated that the score reports they received were actually different than the example DLM reports shared in the meeting.

All teachers who received reports indicated receiving them in the fall, typically from their district or building test coordinator. Several mentioned their district test coordinator delivered reports at an annual meeting that also included required annual test administrator training. Fewer indicated receiving the reports as part of a meeting intended to discuss results. Others reported receiving only an email to notify them score reports were ready, with no additional explanation or interpretive materials provided. A review of consortium practice indicated 11 states made reports available to building test coordinators, while only three states made individual student score reports available to teachers in Educator Portal.

9.4.2.2. Using Reports to Inform Instruction

Participant discussion revealed varying levels of utility for using results to plan instruction. Teachers of elementary and middle school students whose accountability requirements included annual assessment found reports to be more useful than high school teachers, where students are typically only required to assess in a single grade for state accountability purposes (e.g., eleventh grade). Teachers noted challenges when the most recent summative score report available was from several years prior, particularly for their eleventh grade students who only had eigth grade reports available. Teachers also noted that often the curriculum in twelfth grade, as students prepared to transition, was markedly different from the eleventh grade curriculum, and therefore results from the prior year were not as useful. In contrast, elementary and middle school teachers, and especially those who instruct the same students year to year, reported much more utility in using reports for planning instruction, specifying individualized education program (IEP) goals, and planning instructional groupings.

Teachers shared varied feedback for using contents of the Performance Profile to inform instructional planning. One strategy included using the performance level descriptors to know the skills typical of students in the performance level. Another strategy involved the use of the conceptual area bar graphs to know the percent of skills mastered in each area of related content standards. Finally, teachers described relating the information on the DLM score reports to information obtained from local assessment results as an additional source of information about student performance.

Teachers also described using the performance level descriptors included on the Perofrmance Profile for IEP goal planning. One teacher mentioned using the conceptual area bar graphs combined with results from a district assessment to frame IEP goals. The teacher stated, "We have a district assessment in the fall, winter, and spring, so in the fall, they provide a report and summary. I try to see if there is still a deficiency based on the DLM [results from] the spring and in the new report in the fall to see if that is an area that there's still a weakness, and if there is then that's definitely something I would spend more time on." However, other teachers indicated the report was not specific enough to frame IEP goals, particularly because the year-end assessment model does not capture growth over time. These teachers reported using data from other progress monitoring and district tools to inform IEP goal development.

In instances where multiple students were assessed in the same grade, teachers described the benefit of being able to plan instructional groupings from reports. One teacher expressed a desire for an aggregated report that made instructional groupings more clear, particularly around standards and levels students were working on in common.

9.4.2.3. Talking With Parents

Teachers highlighted the importance of understanding the assessment and student results when talking to parents. As one teacher stated, "That first year…I wasn't able to give the parents a lot other than, 'Here's your score report,' " and indicate the performance level. By the second year, the teacher mentioned knowing more about the content measured by the assessment. She stated, "I know more about where they are going and what they're doing so I can share that with parents….This is the academic focus, this is what we're hoping they get out of reading that aligns with their IEP goals, which aligns with the DLM testing. It is a better conversation about why this testing format is."

For parents of students new to the DLM assessment system, teachers reported some confusion about the reports. "Parents seemed a little confused because they had never seen a report before. So I don't think they really knew exactly what they were looking at since it was something so new presented to them." The teacher went on to share, "We just went over exactly what was on the report step by step. I pointed out some of the IEP objectives and how they were related to what was on the report."

Teachers reported that parents seemed unsure how the student performance level was determined. As one stated, "The mathematical formula was not very cut and dry, so it was very difficult to explain it to them." While the Performance Profile contains narrative text in addition to a visual representation of performance levels, these teacher comments indicate the report likely needs to go further in explaining how the performance level was determined to be informative to parents.

Overall, teachers reported that, with a few exceptions, parents did not ask questions about the DLM assessment or score reports, so the extent of information parents received about the assessment and its use for instruction in the subsequent year was dependent upon what the teacher offered. As one teacher indicated, "Unfortunately, I just don't think that our parents know what to ask. They're not educated about the test. They only have the information that I give them and so, this year I was able to give them more, but will I be able to give them even more information at the end of the year when we transition their child off to middle school? Oh yeah, because I've looked at it better so I could give more information."

Findings from the focus groups provide some evidence of appropriate use of DLM assessment results for informing instruction. However, the challenge of identifying teachers who used reports in the subsequent academic year indicates a need for further instructional supports around appropriate use of results. Next steps are described in Chapter 11.

9.5. Conclusion

This chapter presents additional studies as evidence to support the overall validity argument for the DLM Alternate Assessment System. The studies are organized into categories where available (content, response process, internal structure, external variables, and consequences of testing), as defined by the *Standards for Educational and Psychological Testing* (AERA et al., 2014), the professional standards used to evaluate educational assessments.

The final chapter of this manual, Chapter 11, references evidence presented through the technical 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; DLM Consortium, 2017b), 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 2017–2018 instructional professional development. No changes were made to training in 2017–2018.

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[12]. No new modules were added in 2017–2018.

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.

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.

[12] http://dlmpd.com

10.1.1. Professional Development Participation and Evaluation

As reported in Table 10.1, a total of 9,238 modules were completed in the self-directed format from September 1, 2017, to August 31, 2018. 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 2017–2018 by Educators in DLM States and Other Localities (*N* = 9,238)

State	Self-directed modules completed
Kansas	2,410
Colorado	1,692
Wisconsin	678
Rhode Island	627
New Jersey	417
Iowa	406
Missouri	353
Illinois	294
Utah	210
Oklahoma	186
New York	170
Delaware	40
Maryland	40
New Hampshire	36
West Virginia	7
Alaska	2
North Dakota	2
Non-DLM states and other locations	1,668

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 2017–2018 year, educators completed the evaluation questions 85% of the time. The responses were consistently positive, as illustrated in Table 10.2. Across all modules approximately 80% of respondents either agreed or strongly agreed with each statement.

Table 10.2. Response Rates and Rate of *Agree* or *Strongly Agree* on 2017–2018 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	211	.85	.80	.78	.79	.78
Basic Geometric Shapes	260	.73	.68	.67	.66	.68
Beginning Communicators	433	.84	.80	.80	.79	.80
Calculating Accurately with Addition	190	.77	.73	.71	.73	.73
Calculating Accurately With Division	85	.94	.88	.84	.87	.88
Calculating Accurately With Multiplication	108	.91	.85	.83	.85	.84
Calculating Accurately With Subtraction	93	.87	.83	.83	.84	.83
Common Core Overview	315	.92	.86	.77	.83	.85
Composing and Decomposing Shapes and Area	88	.86	.80	.80	.78	.80
Composing, Decomposing, and Comparing Numbers	174	.90	.87	.86	.86	.86
Core Vocabulary and Communication	374	.89	.86	.83	.81	.83

Table 10.2. Response Rates and Rate of *Agree* or *Strongly Agree* on 2017–2018 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. (%)**
Counting and Cardinality	244	.91	.87	.85	.86	.86
DLM Claims and Conceptual Areas	177	.93	.84	.77	.81	.80
DLM Essential Elements Overview	544	.82	.76	.76	.74	.76
DR-TA and Other Text Comprehension Approaches	109	.84	.83	.81	.79	.80
Effective Instruction in Mathematics	155	.90	.85	.83	.85	.86
Emergent Writing	352	.83	.82	.81	.80	.81
Exponents and Probability	46	.89	.83	.78	.78	.80
Forms of Number	171	.55	.51	.48	.49	.49
Fraction Concepts and Models Part I	47	.77	.72	.70	.68	.70
Fraction Concepts and Models Part II	49	.76	.69	.65	.63	.67
Functions and Rate	19	.79	.63	.63	.63	.68

Table 10.2. Response Rates and Rate of *Agree* or *Strongly Agree* on 2017–2018 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. (%)**
Generating Purposes for Reading	112	.86	.83	.82	.81	.83
IEPs Linked to DLM Essential Elements	288	.82	.77	.74	.74	.78
Measuring and Comparing Lengths	122	.86	.81	.80	.79	.81
Organizing and Using Data to Answer Questions	58	.84	.79	.74	.76	.76
Patterns and Sequences	32	.66	.56	.47	.47	.53
Perimeter, Volume, and Mass	54	.76	.72	.70	.72	.70
Place Value	71	.63	.58	.59	.58	.59
Predictable Chart Writing	136	.92	.89	.88	.88	.90
Principles of Effective Instruction ELA	198	.82	.81	.79	.79	.80
Properties of Lines and Angles	37	.89	.78	.78	.76	.78
Shared Reading	480	.86	.82	.80	.80	.81
Speaking and Listening	207	.84	.82	.82	.80	.81

Table 10.2. Response Rates and Rate of *Agree* or *Strongly Agree* on 2017–2018 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. (%)**
Standards of Mathematical Practice	9	1.00	1.00	1.00	1.00	1.00
Strategies and Formats for Presenting Ideas	138	.86	.80	.78	.78	.76
Supporting Participation in Discussions	146	.82	.78	.77	.75	.77
Symbols	80	.90	.85	.82	.82	.84
Teaching Text Comprehension: Anchor-Read-Apply	245	.84	.80	.78	.78	.78
The Power of Ten-Frames	61	.92	.87	.84	.84	.85
Time and Money	65	.82	.74	.68	.68	.74
Unitizing	65	.60	.54	.51	.49	.52
Units and Operations	41	.80	.73	.71	.71	.73
Universal Design for Learning	462	.94	.91	.90	.87	.91
Who are Students with Significant Cognitive Disabilities?	1,157	.90	.89	.82	.84	.86

Table 10.2. Response Rates and Rate of *Agree* or *Strongly Agree* on 2017–2018 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 Information Texts	83	.86	.78	.76	.78	.76
Writing With Alternate Pencils	335	.89	.87	.85	.85	.84
Writing: Getting Started in Narrative Writing	26	.81	.73	.73	.73	.77
Writing: Getting Started Writing Arguments	18	.78	.61	.61	.56	.67
Writing: Production and Distribution	27	.89	.74	.74	.74	.78
Writing: Research and Range of Writing	118	.90	.86	.82	.86	.87
Writing: Text Types and Purposes	123	.81	.80	.76	.76	.80
Total	*9,238*	*.85*	*.81*	*.79*	*.79*	*.80*

Note: SWSCDs = students with significant cognitive disabilities.

To better understand teacher use of modules, the spring 2018 teacher survey asked teachers to indicate how many professional development modules they had completed in the last two years. Results are summarized in Table 10.3. Around 36% of respondents indicated they had completed between one and five modules in the last two years, while 11% indicated they had not completed any modules.

Table 10.3. Number of Professional Development Modules Completed in the Last 2 Years

Number of modules	***n***	**%**
0	6,407	11.1
1-5	20,529	35.7
6-10	8,840	15.4
11-15	4,596	8.0
16-20	3,287	5.7
>20	3,705	6.4
Missing	10,176	17.7

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 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. 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 fourth operational administration year in 2017–2018. This technical manual update provides updated evidence from the 2017–2018 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; DLM Consortium, 2017b). 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 2017–2018 year
2	Not updated for 2017–2018
3, 4, 10	Provides procedural evidence collected during 2017–2018 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 2017–2018
7, 8	Describes results and analyses from the fourth 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 2017–2018 focused on specific topics related to validity and to evaluate the score propositions and intended uses

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 2017–2018 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 2017–2018

Claim	Sources of evidence*				
	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, 7.1, 7.2, 9.1	4.1, 4.2, 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 2017–2018
3.4	3	Field Testing
4.1	4	User Experience With the DLM System
4.2	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, 2017–2018 was the fourth year the DLM Alternate Assessment System was operational. While the 2017–2018 assessments were carried out in a manner that supports the validity of inferences made from results for the intended purposes, the DLM Alternate Assessment Consortium is committed to continual improvement of assessments, teacher and student experiences, and technological delivery of the assessment system. Through formal research and evaluation as well as informal feedback, some improvements have already been implemented for 2018–2019. This section describes significant changes from the third to fourth year of operational administration, as well as examples of improvements to be made during the 2018–2019 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 2017–2018.

Based on an ongoing effort to improve KITE® system functionality during 2017–2018, Educator Portal was enhanced to support creation and delivery of data files and score reports to maintain faster delivery timelines. This included automated creation of all aggregated reports provided at the class, school, district, and state levels; and delivery of the final General Research File in the interface.

The validity evidence collected in 2017–2018 expands upon the data compiled in the first three 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 products. 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-survey responses also provided evidence based on consequences of testing, as well as a summary of findings from score-report focus groups collecting teacher feedback on their use of summative reports in the subsequent academic year. Studies planned for 2018–2019 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 2018–2019 and beyond. The manual identifies some areas for further investigation.

DLM staff members are planning several studies for spring 2019 to collect data from teachers in the DLM Consortium states. Teachers will be recruited to participate in a study to collect additional evidence based on other variables, whereby teacher ratings of student mastery will be correlated with model-derived mastery. Finally, teacher-survey data collection will also continue during spring 2019 to obtain the third year of data for longitudinal survey items as further validity evidence.

Teachers will continue to compile and rate student writing samples to expand the collection and evaluation of interrater agreement of writing products. The process for collecting test administration observations is also being updated to expand the collection of protocols to a more representative sample. State partners will continue to collaborate with additional data collection as needed.

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.

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 expressive communication subgroups, as identified by the First Contact survey.

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. doi:10.1037/1040-3590.6.4.284[13]

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). *2014–15 Technical Manual—Year-End Model*. University of Kansas, Center for Educational Testing and Evaluation. Lawrence, KS.

Dynamic Learning Maps Consortium. (2017a). *2015–16 Technical Manual Update—Year-End Model*. University of Kansas, Center for Educational Testing and Evaluation. Lawrence, KS.

Dynamic Learning Maps Consortium. (2017b). *2016–17 Technical Manual Update—Year-End Model*. 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*. University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS). Lawrence, KS.

Dynamic Learning Maps Consortium. (2017d). *Educator Portal User Guide*. University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS). Lawrence, KS.

Dynamic Learning Maps Consortium. (2017e). *Test Administration Manual 2017–2018*. University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS). Lawrence, KS.

Dynamic Learning Maps Consortium. (2018). *2017–2018 Technical Manual Update—Science*. 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.

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.

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

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.

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

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. doi:10.1007/s00357-013-9129-4[14]

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

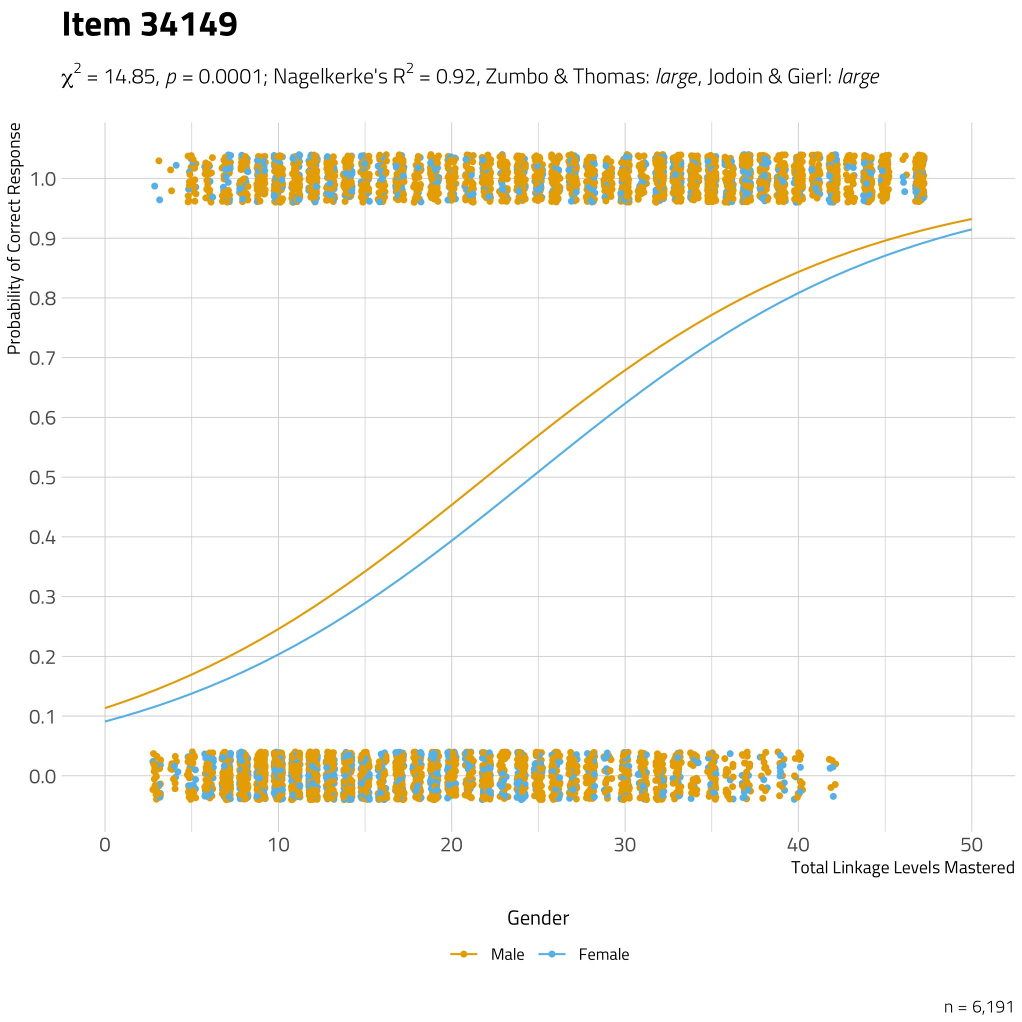
[14] https://dx.doi.org/10.1007/s00357-013-9129-4

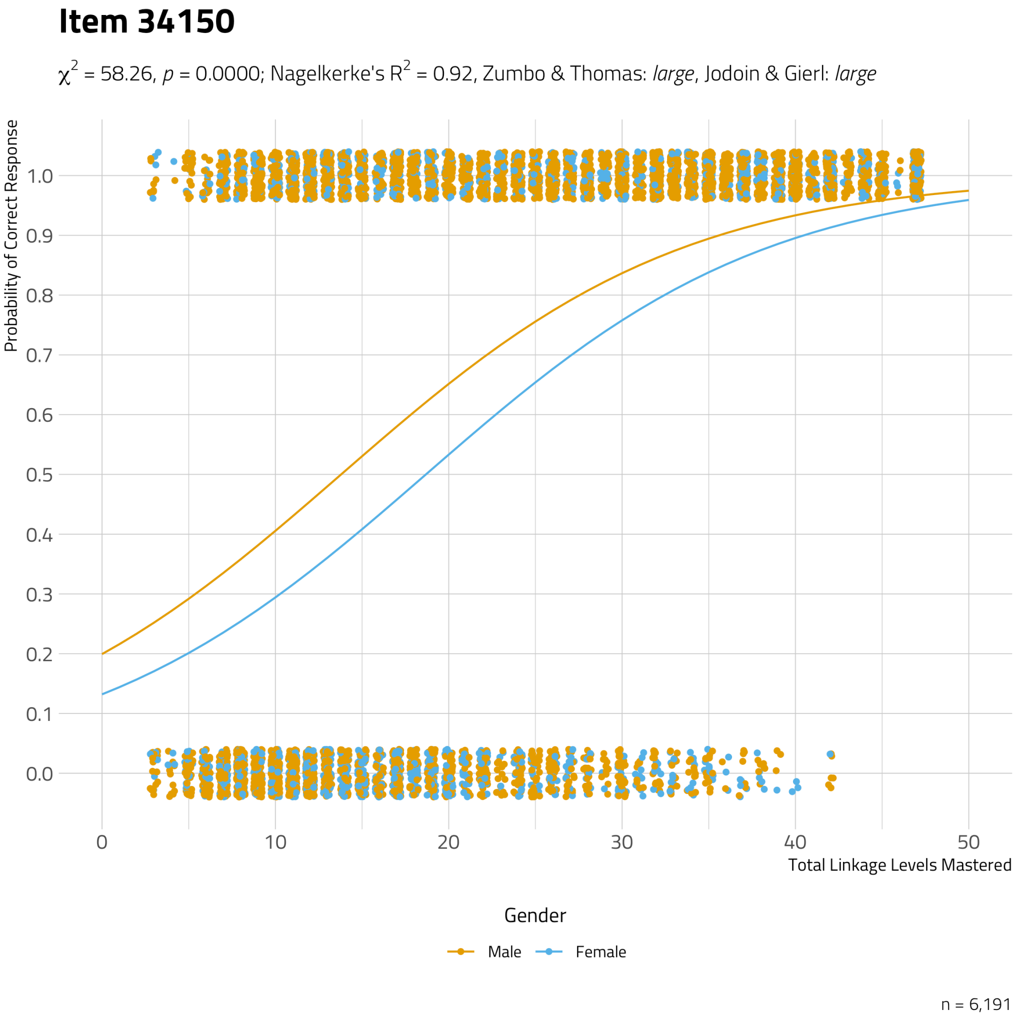
A. Differential Item Functioning Plots

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.



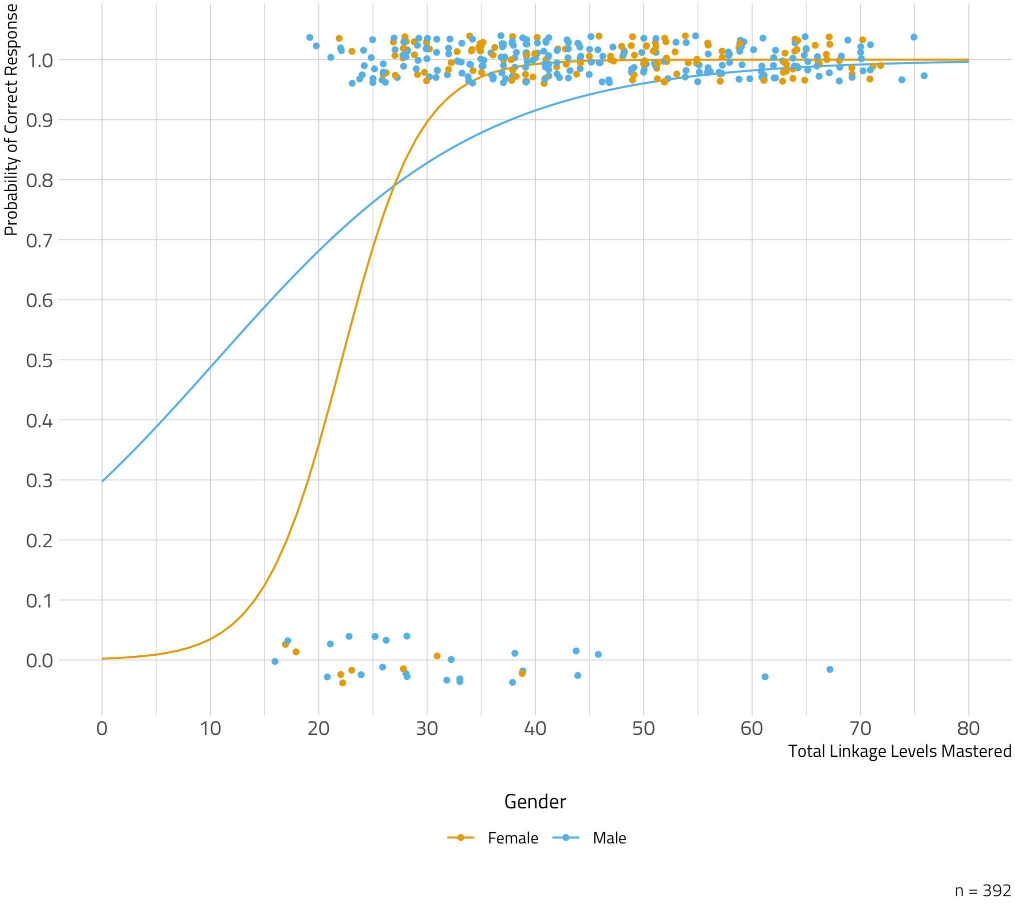


A.2. Combined Model

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.

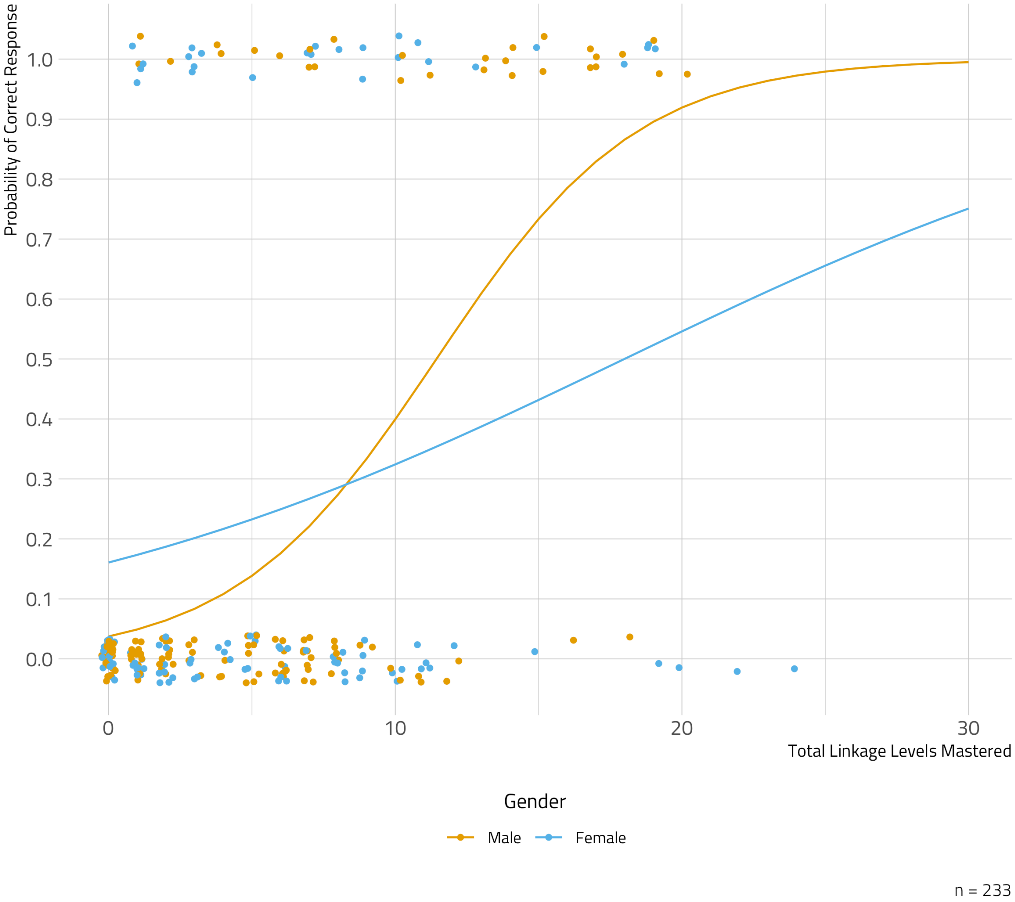
Item 28628

χ^2 = 8.63, *p* = 0.0033; Nagelkerke's R^2 = 0.04, Zumbo & Thomas: *negligible*, Jodoin & Gierl: *moderate*



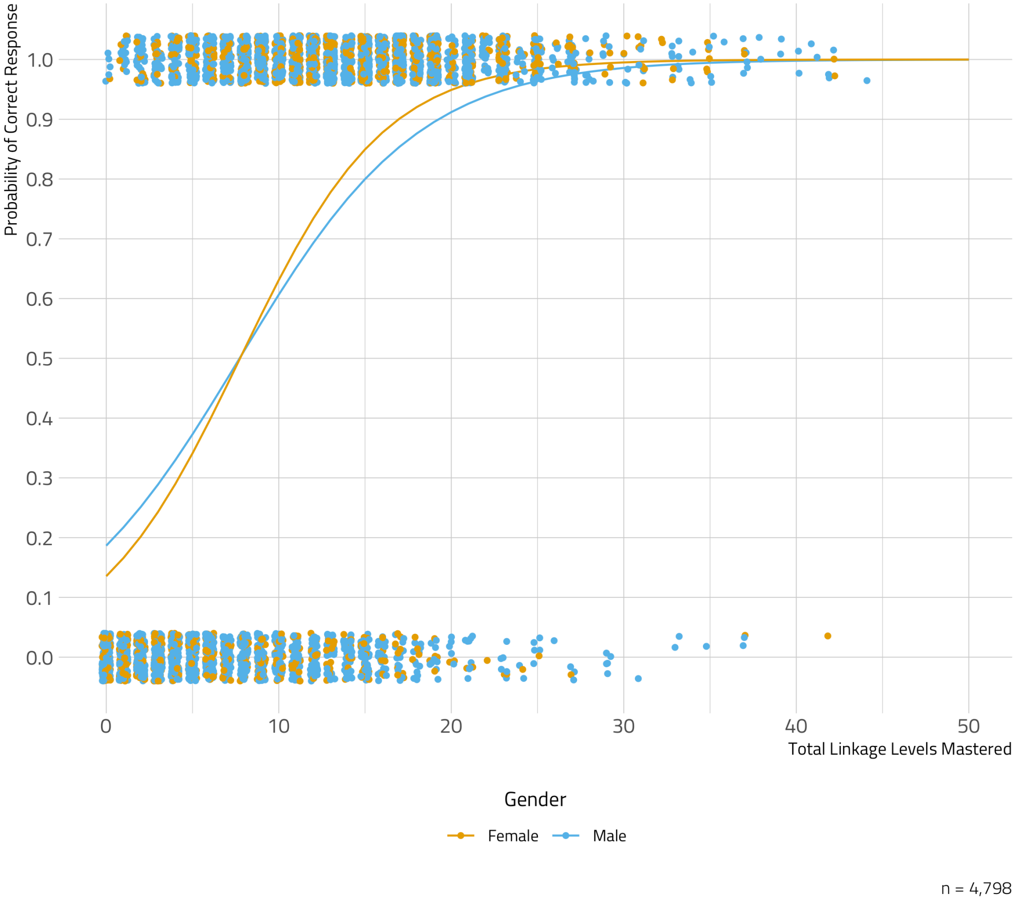
Item 39680

χ^2 = 9.26, *p* = 0.0023; Nagelkerke's R^2 = 0.04, Zumbo & Thomas: *negligible*, Jodoin & Gierl: *moderate*



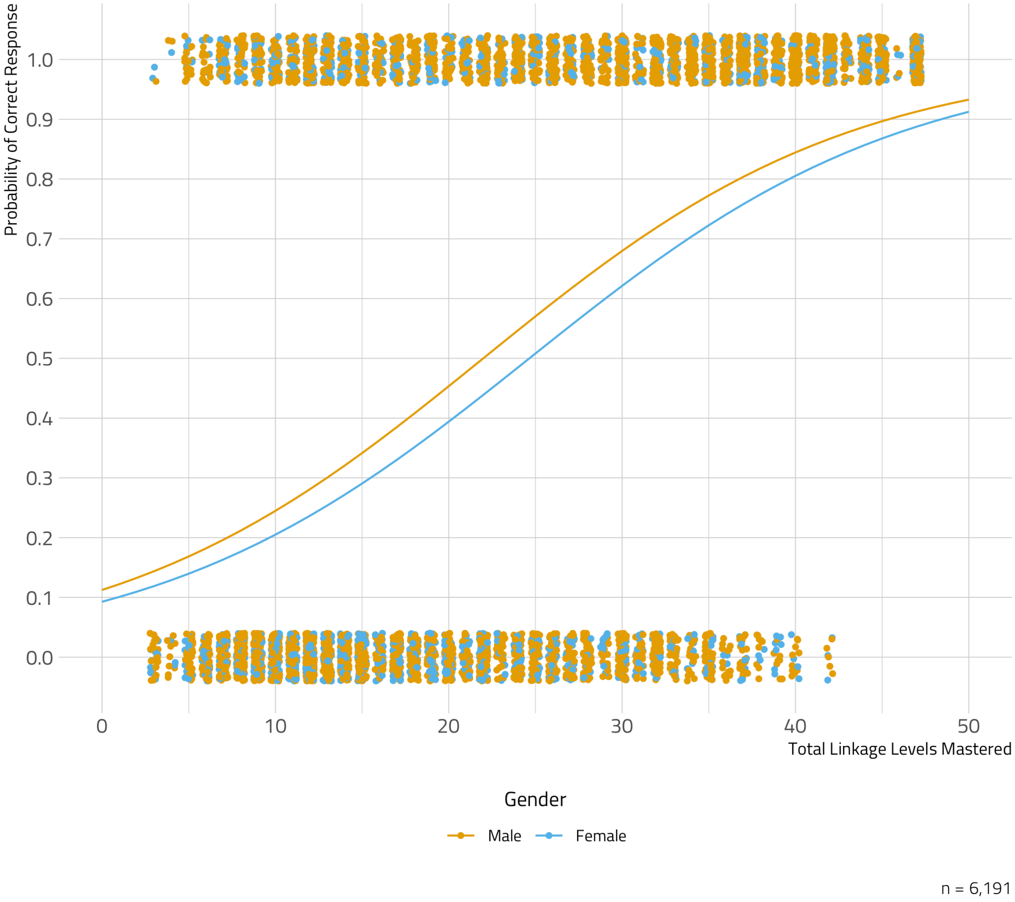
Item 30547

χ^2 = 11.97, p = 0.0005; Nagelkerke's R^2 = 0.93, Zumbo & Thomas: *large*, Jodoin & Gierl: *large*



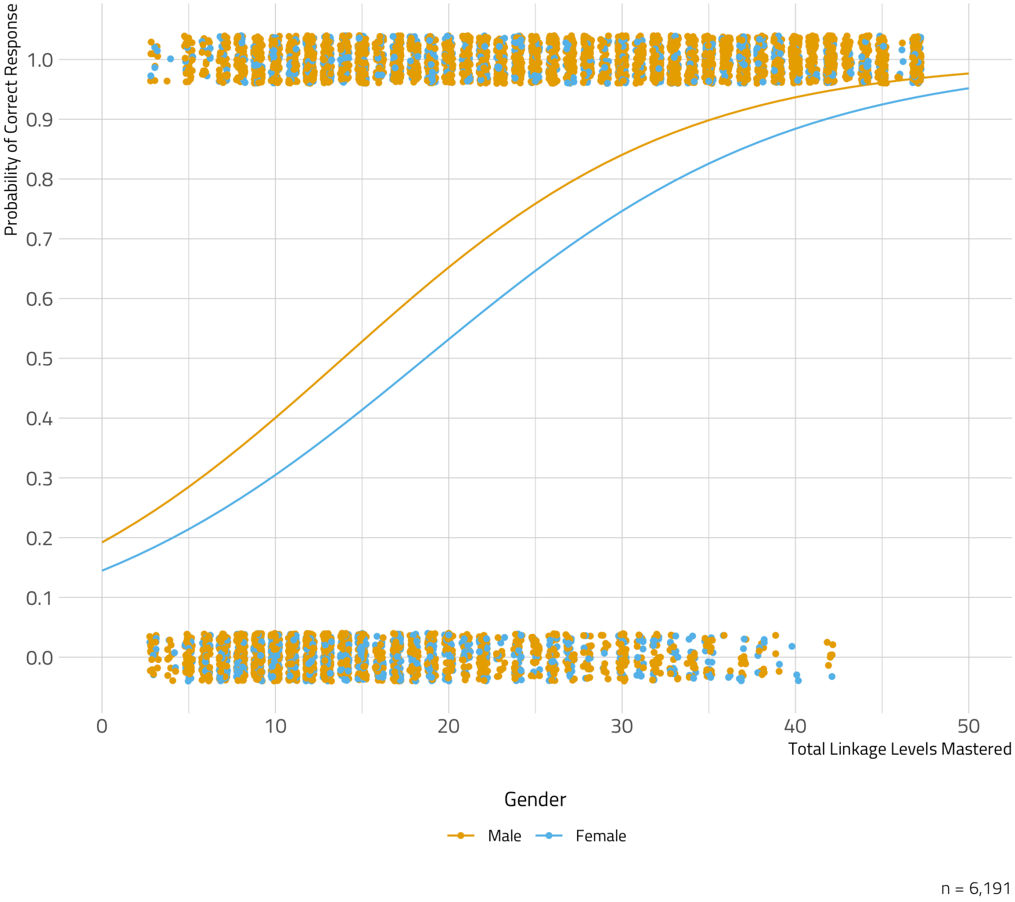
Item 34149

χ^2 = 14.91, *p* = 0.0001; Nagelkerke's R^2 = 0.92, Zumbo & Thomas: *large*, Jodoin & Gierl: *large*



Item 34150

χ^2 = 59.71, *p* = 0.0000; Nagelkerke's R^2 = 0.92, Zumbo & Thomas: *large*, Jodoin & Gierl: *large*


Probability of Correct Response
Total Linkage Levels Mastered
Gender
Male
Female
n = 6,191