DYNAMIC®
LEARNING MAPS

2016-17 Technical Manual Update

Year-End

December 2017

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

Acknowledgements

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

Many contributors made the writing of this technical manual update possible. We are especially grateful for the contributions of the members of the Dynamic Learning Maps (DLM) Technical Advisory Committee who graciously provided their expertise and feedback. Members of the Technical Advisory Committee include:

Jamal Abedi, Ph.D., *University of California-Davis*
Russell Almond, Ph.D., *Florida State University*
Greg Camilli, Ph.D., *Rutgers University*
Karla Egan, Ph.D., *Independent Consultant*
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*

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

Amy Clark, Ph.D., *Psychometrician Senior*
Brooke Nash, Ph.D., *Psychometrician Senior*
William Jacob Thompson, *Psychometrician Assistant*
Meagan Karvonen, Ph.D., *Director*

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

Sue Bechard
Brianna Beitling
Karen Erickson
Lisa Harkrader
Neal Kingston
Allison Lawrence
Lee Ann Mills
Michael Muenks
Haley Olson
Qianqian Pan
Michelle Shipman
Russell Swinburne Romine
Susan K. Thomas
Xiaolin Wang
Lisa Weeks

Table of Contents

List of Tables

List of Figures

I. INTRODUCTION

During the 2016–2017 academic year, the Dynamic Learning Maps® (DLM®) Alternate Assessment System offered assessments of student achievement in mathematics, English language arts (ELA), and science for students with the most significant cognitive disabilities in grades 3 through 8 and high school. Because science was implemented on a separate timeline than ELA and mathematics, a separate technical manual update was prepared for science for the 2016–2017 year (see Dynamic Learning Maps Consortium, 2017b).

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 support inferences about student achievement in the given content area. Results provide information that can be used to guide instructional decisions as well as information appropriate for use with state accountability programs.

The DLM Alternate Assessment System is based on the core belief that all students should have access to challenging, grade-level content. Online DLM assessments give students with the most significant cognitive disabilities opportunities to demonstrate what they know in ways that traditional paper-and-pencil, multiple-choice assessments cannot. The DLM alternate assessment provides optional, instructionally embedded testlets that are available for use in day-to-day instruction. A year-end assessment is administered in the spring, and results from that assessment are reported for state accountability purposes and programs. This design is referred to as the year-end model and is one of two models for the DLM Alternate Assessment System.[1]

A complete technical manual was created for the first year of operational administration, 2014–2015, and a technical manual update provided in 2015–2016. This technical manual provides updates for the 2016–2017 administration; therefore, only sections with updated information are included in this manual. For a complete description of the DLM assessment system, refer to the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b).

I.1. BACKGROUND

In 2016–2017, DLM assessments were administered to students in 16 states: Alaska, Colorado, Illinois, Iowa, Kansas, Maryland, Missouri, New Hampshire, New Jersey, New York, North Dakota, Oklahoma, Utah, Vermont, West Virginia, and Wisconsin.

One state partner, Maryland, did not administer operational assessments in ELA and mathematics in 2016–2017.

In 2016–2017, the Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS) at the University of Kansas continued to partner with the Center for Literacy and Disability

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

Studies at the University of North Carolina at Chapel Hill and the Center for Research Methods and Data Analysis at the University of Kansas. The project was also supported by a Technical Advisory Committee (TAC).

I.2. Assessments

Assessment blueprints consist of 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 model.

- **Integrated model.** In the first of two general testing windows, instructionally embedded assessments occur throughout the fall, winter, and early spring. Educators have some choice of which EEs to assess, within constraints. For each EE, the system recommends a linkage level for assessment and the educator may accept the recommendation or choose another linkage level. During the second testing window in the spring, all students are reassessed on several EEs on which they were taught and assessed earlier in the year. During the spring window the system assigns the linkage level based on student performance on previous testlets; the linkage level for each EE may be the same as or different from what was assessed during the instructionally embedded window. At the end of the year, results used for summative purposes are based on mastery estimates for linkage levels for each EE (including performance on all instructionally embedded and spring testlets). The pools of operational assessments for the instructionally embedded and spring windows are separate. In 2016–2017, the states participating in the integrated model included Iowa, Kansas, Missouri, North Dakota, and Vermont.
- **Year-end model.** In a single operational testing window in the spring, all students take testlets that cover the whole blueprint. Each student is assessed at one linkage level per EE. The linkage level for each testlet varies based on student performance on the previous testlet. The assessment results reflect the student's performance and are used for accountability purposes each school year. The instructionally embedded assessments are available during the school year but are optional and do not count toward summative results. In 2016–2017, the states participating in the year-end model included Alaska, Colorado, Illinois, New Hampshire, New Jersey, New York, Oklahoma, Utah, West Virginia, and Wisconsin, as well as the Miccosukee Indian School.

Information in this manual is common to both models wherever possible and is specific to the Year-End model where appropriate. A separate version of the Technical Manual exists for the Integrated model.

I.3. TECHNICAL MANUAL OVERVIEW

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

Chapter I provides an overview of the assessment and administration for the 2016–2017 academic year and a summary of contents of the remaining chapters. While subsequent chapters describe the essential components of the assessment system separately, several key topics are addressed throughout this manual, including accessibility, and validity.

Chapter II was not updated for 2016–2017. See the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b) for a description of the process by which the DLM maps were developed.

Chapter III outlines procedural evidence related to test content. It includes summaries of external reviews for content, bias, and accessibility. The final portion of the chapter describes the operational and field test content available for 2016–2017.

Chapter IV provides an overview of the fundamental design elements that characterize test administration and how each element supports the DLM theory of action. The chapter provides updated evidence for administration incidents, as well as teacher survey results collected during 2016–2017 regarding educator experience, administration of instructionally embedded assessments, and system accessibility.

Chapter V provides a summary of the psychometric model that underlies the DLM project and describes the process used to estimate item and student parameters from student test data. The chapter includes a summary of calibrated parameters, mastery assignment for students, and evidence of model fit. For a complete description of the modeling method, see *2015–2016 Technical Manual – Year-End Model* (DLM Consortium, 2017c).

Chapter VI was not updated for 2016–2017. See the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b) for a description of the methods, preparations, procedures, and results of the standard setting meeting and the follow-up evaluation of the impact data and cut points based on the 2014–2015 operational assessment administration.

Chapter VII reports the 2016–2017 operational results, including student participation data. The chapter details the percentage of students at each performance level (impact); subgroup performance by gender, race, ethnicity, and English learner status; and the percentage of students who showed mastery at each linkage level. Finally, the chapter provides descriptions of score reports, data files, and quality control methods.

Chapter VIII focuses on reliability evidence, including a summary of the methods used to evaluate assessment reliability and results by performance level, content area, conceptual area, EE, linkage level, and conditional linkage level. For a complete description of the reliability background and methods, see *2015–2016 Technical Manual – Year-End Model* (DLM Consortium, 2017c).

Chapter IX describes additional validation evidence 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 X describes the professional development that was offered across the DLM Consortium in 2016–2017. Participation rates and evaluation results from 2016–2017 instructional professional development are included.

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

II. MAP DEVELOPMENT

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

III. ITEM AND TEST DEVELOPMENT

Chapter III of the *2014–2015 Technical Manual – Year-End Model* (Dynamic Learning Maps® [DLM®] Consortium, 2016b) describes general item- and test-development procedures. This chapter provides an overview of updates to item and test development for the 2016–2017 academic year. The first portion of the chapter provides a supplemental summary of item and testlet information, followed by the 2016–2017 external review of items and testlets for content, bias, and accessibility. The next portion of the chapter describes the operational assessments for 2016–2017, followed by a section describing field tests administered in 2016–2017.

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, 2016b).

III.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 writers for the 2016–2017 year and information on ELA writing testlets. The section on writing testlets provides expanded information about practices in effect beginning in 2014–2015. This material was included in the 2016–2017 update at stakeholder request. For a complete summary of item and testlet development procedures that began in 2014–2015 and were implemented in 2015–2016, see Chapter III of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b).

III.1.A. Item Writers

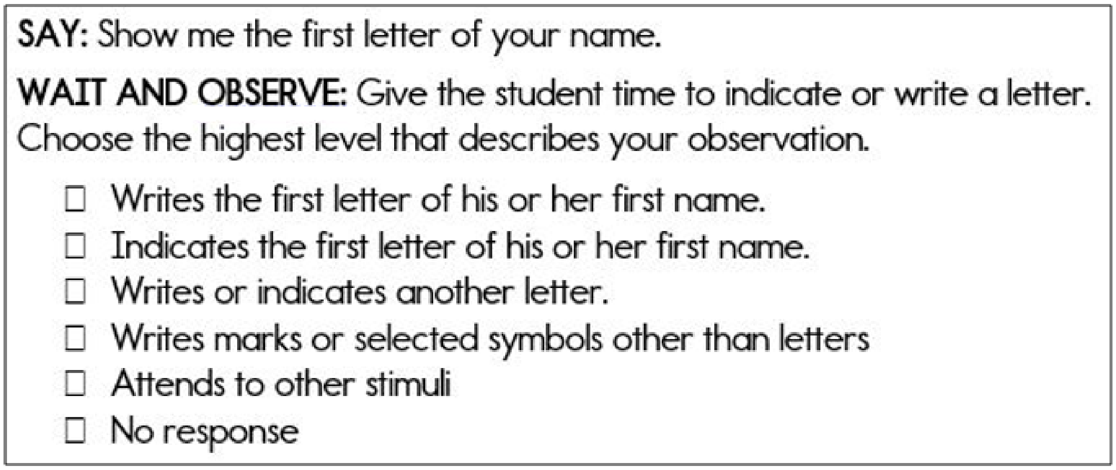
Development of DLM items and testlets began in the summer of 2013. Additional items and testlets were developed during 2014. During these years, most item writing occurred during summer events in which content and special education specialists worked on-site in Lawrence, Kansas, to develop DLM assessments. For the 2016–2017 year, only limited items were written to replenish the pool. One item writer was retained from the previous item writing events. Three graduate research assistants also received item writer training and wrote testlets. In total, four item writers contributed to testlets for the 2016–2017 year, all in ELA. For more information on item writer characteristics, see Chapter III of the *2015–2016 Technical Manual – Year-End Model* (DLM Consortium, 2017c).

III.1.B. English Language Arts Writing Testlets

From 2014–2015 through 2016–2017, every grade level had an emergent and conventional writing testlet available, each of which measures several Essential Elements (EEs). Writing testlets include EEs in the Writing strand, and in some grades, EEs in the Language strand. Emergent writing testlets measure the Initial Precursor and Distal Precursor linkage levels, while conventional writing testlets measure the Proximal Precursor, Target, and Successor linkage levels. Because writing testlets measure multiple EEs and linkage levels, the structure of writing testlets differs from that of other testlets.

All writing testlets are teacher-administered. The testlet engagement activity is followed by items that require the test administrator to evaluate the student's writing process. Some writing testlets also evaluate the student's writing product. Item types are either single-select multiple choice or multi-select multiple choice. Both item types ask test administrators to select a response from a checklist of possible responses that best describes what the student did or produced as part of the writing testlet.

Items that assess student writing processes are ratings of the test administrator's observations of the student as they complete items in the testlet. Figure 1 shows an example of a process item from an emergent writing testlet focused on letter identification in support of writing the student's first name. The construct assessed in this item is the student's ability to identify the first letter of his or her own name. In the example, either "Writes the first letter of his or her own name" or "Indicates the first letter of his or her own name" is scored as a correct response (Figure 1). The inclusion of multiple, correct response options at different levels was designed to ensure that this testlet was accessible to emergent writers who were beginning to write letters and emergent writers who had not yet developed writing production skills but were still able to identify the first letter of their first name. As such, each response option is associated with a different EE and linkage level.



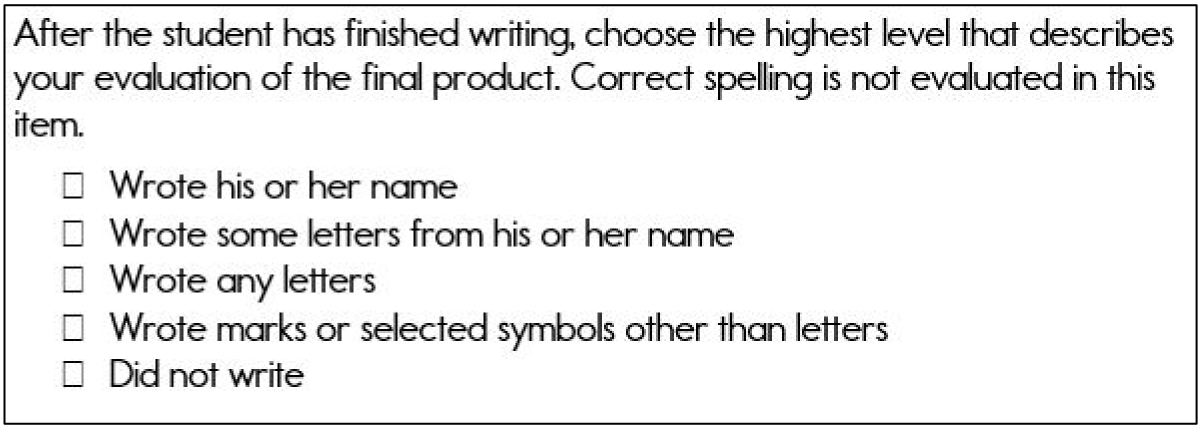
SAY: Show me the first letter of your name.

WAIT AND OBSERVE: Give the student time to indicate or write a letter. Choose the highest level that describes your observation.

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

Figure 1. Example of English language arts emergent writing item focused on process.

Items that assess writing products are the test administrator's ratings of the product created by the student as a result of the writing processes completed in the administration of the testlet. Figure 2 provides an example of an item that evaluates a student's writing product. For some product items, administrators choose all the responses in the checklist that apply to the student's writing product. A complete description of writing testlets can be found in Chapter III of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b).



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

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

Figure 2. Example of English language arts conventional writing item focused on product.

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

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

During spring 2017 administration, writing products were collected to evaluate consistency in scoring across teachers. For a full description of this study, see Chapter IX of this manual.

III.2. External Reviews

The purpose of external review is to evaluate items and testlets developed for the DLM Alternate Assessment System. Using specific criteria established for DLM assessments, reviewers decided whether to recommend that content be accepted, revised, or rejected. Feedback from external reviewers was used to make final decisions about assessment items before they were field-tested.

Overall, the process and review criteria for external review in 2016–2017 remained the same as those used in the previous two review cycles. Minor changes were made, including using fewer reviewers who completed more assignments.

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

In 2016–2017, volunteers completed a survey to express interest in serving as external review panelists. The Qualtrics survey captured demographic information and information about their education and experience. These data were then used to identify panel types (content, bias and sensitivity, and accessibility) for which the volunteer would be eligible. A total of 27 people from year-end model states completed the required training, 24 of whom were placed on external review panels.

Of the 24 reviewers placed on panels, 16 completed reviews. Each reviewer was assigned to one of the three panel types. There were six ELA reviewers: one on an accessibility panel, three on content panels, and two on bias and sensitivity panels. There were seven mathematics reviewers: two on accessibility panels, three on content panels, and two on bias and sensitivity panels. Also, three power reviewers and two hourly reviewers reviewed all three panel types as needed for each content area.

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

Table 1 presents the professional roles reported by the 2016–2017 volunteer reviewers. Reviewers who reported other roles included educational services staff, education specialists, state administrators, and site supervisors.

Table 1. Professional Roles of External Reviewers

	English language arts (*N* = 6)		**Mathematics* (*N* = 7)**	
Role	***n***	**%**	***n***	**%**
Classroom teacher	3	50.0	4	57.1
District staff	1	16.7	0	0.0
Other	2	33.3	2	28.6

Note. *One reviewer did not provide a professional role.

Reviewers had varying experience teaching students with the most significant cognitive disabilities. ELA reviewers had a median of 10.5 years of experience, with a minimum of 5 and a maximum of 29 years of experience. Mathematics reviewers had a median of 7 years of experience teaching students with the most significant cognitive disabilities, with a minimum of 3 and a maximum of 25 years of experience.

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

Table 2. Population Density for Schools of External Reviewers

	English language arts (*N* = 6)		**Mathematics (*N* = 7)**	
Population density	***n***	**%**	***n***	**%**
Rural	4	66.7	1	14.3
Suburban	1	16.7	4	57.1
Urban	1	16.7	2	28.6

Review assignments were given throughout the year. Reviewers were notified by email each time they were assigned collections of testlets. Reviewers were notified by 1½ to 2 hours to complete. In most cases, reviewers had between 10 days and 2 weeks to complete an assignment.

III.2.B. Results of Reviews

Content externally reviewed during the 2016–2017 academic year was either included in the spring testing window or examined for the upcoming 2017–2018 school year. In ELA, 91 items and 22 testlets were reviewed. In mathematics, 147 items and 27 testlets were reviewed, based on availability of content that had already been developed. For both content areas, 100% of items were rated as *accept* or *revise*. No content was recommended for rejection. A summary of the test development team decisions and outcomes based on external review recommendations is provided here.

III.2.B.i. Test Development Team Decisions

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

DLM test development teams documented the decision category applied by external reviewers to each item and testlet. Following this process, test development teams made a final decision to accept, revise, or reject each of the items and testlets. Table 3 summarizes the test development decisions following their review. The ELA team retained 100% of items and testlets sent out for external review; no items or testlets were revised. The mathematics team also retained 100% of items and testlets sent out for external review. Of the items and testlets that were revised, all required only minor changes (e.g., minor rewording but concept remained unchanged), as opposed to major changes (e.g., stem or option replaced). The mathematics team made 48 minor revisions to items and 17 minor revisions to testlets.

Table 3. Test Development Team Decisions

	English language arts				**Mathematics**			
	Items (*N* = 91)		**Testlets (*N* = 22)**		**Items (*N* = 147)**		**Testlets (*N* = 27)**	
Decision	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Accept	91	100.0	22	100.0	99	67.3	10	37.0
Revise	0	0.0	0	0.0	48	32.7	17	63.0
Reject	0	0.0	0	0.0	0	0.0	0	0.0

III.3. Operational Assessment Items for 2016–2017

A total of 969,638 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 the testing window counted toward the total test sessions.

Testlets were made available for operational testing in 2016–2017 based on the 2015–2016 operational pool and the promotion of testlets field-tested during 2015–2016 to the operational pool following their review. Table 4 and Figure 9 summarize the total number of operational testlets by content area for 2016–2017 for ELA and mathematics, respectively. There were 717 operational testlets available across grades and content areas. This total also included 219 (72 mathematics, 147 ELA) sections for which both a general version and a version for students who are blind or visually impaired were available.

Table 4. 2016–2017 English Language Arts Operational Testlets (*N* = 328)

Grade	***n***
3	48
4	47
5	43
6	36
7	34
8	28
9	31
10	36
11	25

Table 5. 2016–2017 Mathematics Operational Testlets (*N* = 389)

Grade	***n***
3	38
4	47
5	41
6	42
7	40
8	41
9	48
10	45
11	47

Similar to prior years, *p* values were calculated for all operational items to summarize information about item difficulty.

Figure 3 and Figure 4 include the *p* values for each operational item for ELA and mathematics, respectively. To prevent items with small sample size 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 more items falling in the higher bin (*p* value) ranges.

Writing items were omitted from this plot because scoring occurred at the option level rather than item level.

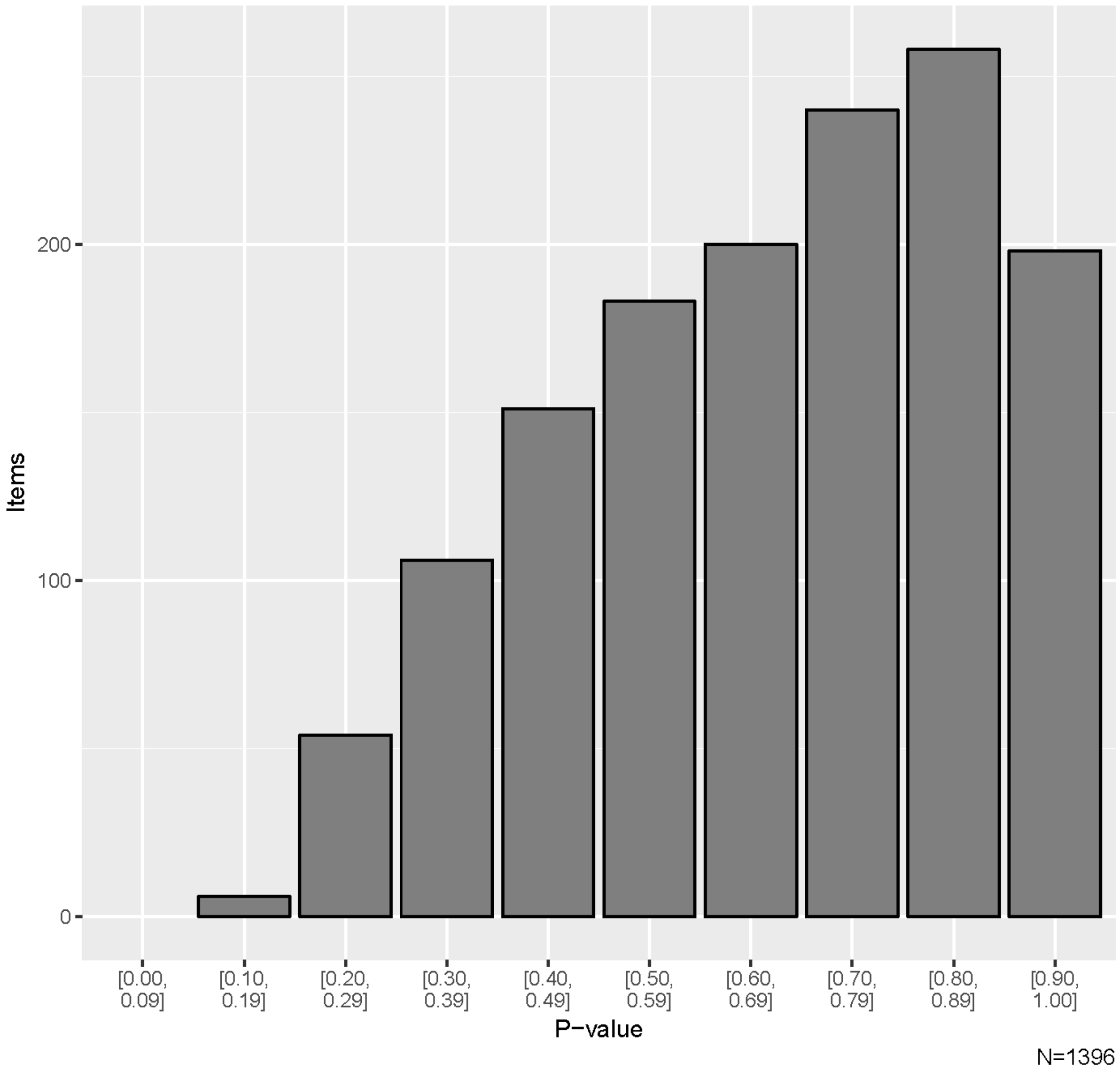


Figure 3. *p* values for English language arts 2016–2017 operational items.
Note. Writing items and items with a sample size of less than 20 were omitted.

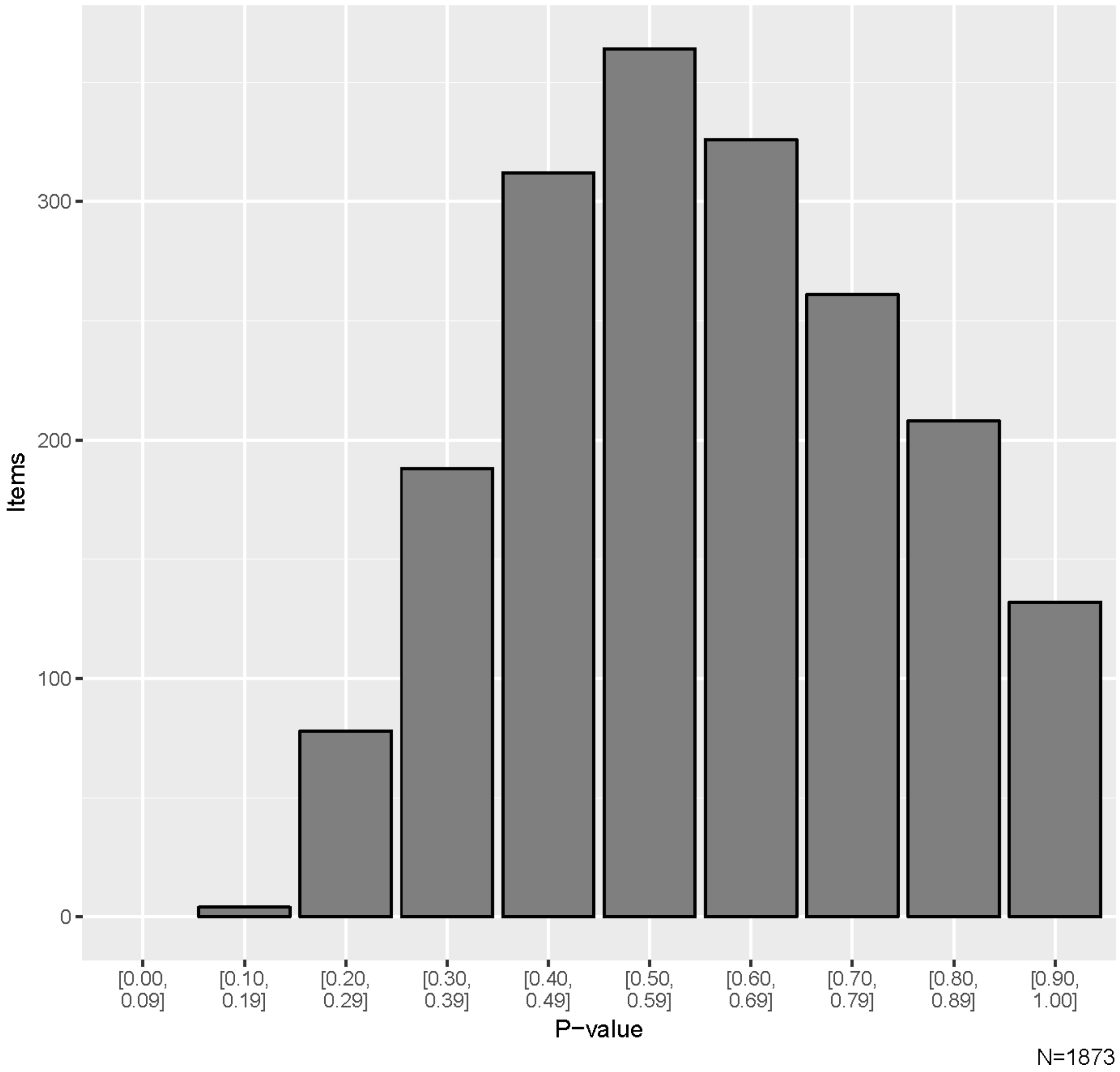


Figure 4. *p* values for mathematics 2016–2017 operational items.
Note. Items with a sample size of less than 20 were omitted.

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 combination. The standardized difference values provide one source of evidence of internal consistency. See Chapter IX in this manual for additional information.

Figure 5 and Figure 6 summarize the standardized difference values for operational items for ELA and mathematics, respectively. Most items fell within 2 standard deviations of the mean of all items measuring the EE and linkage level. As additional data are collected and decisions are made regarding item-pool replenishment, item standardized difference values will be considered along with item misfit analyses to determine which items and testlets are recommended for retirement.

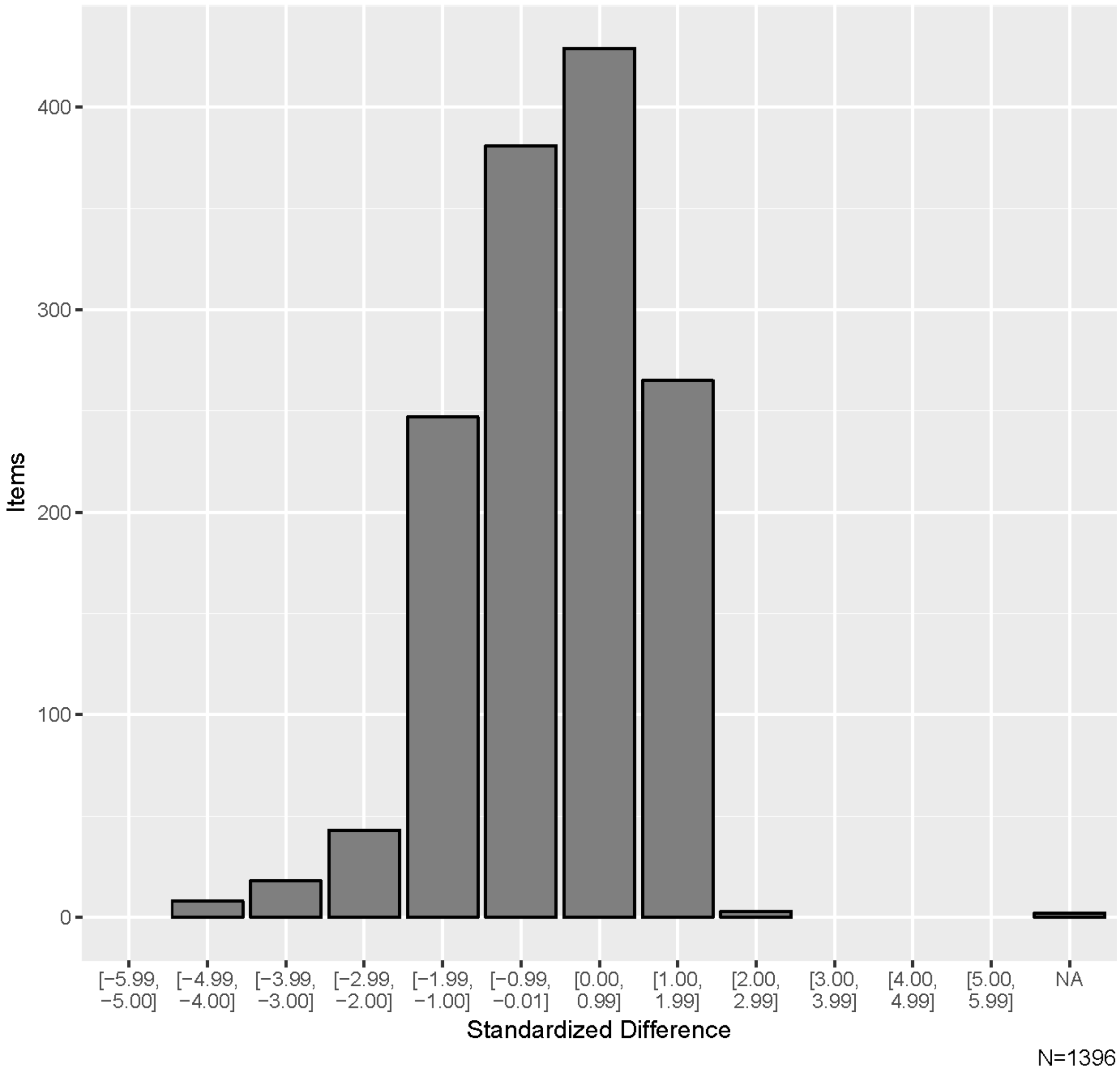


Figure 5. Standardized difference *z* scores for English language arts 2016–2017 operational items.
Note. Writing items and items with a sample size of less than 20 were omitted.

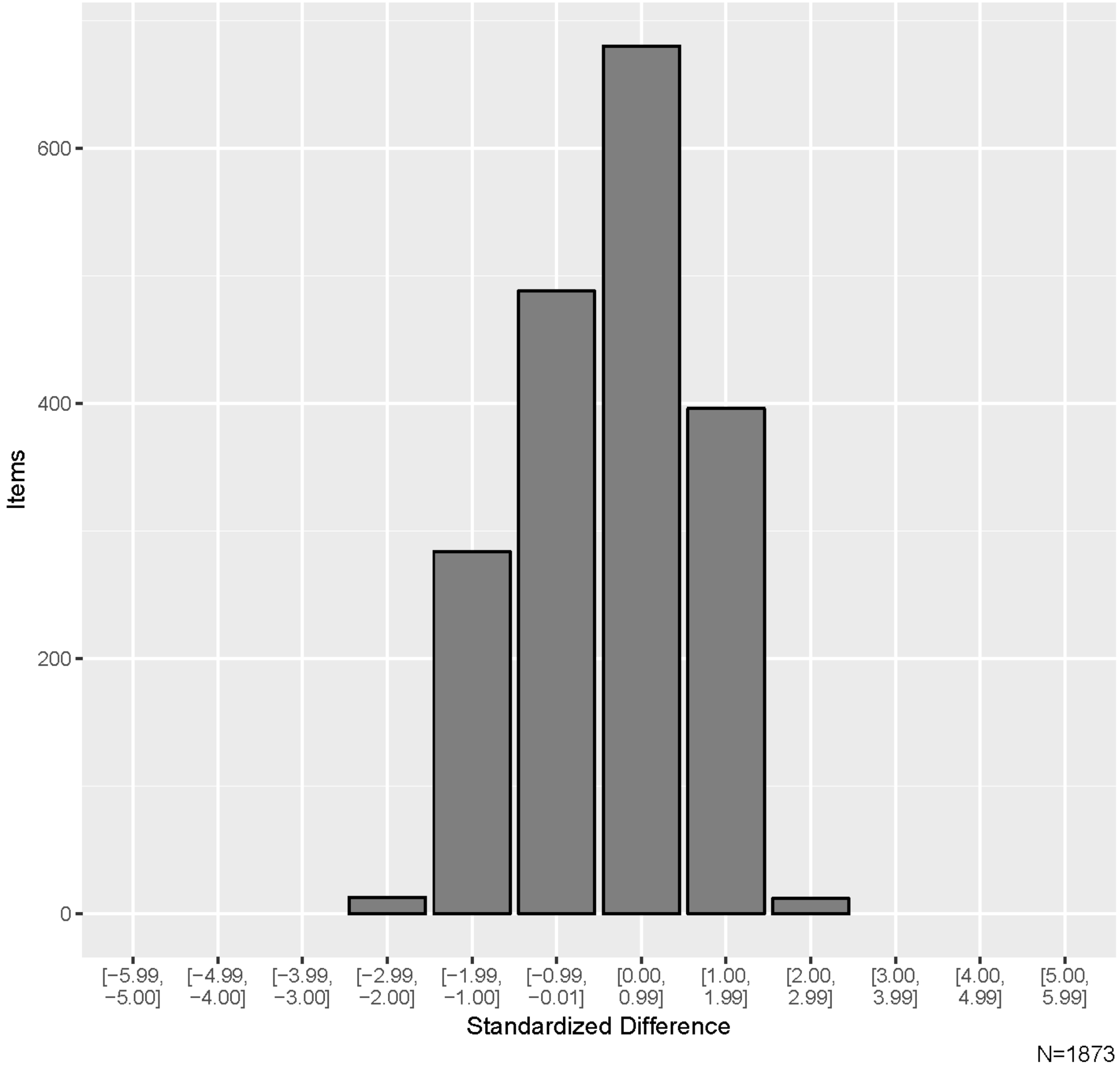


Figure 6. Standardized difference *z* scores for mathematics 2016–2017 operational items.
Note. Items with a sample size of less than 20 were omitted.

III.4. Field Testing

During the 2016–2017 academic year, DLM field tests were administered to evaluate item quality for EEs assessed at each grade level for ELA and mathematics. Field testing is conducted to deepen operational pools so that multiple testlets are available in spring window. By deepening the operational pools, testlets can also be evaluated for retirement in instances where other testlets perform better. Further, the additional data collected during field tests serve to inform modeling research, described in Chapter V of this manual.

A complete summary of prior field test events can be found in *Summary of Results from the 2014 and 2015 Field Test Administrations of the Dynamic Learning Maps® Alternate Assessment System* (Clark, Karvonen, & Wells-Moreaux, 2016), and in Chapter III of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b) and *2015—2016 Technical Manual Update – Year-End Model* (DLM Consortium, 2017c).

III.4.A. Description of Field Tests

During the spring window, all students received one field test testlet for each content area upon completion of all operational testlets. Similar to the prior year, the spring field test administration was designed to collect data for each participating student at more than one linkage level for an EE to support future modeling development (see Chapter V of this manual). As such, the field test testlet for each content area was assigned at one linkage level below the last linkage level at which the student was assessed. Because the process assigns the testlet one linkage level lower than the last testlet, no Successor-level testlets were field-tested during the spring 2017 window.

Testlets were made available for spring field testing in 2016–2017 based on the availability of field test content for each section of the assessment. Table 6 and Table 7 summarize the total number of field test testlets by content area and grade level for 2016–2017. A total of 195 field test testlets were available across grades and content areas. This number included one ELA test section for which both a general version and a version for students who are blind or visually impaired were available.

Table 6. 2016–2017 English Language Arts Field Test Testlets (*N* = 22)

Grade	*n*
3	2
4	4
5	3
6	2
7	3
8	2
9	2
10	2
11	2

Table 7. 2016–2017 Mathematics Field Test Testlets (*N* = 173)

Grade	*n*
3	17
4	20
5	23
6	16
7	15
8	17
9	22
10	22
11	21

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 2016–2017 are shown in Table 8. Because ELA had less content pending field test data collection, fewer students received an ELA field test assignment than mathematics. The large number of students responding to field test testlets allowed for more testlets to meet sample size requirements (responses from at least 20 students) and thus undergo statistical and content review prior to moving to the operational pool for the 2018 administration. Testlets that did not meet sample size requirements were scheduled for additional field testing.

Table 8. 2016–2017 Participation Rates in Spring Field Testing, by Content Area (*N* = 80,660)

Content area	*n*	%
English language arts	24,972	36.0
Mathematics	55,688	80.3

Of the 173 mathematics testlets available during spring testing, five (3%) did not meet the sample size threshold and required additional field testing prior to data analysis. Of the 22 ELA testlets available during spring testing, all met the sample size threshold and did not require additional field testing prior to data analysis. The testlets that did not meet the sample size threshold were scheduled for additional field testing in a subsequent administration.

III.4.B. Field Test Results

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

III.4.B.i. Item Flagging

Criteria used for item flagging during previous field test events were retained for 2016–2017. Items were flagged for review by test development teams if they met any of the following statistical criteria:

- The item was too challenging, as indicated by a proportion correct (*p* value) of less than .35. This value was selected as the threshold for flagging because most DLM items offer three response options, so a value of less than .35 may indicate chance selection of the option.
- The item was significantly easier or harder than other items assessing the same EE and linkage level, as indicated by a weighted standardized difference greater than 2 standard deviations from the mean *p* value for that EE and linkage level combination.

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

Figure 7 and Figure 8 summarize the *p* values for items field-tested during the spring 2017 window for ELA and mathematics, respectively. Most items fell above the .35 threshold for flagging. Test development teams for each content area reviewed items below the threshold.

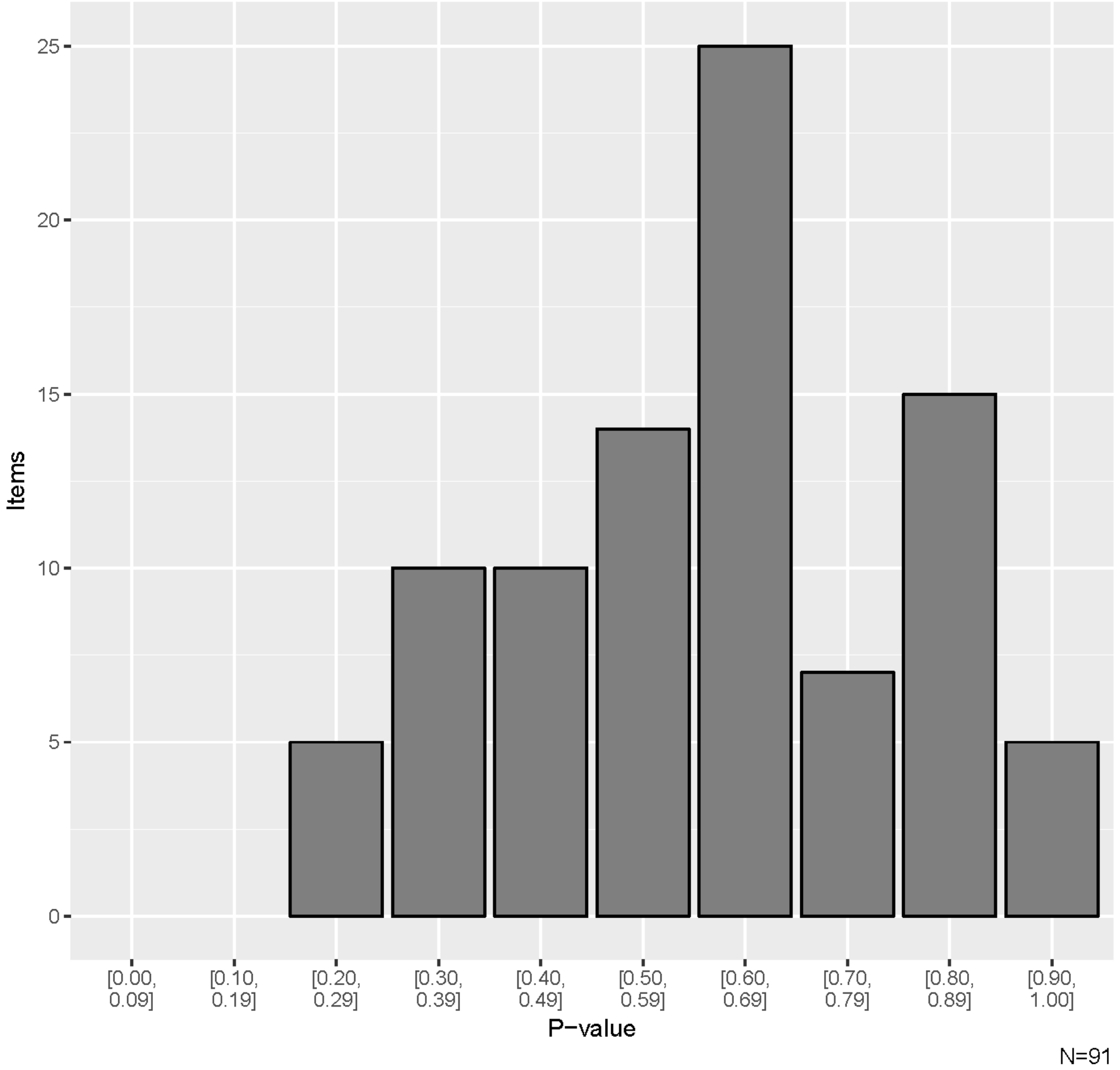


Figure 7. *p* values for English language arts items field-tested during the spring 2017 window. *Note*. Items with a sample size of less than 20 were omitted.

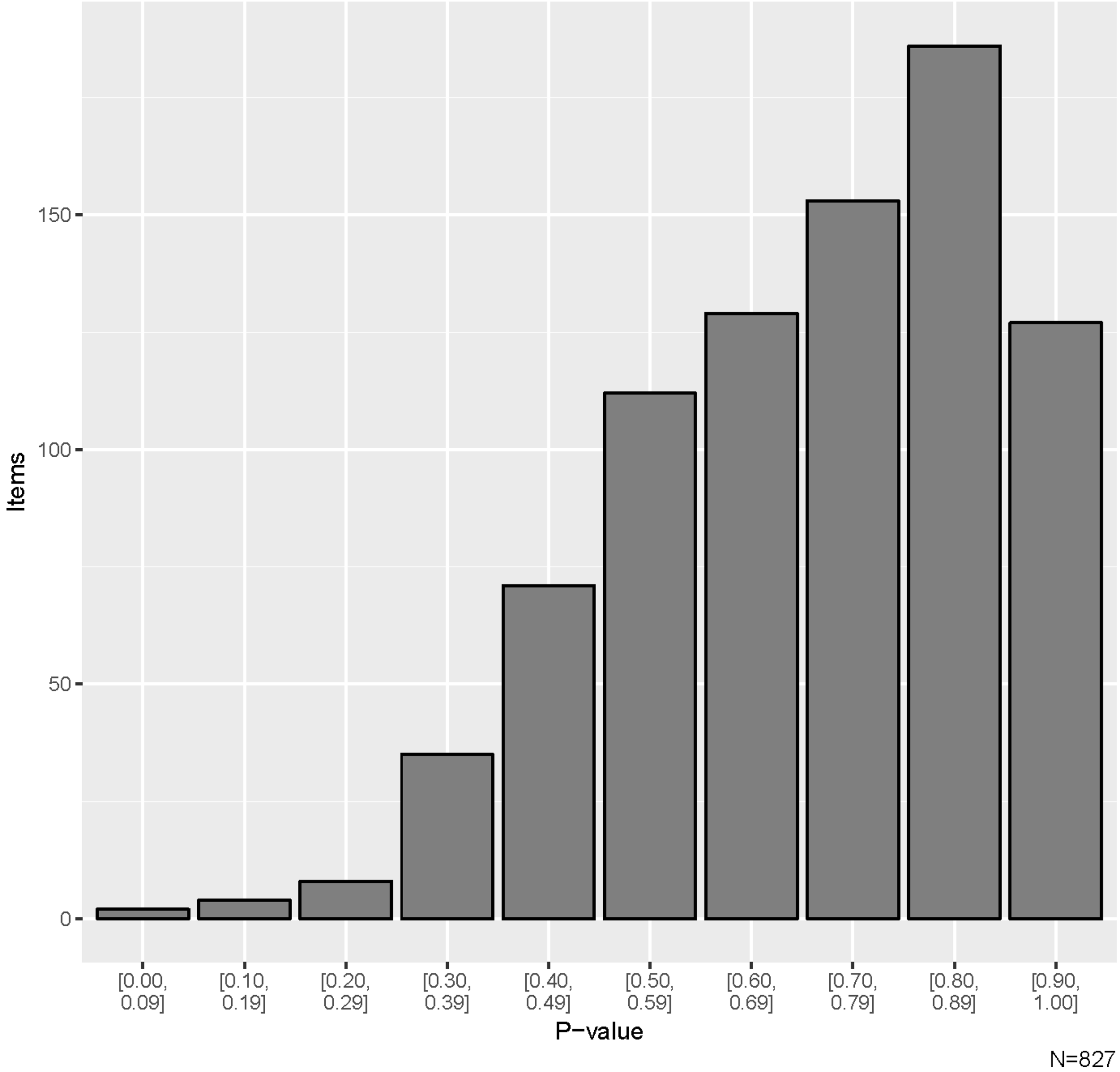


Figure 8. Shown are p values for mathematics items field-tested during the spring 2017 window. *Note*. Items with a sample size of less than 20 were omitted.

Figure 9 and Figure 10 summarize the standardized difference values for items field-tested during the spring 2017 window. Most items fell within 2 standard deviations of the mean for the EE and linkage level. Items beyond the threshold were reviewed by test development teams for each content area.

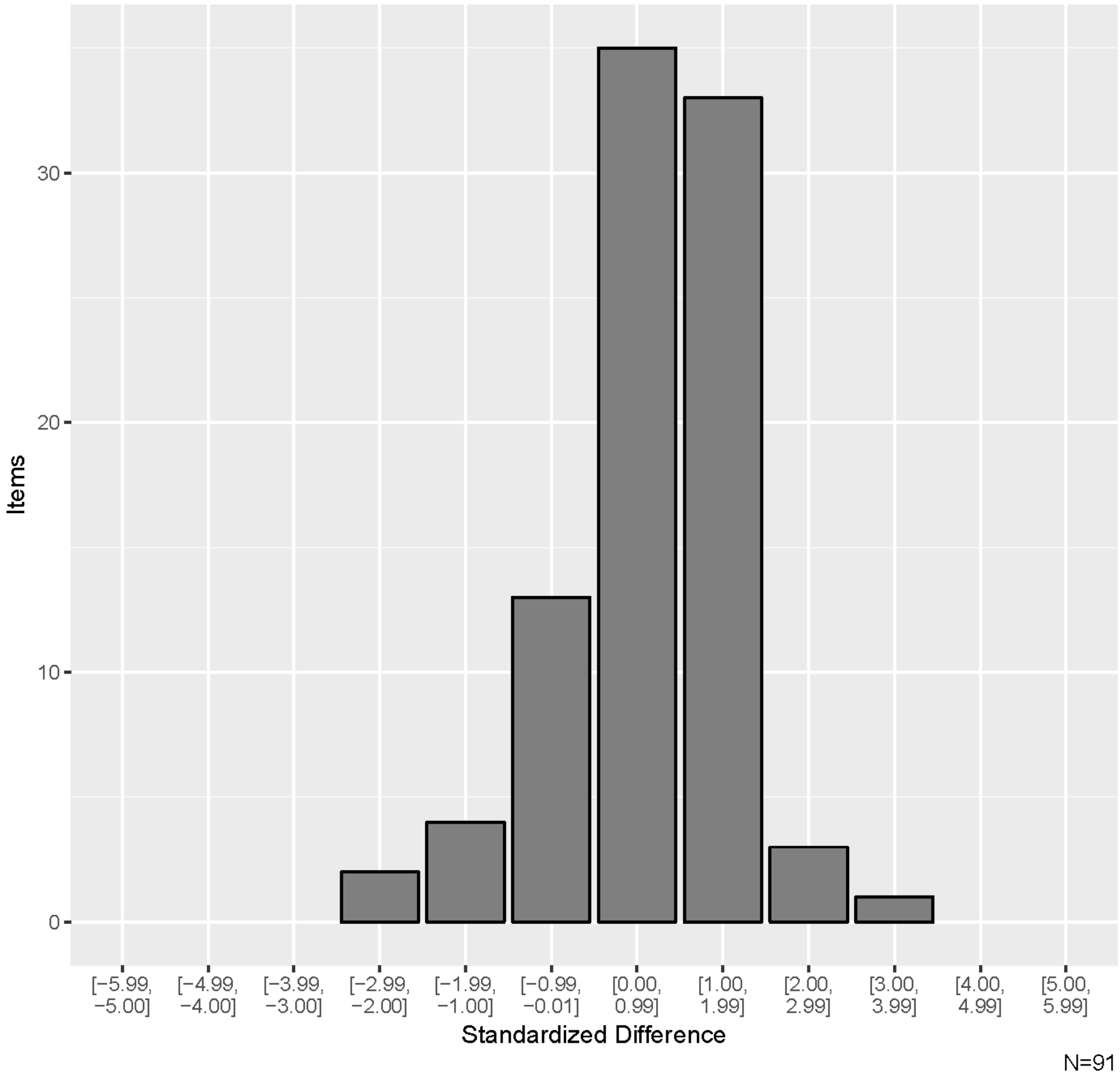


Figure 9. Standardized difference *z* scores for English language arts items field-tested during the spring 2017 window.
Note. Items with a sample size of less than 20 were omitted.

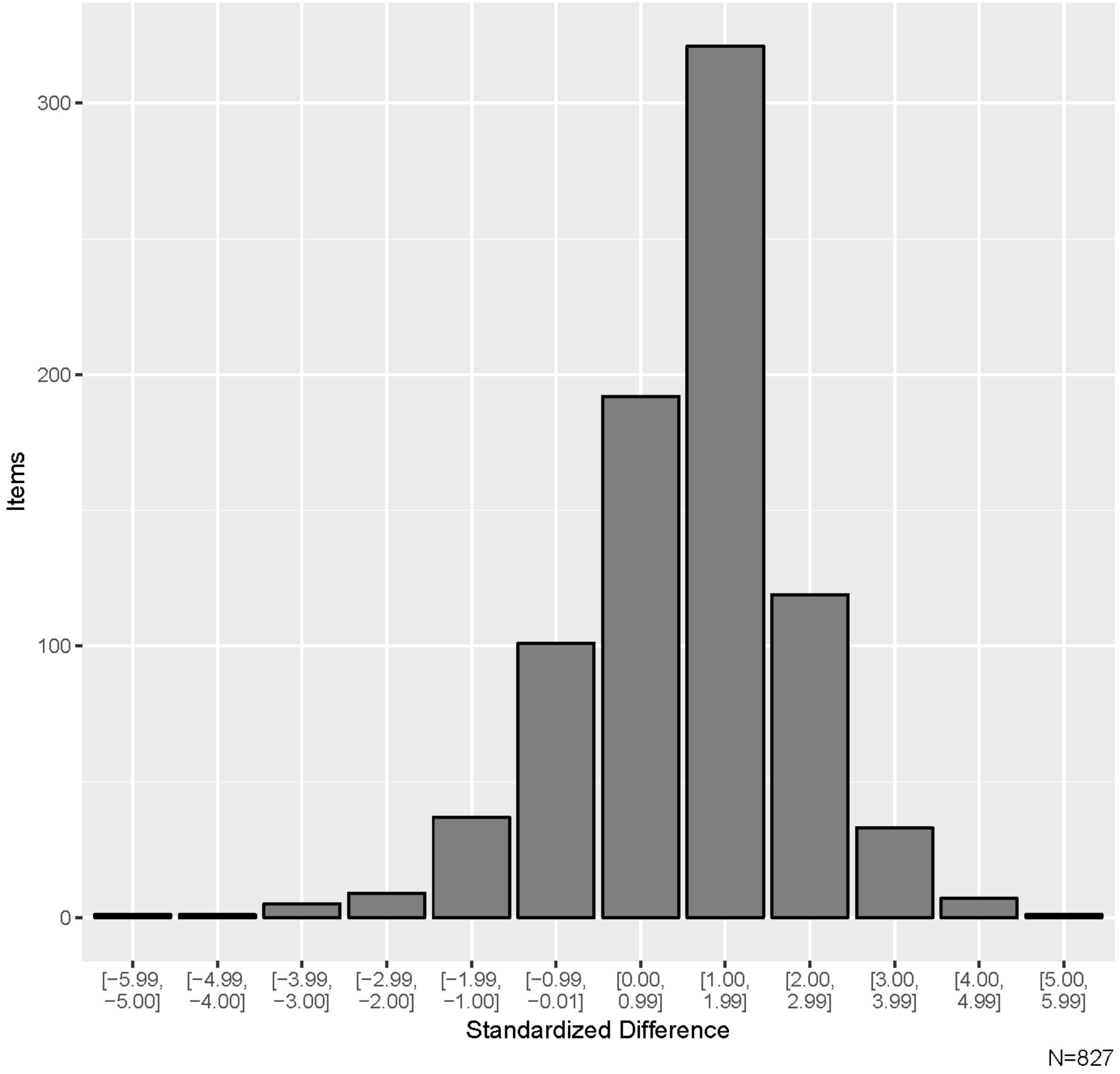


Figure 10. Standardized difference *z* scores for mathematics items field-tested during the spring 2017 window.
Note. Items with a sample size of less than 20 were omitted.

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

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

1. No changes made to item: Test development team decided item can go forward to operational assessment.

2. Test development team identified concerns that required modifications. Modifications were clearly identifiable and were likely to improve item performance.
3. Test development team identified concerns that required modifications: The content was worth preserving rather than rejecting. Item review may not have clearly pointed to specific edits that were likely to improve the item.
4. Reject item: Test development team determined the item was not worth revising.

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

Common reasons for flagging an item for modification included items that were misaligned to the node, distractors that could be argued as partially correct, or unnecessary complexity in the language of the stem.

After reviewing flagged items, the reviewers looked at all items rated as three or four within the testlet to help determine whether to retain or reject the testlet. Here, the test development team could elect to keep the testlet (with or without revision) or reject it. If a revision was needed, it was assumed the testlet needed retesting. The entire testlet was rejected if the test development team determined the flagged items could not be adequately revised.

III.4.B.iii. Results of Item Analysis and Test Development Team Review

A total of 15 ELA items and 190 mathematics items were flagged due to their *p* values and/or standardized difference values. Test development teams reviewed all flagged items and their context within the testlet to identify possible reasons for the flag and to determine whether an edit was likely to resolve the issue.

Table 9 and Table 10 provide the test development team accept, revise, and reject counts for all field test flagged items, for ELA and mathematics, respectively. In ELA, no items were rejected, whereas in mathematics, 28 items were rejected. Items were rejected in instances where test development team review indicated the item had more than one correct response option, no correct response option, or in cases where items were determined not to have met guidelines used for test development.

Table 9. English Language Arts Team Response to Item Flags, by Grade

Grade	Flagged items (*N*)	Accept		Revise		Reject	
		n	%	*n*	%	*n*	%
3	1	1	100.0	0	0.0	0	0.0
4	5	5	100.0	0	0.0	0	0.0
5	0	0	0.0	0	0.0	0	0.0
6	0	0	0.0	0	0.0	0	0.0
7	1	1	100.0	0	0.0	0	0.0
8	1	1	100.0	0	0.0	0	0.0
9	4	4	100.0	0	0.0	0	0.0
10	3	3	100.0	0	0.0	0	0.0
11	0	0	0.0	0	0.0	0	0.0

Table 10. Mathematics Team Response to Item Flags, by Grade

Grade	Flagged items (*N*)	Accept		Revise		Reject	
		n	%	*n*	%	*n*	%
3	22	21	95.5	0	0.0	1	4.5
4	20	20	100.0	0	0.0	0	0.0
5	23	19	82.6	0	0.0	4	17.4
6	14	13	92.9	0	0.0	1	7.1
7	19	17	89.5	0	0.0	2	10.5
8	22	17	77.3	0	0.0	5	22.7
9	17	15	88.2	0	0.0	2	11.8
10	34	27	79.4	0	0.0	7	20.6
11	19	13	68.4	0	0.0	6	31.6

Decisions to recommend testlets for retirement occur on an annual basis following the completion of the operational testing year. In instances where multiple testlets are available for an EE and linkage level combination, test development teams may recommend retiring testlets that perform poorly compared to others measuring the same EE and linkage level. The retirement process will begin following the 2016–2017 academic year and reported in the *2017–2018 Technical Manual Update – Year-End Model*.

IV. TEST ADMINISTRATION

Chapter IV of the *2014–2015 Technical Manual – Year-End Model* (Dynamic Learning Maps® [DLM®] Consortium, 2016b) describes general test administration and monitoring procedures. This chapter describes procedures and data collected in 2016–2017, including a summary of adaptive routing, administration incidents, 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 *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b).

IV.1. Overview of Key Administration Features

This section describes the testing windows for DLM test administration for 2016–2017. For a complete description of key administration features, including information on assessment delivery, the KITE® system, and linkage level selection, see Chapter IV of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b). Additional information about administration can be found in the *Test Administration Manual 2016–2017* (DLM Consortium, 2016a) and the *Educator Portal User Guide* (Dynamic Learning Maps Consortium, 2017a).

IV.1.A. Test Windows

Instructionally embedded assessments were available optionally for teachers to administer between September 21 and December 18, 2016, and between January 4 and February 28, 2017. During the consortium-wide spring testing window, which occurred between March 15 and June 9, 2017, students were assessed on each Essential Element (EE) on the blueprint. Each state set its own testing window within the larger consortium spring window.

IV.2. Implementation Evidence

This section describes evidence collected during the spring 2017 operational implementation of the DLM Alternate Assessment System. The categories of evidence include data relating to the adaptive delivery of testlets in the spring window, administration incidents, user experience, and accessibility.

IV.2.A. Adaptive Delivery

During the spring 2017 test administration, the ELA and mathematics assessments were adaptive between testlets, following the same routing rules applied in the previous two 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, skill, and ability 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 11.

Table 11. 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 IV of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b).

Following the spring 2017 administration, analyses were conducted to determine the mean percentage of testlets that adapted up a linkage level, stayed at the same linkage level, or adapted down a linkage level from the first to second testlet administered for students within a grade, content area, and complexity band. The aggregated results can be seen in Table 12 and Table 13 for ELA and mathematics, respectively.

Overall, results were similar to those found in the previous two 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 (range of 67.7% to 93.5% across both content areas). 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 content areas. Further investigation is needed to evaluate reasons for these different patterns.

The 2016–2017 results build on earlier findings from the pilot study and previous two years of operational assessment administration (see Chapter III and Chapter IV of the *2014–2015 Technical Manual – Year-End Model*, respectively, as well as Chapter III and Chapter IV of the *2015–2016 Technical Manual Update – Year-End Model*), and suggest that the First Contact survey complexity band assignment was 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 12. Adaptation of Linkage Levels Between First and Second English Language Arts Testlets (*N* = 69,408)

	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 (%)
3	19.7	80.3	32.2	38.8	29.0	72.1	13.2	14.8	93.4	3.3	3.3
4	32.0	68.0	20.2	42.7	37.0	34.3	42.0	23.6	54.6	43.9	1.5
5	21.4	78.6	26.3	30.8	42.9	61.0	26.4	12.6	64.3	29.1	6.6
6	17.5	82.5	23.3	9.3	67.3	41.5	21.8	36.7	37.6	20.2	42.1
7	18.7	81.3	19.0	32.1	48.9	30.8	35.1	34.1	41.3	29.2	29.5
8	32.3	67.7	29.4	41.9	28.7	50.9	38.7	10.5	84.8	12.2	3.0
9	17.4	82.6	19.2	9.4	71.5	33.1	12.9	54.0	44.6	9.7	45.7
10	15.5	84.5	20.2	39.3	40.5	27.9	45.8	26.3	50.8	41.4	7.7
11	14.9	85.1	3.2	26.5	70.2	23.3	41.6	35.1	39.3	44.5	16.2

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

Table 13. Adaptation of Linkage Levels Between First and Second Mathematics Testlets (*N* = 69,344)

	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 (%)**
3	6.5	93.5	7.4	30.7	61.9	15.7	27.1	57.2	7.5	53.1	39.3
4	16.5	83.5	50.3	12.9	36.8	62.0	18.4	19.6	53.5	23.7	22.7
5	25.5	74.5	11.1	17.2	71.7	16.2	8.7	75.1	57.3	5.4	37.3
6	13.6	86.4	13.9	25.1	61.0	16.1	34.4	49.4	43.2	28.2	28.5
7	11.5	88.5	8.7	18.4	73.0	32.3	33.9	33.8	38.7	9.2	52.0
8	16.2	83.8	13.4	6.2	80.4	3.2	10.3	86.6	12.2	17.1	70.7
9	17.2	82.8	8.8	30.2	61.0	7.1	42.0	51.0	14.5	42.3	43.2
10	16.2	83.8	4.0	22.4	73.5	3.3	28.2	68.6	24.5	52.7	22.8
11	11.7	88.3	2.1	25.1	72.8	1.7	23.6	74.7	8.5	55.8	35.7

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

IV.2.B. Administration Incidents

As in the previous two years, testlet assignment during the spring 2017 assessment window was monitored to ensure students were correctly assigned to testlets. Improving on the previous two years, only four incidents were observed that had the potential to impact scoring.

Table 14 provides a summary of the number of students potentially affected by each of the incidents, as delivered to states in the Incident File (see Chapter VII of this manual for more information). Following delivery of the Incident File on the predetermined delivery timeline, a script was created to identify students who were actually impacted by each incident, narrowing from the list of those potentially impacted. These values are also reported in Table 14. This script will be modified such that the 2018 Incident File and beyond reports students impacted by the incident rather than students potentially impacted.

The most frequent incident during the spring 2017 administration was potential misrouting caused by two items with multiple correct answer options. Overall, it was determined that each administration incident affected less than 0.1% of students.

Table 14. Number of Students Affected by Each 2017 Incident, Year-End Model

Incident code	Incident description	Potential impact (reported in Incident File)		Actual impact	
		n	%	*n*	%
2	Potential misrouting due to multiple correct answer options (two items).	162	0.23	70	0.10
3	Potential misrouting due to misspecified Essential Element (one item).	1,796	2.51	0	0.00
4	Potential incorrect scoring and misrouting due to KITE database load issue.	166	0.23	0*	0.00*
5	Misrouting due to testlet re-administration after student transfer.	2	<0.01	2	<0.01

Note: *Estimated actual impact. There is no evidence that the database did not record student responses.

Upon identification of the incidents coded 4 and 5 (i.e., “Potential incorrect scoring and misrouting due to the KITE database load issue” and “Misrouting due to testlet re-administration after student transfer”), states were provided with lists of students potentially impacted and given the option to revert each student’s assessment back to the end of the last correctly completed testlet (i.e., the point at which routing failed) and have the students complete the remaining testlets as intended. Additional details about the four incidents are described in Table 15, including a brief summary of the incident.

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

Table 15. Incident Summary for 2016–2017 Operational Testing

Incident no.	Issue	Type	Summary
2	Potential misrouting due to multiple correct answer options	Assessment: Content	Item had a correct distractor option, which was marked incorrect at the time of testing due to not being the system key, even though the response was also correct. Occurred for two items. Scoring was corrected prior to release of results, but the issue potentially impacted routing to subsequent testlets at the time of administration.
3	Potential misrouting due to misspecified Essential Element	Assessment: Content	One Initial Precursor item was not assigned to the correct Essential Element in the system. Scoring was corrected prior to the release of results, but the issue potentially impacted routing to subsequent testlets at the time of administration.
4	Potential incorrect scoring and misrouting due to KITE database load issue	Technology: Capacity	An integration server used by the test delivery application experienced server load issues April 4, 2017, from 8:50 a.m.–2:05 p.m. CST and April 5, 2017, from 8:45–10:55 a.m. CST. As a matter of regular practice, if the database times out before a student's response is submitted, the system starts the student over at the beginning of the testlet the next time the testlet is opened. There is no evidence that the database did not record all student responses during the two impacted periods. Because items may be intentionally skipped as a matter of practice or student choice, and because there is no evidence that the database did no record student responses, it is

Incident no.	Issue	Type	Summary
			assumed that all responses were recorded by the database as intended. However, out of an abundance of caution, testlets with one or more missing responses submitted during the two time periods were identified and provided to states for review. States were given the option to revert students to the end of the previously submitted testlet and resume testing, or to let students proceed forward as usual. Of the students reported in the Incident File as testing during the two impacted time periods, one student was reset to the end of the previously submitted testlet.
5	Misrouting due to testlet re-administration after student transfer	Technology: Administration	The student transferred to a different school, district, and/or teacher, and the system reassigned a previously taken testlet. During the second administration, the student provided different responses, resulting in a different percentage correct for routing purposes.

IV.2.C. User Experience With the Dynamic Learning Maps System

User experience with the system was evaluated through the spring 2017 survey disseminated to teachers who had administered a DLM assessment during the spring window. In 2017, the survey was distributed to teachers in KITE Client where students complete assessments. Each student was assigned a survey for his or her teacher to complete. The survey included three sections. The first and third sections were fixed, while the second section was spiraled, with teachers responding to blocks 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 7,875 teachers responded to the survey (response rate of 83.6%[2]) for 26,474 students. This reflects a substantial increase in the rate of responding teachers compared to those observed during previous delivery of surveys in Qualtrics (e.g., 2016 response rate was 11.5%). Because of the difference in response rates over years and changes to the structure and content of the survey, the spring 2017 administration was treated as a baseline data collection. Comparisons of data collected from 2016 are not included in this manual.

[2] Excluding two states, which opted out of the survey.

Participating teachers responded to surveys for between one and 25 students. Teachers most frequently reported having 0 to 5 years of experience in ELA, mathematics, and in teaching students with significant cognitive disabilities. The median years of experience in each of these areas ranged from 6 to 10. Approximately 56% of respondents indicated they had experience administering the DLM assessment in all three operational years.

The sections that follow summarize user experience with the system and accessibility. Additional survey results are summarized in Chapter IX of this manual. For responses to the 2014–2015 and 2015–2016 teacher surveys, see Chapter IV and Chapter IX in the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b) and the *2015–2016 Technical Manual Update – Year-End Model* (DLM Consortium, 2017c), respectively.

IV.2.C.i. Educator Experience

Respondents were asked to reflect on their own experience with the assessments and their comfort level and knowledge with regard to administering them. Most of the questions required respondents to rate results on a 4-point scale: *strongly disagree*, *disagree*, *agree*, or *strongly agree*. Responses are summarized in Table 16.

Most teachers (97.0%) agreed or strongly agreed that they were confident in administering DLM testlets. Most respondents (89.5%) 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 (89.3%), they knew how to use accessibility supports, allowable supports, and options for flexibility (93.7%), and that the Testlet Information Pages helped them deliver the testlets (86.5%).

Table 16. Teacher Responses Regarding Test Administration

	SD		D		A		SA		A+SA	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Confidence in ability to deliver DLM testlets	32	0.8	94	2.2	1,782	42.4	2,294	54.6	4,076	97.0
Test administrator training prepared respondent for responsibilities of test administrator	101	2.4	336	8.1	2,211	53.1	1,518	36.4	3,729	89.5
Manuals and DLM Educator Resources Page materials helped respondent understand how to use assessment system	86	2.1	359	8.6	2,348	56.3	1,380	33.1	3,728	89.3
Respondent knew how to use accessibility features, allowable supports, and options for flexibility	45	1.1	216	5.2	2,354	56.4	1,559	37.4	3,913	93.7
Testlet Information Pages helped respondent deliver the testlets	116	2.8	448	10.7	2,309	55.3	1,303	31.2	3,612	86.5

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

IV.2.C.i.a 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 5-point scale: *very hard*, *somewhat hard*, *neither hard nor easy*, *somewhat easy*, or *very easy*. Table 17 summarizes teacher responses to these questions.

Respondents found it to be either Somewhat Easy or Very Easy to enter the site (79.5%), to navigate within a testlet (82.2%), to record a response (85.6%), to submit a completed testlet (85.9%), and to administer testlets on various devices (73.4%). Open-ended survey response feedback indicated testlets were easy to administer and that technology had improved compared to previous years.

Table 17. Ease of Using KITE Client

	VH		SH		N		SE		VE		SE+VE	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Enter the site	58	1.4	245	5.7	577	13.4	1,294	30.1	2,122	49.4	3,416	79.5
Navigate within a testlet	49	1.1	194	4.5	519	12.1	1,283	29.9	2,246	52.3	3,529	82.2
Record a response	32	0.7	112	2.6	470	11.0	1,183	27.7	2,481	58.0	3,664	85.6
Submit a completed testlet	39	0.9	115	2.7	446	10.5	1,112	26.1	2,541	59.7	3,653	85.9
Administer testlets on various devices	78	1.8	244	5.7	814	19.1	1,258	29.5	1,870	43.9	3,128	73.4

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 the software used to store and manage student data and to enter PNP and First Contact information. Teachers were asked to assess the ease of navigating and using Educator Portal for its intended purposes. The data are summarized in Table 18 using the same scale used to rate experiences with KITE Client. Overall, respondents' feedback was generally positive: most teachers found it to be either Somewhat Easy or Very Easy to navigate the site (62.3%), to enter PNP and First Contact information (70.2%), to manage student data (61.0%), to manage their accounts (64.4%), and to manage tests (62.2%).

Table 18. Ease of Using Educator Portal

	VH		SH		N		SE		VE		SE+VE	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Navigate the site	153	3.6	690	16.1	771	18.0	1,455	34.0	1,214	28.3	2,669	62.3
Enter Access Profile and First Contact information	77	1.8	427	10.0	770	18.0	1,627	38.1	1,371	32.1	2,998	70.2
Manage student data	135	3.2	606	14.2	926	21.7	1,536	36.0	1,068	25.0	2,604	61.0
Manage my account	105	2.5	475	11.1	945	22.1	1,599	37.3	1,158	27.0	2,757	64.4
Manage tests	143	3.3	627	14.7	842	19.7	1,484	34.8	1,174	27.5	2,658	62.2

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.

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

Finally, respondents were asked to rate their overall experience with KITE Client and Educator Portal on a 4-point scale: *poor*, *fair*, *good*, and *excellent*. Results are summarized in Table 19. The majority of respondents reported a positive experience with KITE Client. A total of 79.6% of respondents rated their experience as good or excellent, while 70.2% rated their overall experience with Educator Portal as good or excellent.

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 that will be considered for technology development for 2017–2018 and beyond.

Table 19. Overall Experience With KITE Client and Educator Portal

	Poor		**Fair**		**Good**		**Excellent**	
Interface	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
KITE Client	161	3.7	718	16.7	2,138	49.7	1,289	29.9
Educator Portal	274	6.4	1,007	23.4	2,198	51.0	828	19.2

IV.2.C.ii. Accessibility

Accessibility supports provided in 2016–2017 were the same as those available in the previous two years. DLM accessibility guidance (Wells-Moreaux, Bechard, & Karvonen, 2016) distinguishes between accessibility supports that: (a) are provided in KITE Client via the Access Profile, (b) require additional tools or materials, and (c) are provided by the test administrator outside the system.

Table 20 shows selection rates for three categories of accessibility supports, sorted by rate of use within each category. 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 IV in *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b).

Table 20. Accessibility Supports Selected for Students (*N* = 71,731)

Support	***n***	**%**
Supports provided in KITE Client via Access Profile		
Spoken audio	21,412	29.9
Magnification	15,525	21.6
Color contrast	13,698	19.1
Overlay color	12,574	17.5
Invert color choice	11,452	16.0
Supports requiring additional tools/materials		
Individualized manipulatives	36,516	50.9
Calculator	26,356	36.7
Single-switch system	12,233	17.1
Alternate form – visual impairment	10,277	14.3
Two-switch system	9,714	13.5
Uncontracted braille	8,915	12.4
Supports provided outside the system		
Human read aloud	63,983	89.2
Test administrator enters responses for student	40,613	56.6
Partner-assisted scanning	14,338	20.0
Sign interpretation of text	9,887	13.8
Language translation of text	10,042	14.0

Table 21 describes teacher responses to survey items that asked about the accessibility supports used during administration. Teachers were asked to respond to two items using a 4-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 the student was able to effectively use accessibility supports (81.8%), and that accessibility supports were similar to ones the student used for instruction (82.7%). 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 2018 survey to determine whether results improve over time.

Table 21. 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.	118	2.6	158	3.4	1,925	41.8	1,839	40.0	3,764	81.8	561	12.2
Accessibility features were similar to ones student uses for instruction.	101	2.2	177	3.9	1,872	40.9	1,911	41.8	3,783	82.7	516	11.3

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

IV.3. Conclusion

During the 2016–2017 academic year, the DLM system was available during two testing windows: instructionally embedded (optional) and spring. Implementation evidence was collected in the form of testlet adaptation analyses, a summary of students affected by incidents during operational testing, and teacher survey responses regarding user experience, and accessibility. Results indicated that teachers felt confident administering testlets in the system and that KITE Client was easy to use but that Educator Portal posed some challenges, but had improved since the prior year. Further, teacher surveys will continue to be distributed in KITE Client based on the substantial improvement in teacher response rates as compared to using Qualtrics.

V. MODELING

Chapter V of the *2014–2015 Technical Manual – Year-End Model* (Dynamic Learning Maps® [DLM®] Consortium, 2016b) describes the basic psychometric model that underlies the DLM assessment system, while the *2015–2016 Technical Manual Update – Year-End Model* (DLM Consortium, 2017c) 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 2016–2017 administration year. Additional evidence provided includes a description of model-fit analyses and results.

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, 2017c).

V.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 ELA and mathematics. Because of the underlying map structure and the goal to provide more fine-grained information beyond a single raw or scale score value when reporting student results, the assessment system provides a profile of skill mastery to summarize student performance. This profile is created using a form of diagnostic classification modeling, called latent class analysis, to provide information about student mastery on multiple skills measured by the assessment. Results are reported for each alternate content standard, called Essential Elements (EEs), 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, where overlapping data from students taking testlets at multiple levels within an EE are uncommon. Instead, each linkage level was calibrated separately for each EE using separate latent class analyses. Additionally, 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.

The DLM scoring model for the 2016–2017 administration was as follows. Using latent class analysis, a probability of mastery was calculated on a scale of 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: either master or non-master. As described in Chapter VI of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b), a posterior probability of at least .8 was required for mastery classification. As per the assumption of item fungibility, a single set of probabilities of providing a correct response for masters and non-masters was estimated for all items within a linkage level. Finally, a structural parameter was also estimated, which is the proportion of masters for the linkage level (i.e., the analogous map parameter). 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, results for each linkage level were combined to determine the highest linkage level mastered for each EE. Although the connections between linkage levels were not modeled empirically, they were used in the scoring procedures. In particular, if the latent class analysis determined a student was judged to have mastered a given linkage level within an EE, then the student was assumed to have mastered all lower levels within that EE.

In addition to the calculated posterior probability of mastery, students were able to demonstrate mastery of each EE in two additional ways: (a) having correctly answered 80% of all items administered at the linkage level, or (b) 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.

V.2. Calibrated Parameters

As mentioned in the previous section, for diagnostic assessments, the comparable *item parameters* are conditional probabilities of 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. Parameters include a conditional probability of providing a correct response for non-masters and a conditional probability of providing a correct response for masters. Across all linkage levels, it is generally expected that the conditional probability of providing a correct response will be high for masters and low for non-masters. A summary of the operational parameters used to score the 2016–2017 assessment is provided in the following sections.

V.2.A. Probability of Master 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. Figure 11 depicts the conditional probability of masters providing a correct response to items measuring each of the 1,210 linkage levels based on the spring 2017 calibration. Because the point of maximum uncertainty is 0.5, masters should have a greater than 0.5 chance of providing a correct response. The results in Figure 11 demonstrate that most linkage levels performed as expected.

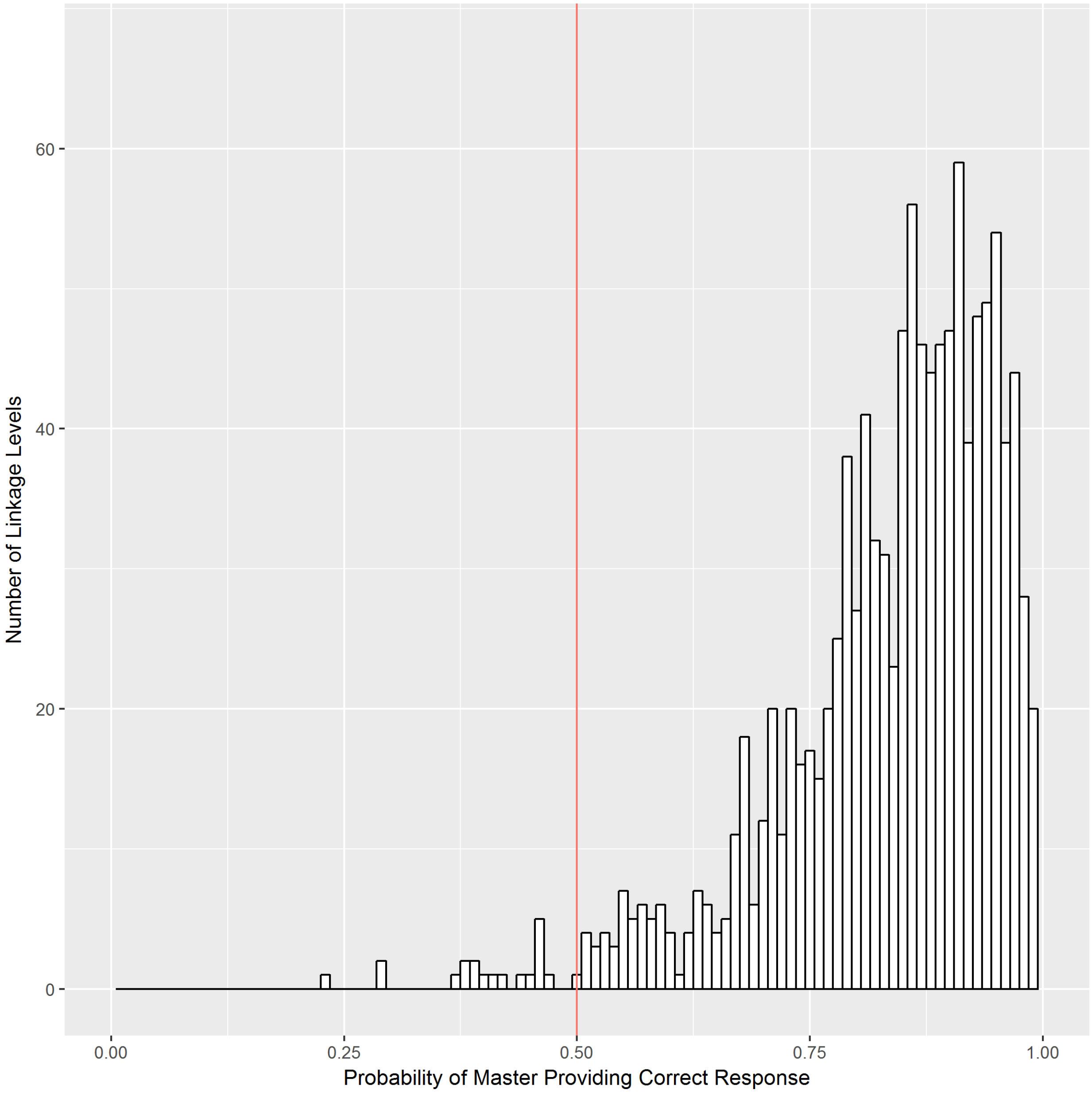


Figure 11. Probability of masters providing a correct response to items measuring each linkage level.

Note. Histogram bins are in increments of 0.01. Reference line indicates 0.5.

V.2.B. Probability of Non-Master 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 the linkage level does not measure what it intends to measure, or the correct answers to items measuring the level are easily guessable. This may result in students who have not mastered the content providing correct responses and potentially being classified as masters incorrectly, which has implications for the validity of inferences that can be made from results and the utility of teachers using results to inform instructional planning.

Figure 12 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, with the histogram in Figure 12 indicating that non-masters sometimes have a greater than chance (>0.5) 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) are easily guessable or do not discriminate well between the two groups of students.

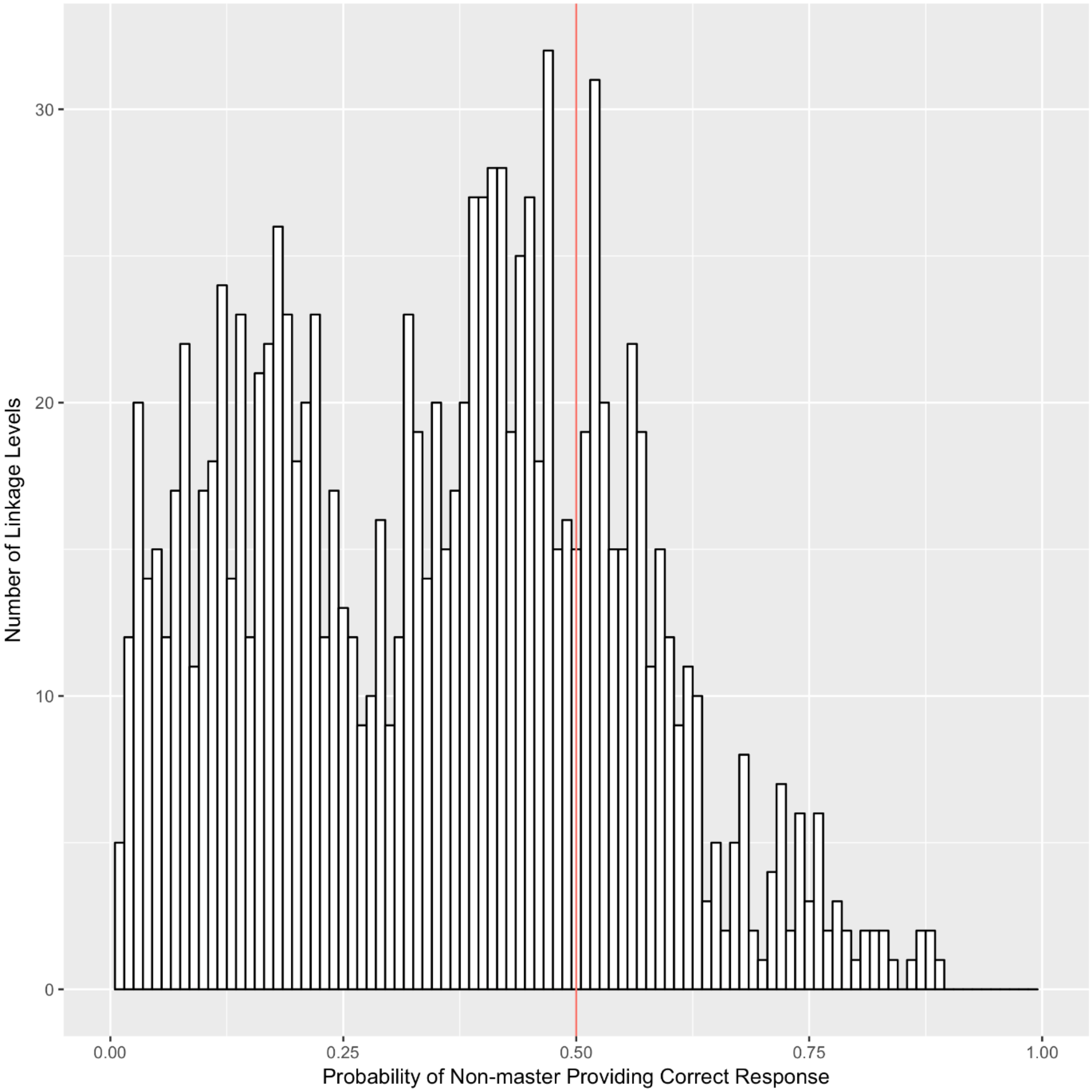


Figure 12. Probability of non-masters providing a correct response to items measuring each linkage level.
Note. Histogram bin size is in increments of 0.01. Reference line indicates 0.5.

V.3. Mastery Assignment

As mentioned, in addition to the calculated posterior probability of mastery, students were able to demonstrate mastery of each EE in two additional ways: (a) having answered 80% of all items administered at the linkage level correctly, or (b) the two-down scoring rule. To evaluate the degree to which each mastery assignment rule contributed to students' linkage level mastery status during the 2016–2017 administration of DLM assessments, the percentage of mastery statuses obtained by each scoring rule was calculated, as shown in Figure 13. Posterior probability was given first priority. If mastery was not demonstrated by the posterior probability threshold being met, the next two scoring rules were imposed. Approximately 60% to 75% of mastered linkage levels were derived from the posterior probability obtained from the modeling procedure. The other approximately 25% to 40% of linkage levels were assigned mastery status by the minimum mastery, or two-down rule, and the remaining percentages at each grade were determined by the percentage-correct rule. These results indicate that the percentage-correct rule likely had strong overlap (but was ordered second in priority) with the posterior probabilities, in that correct responses to all items measuring the linkage level were likely necessary to achieve a posterior probability above the 0.8 threshold. The percentage-correct rule does, however, provide mastery status in those instances, where entering correct responses to all or most items still resulted in a posterior probability below the mastery threshold.

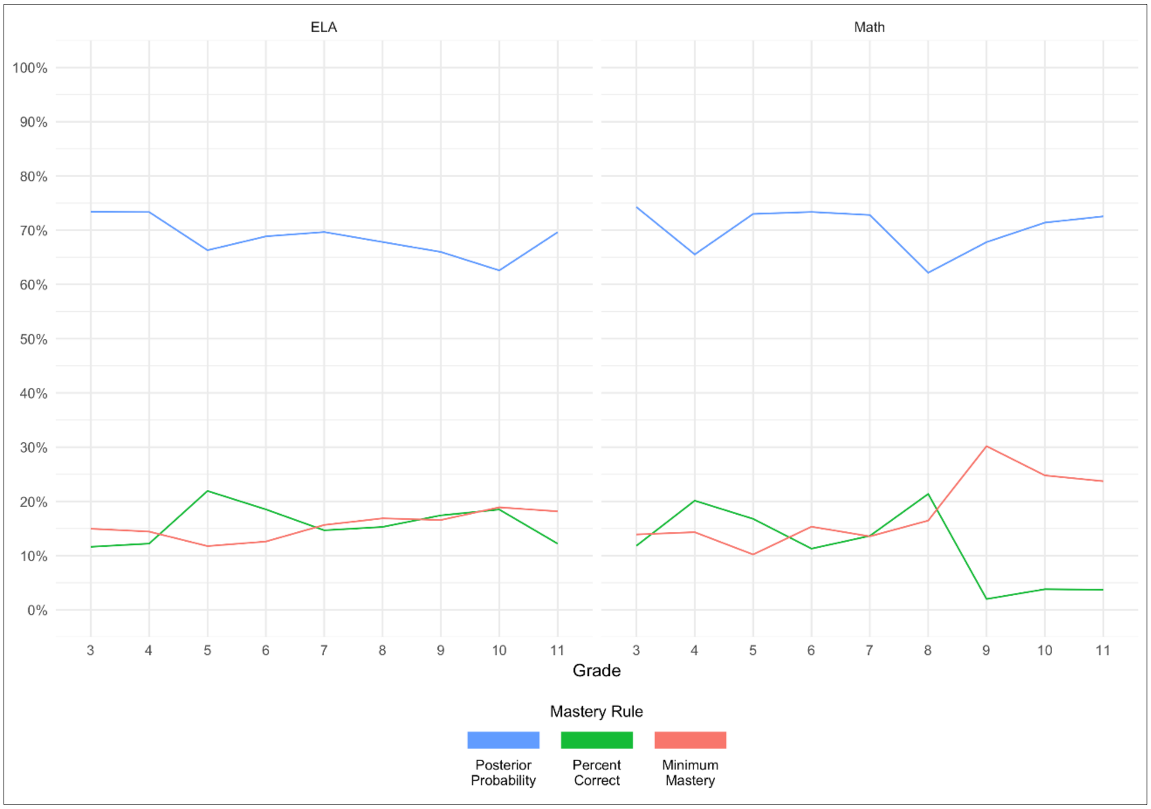


Figure 13. Linkage level mastery assignment by mastery rule for each content area and grade.

V.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. Because one of the assumptions of the DLM assessment system is that items measuring the same linkage level are fungible, or exchangeable, evidence of the degree to which a fungible model fits the data must be evaluated. Also, the fit of the fungible model should be compared to a nonfungible model to evaluate their relative fit.

The following sections provide a detailed description of the methodology used to evaluate model fit using both relative and absolute indices. Results are summarized for the 1,275[3] linkage levels measured by the assessment, which includes 740 linkage levels for ELA based on 148 EEs and 535 linkage levels for mathematics based on 107 EEs.

V.4.A. Description of Methods

To evaluate model fit for DLM assessments, two models were fit to each linkage level: a fungible and a nonfungible model. Definitions of each model follow, where π_{ij} is the probability of a respondent in class *j* providing a correct response to item *i*, η_j is the base rate probability of class *j*, and respondents are subscripted as $h = \{1,2,3,...N\}$, items as $i = \{1,2,3,...I\}$, and classes as $j = \{1,2,...J\}$.

- ***Fungible Model***. In the fungible model, the conditional probabilities for non-masters and masters were held constant for all items measuring the same linkage level.

$$f(\mathbf{x}_h) = \sum_{j=0}^{J} \eta_j \prod_{i=1}^{I} \pi_j^{x_{ih}} (1 - \pi_j)^{1-x_{ih}}$$

  In the previous equation, the probability of a correct response for a respondent in class *j* is denoted as π_j rather than π_{ij}, indicating that π is constant across items for all members of class j.

- ***Nonfungible Model***. In the nonfungible model, the conditional probabilities for non-masters and masters were allowed to vary across all items and linkage levels.

$$f(\mathbf{x}_h) = \sum_{j=0}^{J} \eta_j \prod_{i=1}^{I} \pi_{ij}^{x_{ih}} (1 - \pi_{ij})^{1-x_{ih}}$$

  In the previous equation, the probability of a correct response for a respondent in class *j* is denoted as π_{ij}, indicating that π is specific to each item within class j.

[3] The total of 1,275 includes all EEs and linkage levels measured by the integrated and year-end models. Because calibration is conducted together, model-fit evidence is also provided together to prevent results from appearing to be different by model, with the exception of operational fit results.

Because of the conceptual basis for linkage levels measuring a single skill (see Chapter II of DLM Consortium, 2016b), the fungible model has been used to calibrate and score DLM assessments to date. However, given that the item parameters are allowed to vary in the nonfungible model, it is expected that this model would demonstrate superior model fit. As is the case for the vast majority of statistical models, additional parameters will increase the fit to the data. However, the trade-off is that increasing the number of parameters also increases the risk of overfitting the model to the data. When this happens, the model is not generalizable to data outside of the sample used to estimate the model. Thus, if a more parsimonious model can provide adequate model fit, that simpler model would be preferred. Moreover, because there are fewer parameters to estimate, the fungible model allows for a faster calibration. That all items in the fungible model share the same parameter also means extreme parameter values are less likely. Parameters for items with low sample sizes are not allowed to vary freely but are instead pulled into the fungible parameter, which is calculated on the full sample of available data. This has important implications for scoring in operational assessment systems.

V.4.B. Background on Model Fit Calculation

To provide evidence of model fit for two competing models (e.g., fungible and nonfungible), model fit evidence can be provided in the form of both relative and absolute fit indices. Relative fit compares the fit of two competing models to determine which better fits the data. However, to determine how well each individual model fits the data, absolute fit indices are necessary. The sections that follow describe considerations when calculating both relative and absolute model fit.

V.4.B.i. Relative Fit

The relative fit of two competing models can be evaluated by comparing two nested models in a likelihood ratio test (Neyman & Pearson, 1933). This test provides information about which of two competing models provides better fit to the data when summarizing results across all linkage levels. Relative fit is calculated based on the final log likelihoods from the nested models and the number of parameters in each. Take, for example, a latent class analysis with five items. In the nonfungible model, there are 11 parameters estimated: a conditional probability of a correct response for masters and non-masters (one each for all five items = 10) and one structural parameter that is the base rate probability of mastery. The fungible model has three parameters: one conditional probability of masters providing a correct response shared by all items, one conditional probability of non-masters providing a correct response shared by all items, and one structural parameter. Because the nonfungible model has more parameters, it is expected to always have a larger log likelihood (i.e., better fit). However, the likelihood ratio test tests whether this increase is large enough to justify the additional parameters. The likelihood ratio test is a χ^2 test defined as follows:

$$\chi^2 = 2\ln\left(\frac{\text{likelihood for alternative model}}{\text{likelihood for null model}}\right)$$

$$df = df_{\text{alt}} - df_{\text{null}}$$

In this notation, the null model is the more simplified, or nested, model (the fungible model for DLM scoring). If this test is significant, then the null model is rejected, and it is determined that the additional parameters in the alternative model provide a statistically significant increase in the likelihood.

V.4.B.ii. Absolute Fit

In item response theory, model goodness-of-fit is commonly assessed using residual analysis (see Hambleton, Swaminathan, & Rogers, 1991). At the item level, the continuous theta is split into quadrature nodes, and the fitted item-characteristic curve is used to determine the expected proportion of correct responses at each quadrature point. The observed data are then used to calculate the observed proportion correct for each quadrature node. The difference between these proportions (the residual) is then standardized by dividing by the standard error of the residual. Thus, the prediction errors are essentially turned into *z* scores, which can be summed across all quadrature points for an item. Summed *z* scores follow a χ^2 distribution, with degrees of freedom equal to the number of quadrature points. Thus, for each item, a χ^2 test can be conducted to determine item-level misfit. At the test level, item-characteristic curves can be aggregated into a test characteristic curve, and a similar test can be done across quadrature points to assess test-level model fit.

Because DLM assessments use diagnostic models, where the latent trait is categorical rather than continuous, it is not possible to create item- or test-characteristic curves. Nevertheless, a similar approach can be taken in that a χ^2 can be calculated for each item based on the residuals. However, because the latent trait is categorical, the expected proportion of respondents in each score category can be calculated directly from model parameters instead of breaking the trait into quadrature points.

As an example, consider a dichotomous attribute, where the base rate probability of mastery is .6, and an item that measures this attribute, where masters have a .8 probability and non-masters a .15 probability of providing a correct response. Given these parameters, the proportion of respondents expected to provide a correct response can be calculated as follows:

$$\begin{aligned} P(X_i = 1) &= \eta_1\pi_{i1} + \eta_2\pi_{i2} \\ &= (0.6)(0.8) + (0.4)(0.15) \\ &= 0.54 \end{aligned}$$

Similarly, the proportion of respondents expected to provide an incorrect response can be calculated as follows:

$$\begin{aligned} P(X_i = 0) &= \eta_1(1 - \pi_{i1}) + \eta_2(1 - \pi_{i2}) \\ &= (0.6)(0.2) + (0.4)(0.85) \\ &= 0.46 \end{aligned}$$

These proportions can be converted to frequencies by multiplying the expected proportions by the total number of respondents who took the item. For example, if 100 respondents had taken this item, a contingency table could be constructed showing the number of expected and observed respondents at each score point (Table 22).

Table 22. Univariate Contingency Table

Item 1 score	Expected N	Observed N
0	46	48
1	54	52

Using the data in Table 22, a χ^2 goodness-of-fit test can be calculated ($\chi^2_{(1)}$ = 0.16, p = 0.68). Because the p value is nonsignificant, this test would not indicate item-level misfit.

In addition to looking at a single item, it is also possible to look at the fit of multiple items simultaneously. For example, when using two items, a 2x2 contingency table can be constructed to show the observed and expected frequencies of each response pattern. Table 23 presents these contingency tables (one for expected frequencies and one for observed frequencies) together in one long format table for readability.

Table 23. Bivariate Contingency Table

Item 1 score	Item 2 score	Expected *N*	Observed *N*
0	0	26	30
0	1	20	18
1	0	10	7
1	1	44	45

As with the univariate example in Table 22, a χ^2 goodness-of-fit test can also be conducted on these expected and observed frequencies ($\chi^2_{(3)}$ = 1.74, p = 0.63). This family of tests is known as *limited information goodness-of-fit tests* (see Maydeu-Olivares & Joe, 2006), as they use only subsets of items. This approach can continue to add dimensions (e.g., trivariate tables); however, as more dimensions are added, the number of possible responses increases exponentially (number of response patterns = 2^{items}). Thus, the expected and observed counts at each possible response pattern begin to get too small for there to be a stable χ^2 test.

Because these tests can only use a subset of items, they are unable to give an evaluation of fit for the entire model. Unlike in item response theory, there is no test characteristic curve that can be used to aggregate across items. Theoretically, this could be achieved by using the model parameters to calculate the expected sum score for each latent class (similar to the test characteristic curve indicating the expected sum score for each theta value). However, this is not feasible for the DLM assessment due to the administration design. The number of items tested per linkage level, and thus the total possible sum score, varies between integrated and year-end assessments and by student, depending on which testlet or testlets were administered. Thus, the expected score for a master would be dependent on which assessment design the master was tested in and which testlets the student received.

Because of this, the item-level indices have to be aggregated up to the model level using different methodology. This evaluation takes advantage of the additive properties of χ^2 distributions (Lancaster & Seneta, 2005) in that the sum of χ^2 values is also χ^2 distributed, with degrees of freedom equal to the sum of degrees of freedom from the component χ^2 values. Take, for example, an attribute measured by five items. As seen in Table 24, five univariate χ^2 values could be estimated (one for each item), and each test would have one degree of freedom. Aggregating to the model level, the univariate model-level fit could be assessed by a χ^2 value equal to the sum of the five item-level indices with five degrees of freedom, as illustrated. In Table 24, the χ^2 test at the model level is nonsignificant, indicating acceptable model fit.

Table 24. Example Model-Level Univariate Fit

Item	χ^2	**df**	***p* value**
1	0.16	1	.68
2	1.20	1	.27
3	0.87	1	.35
4	0.98	1	.32
5	1.03	1	.31
Model	4.24	5	.52

Note. df = degrees of freedom

There are several limitations to this approach. First, χ^2 is only perfectly additive asymptotically. Given the low sample sizes on many of the items, this is an assumption that is unlikely to hold. Additionally, for χ^2 to be additive asymptotically, each χ^2 value must be independent of one another. For the univariate index, the assumption holds, as item responses are assumed to be independent conditional upon mastery status. However, this is clearly not met when the bivariate or trivariate indices are aggregated in a manner similar to Table 24. This is because the same item would be included in multiple item-level bivariate and trivariate indices. Therefore, the true sampling distribution of the aggregated χ^2 is unclear. Because of these limitations, Rupp, Templin, and Henson (2010) suggested using only the value of the aggregated χ^2 as an overall index of model fit, with larger values indicating worst fit.

Because of the number of linkage levels that must be estimated for each model (i.e., fungible and nonfungible), it would be difficult to summarize the aggregated χ^2 in any meaningful way. This is due in no small part to the fact that the magnitude of χ^2 is dependent on the number of indices that contributed to the sum. For example, an aggregated χ^2 of 8.00 that came from 10 univariate tests, each with one degree of freedom, seems much more reasonable than if that number came from only two univariate tests. Therefore, to help summarize the findings in a useful way, p values are calculated for the model level χ^2 values, even though the asymptotic distribution is likely incorrect. This p value is used only as a flagging criterion to give a general idea of how much misfit exists across multiple linkage levels (i.e., content area or level), not to make decisions about individual linkage levels specifically. The literature suggests that p values calculated from this reference asymptotic distribution are overly conservative, leading to the rejection of correctly specified models (Maydeu-Olivares & Joe, 2014). Therefore, when using this p value, it is likely that more misfit is identified than is actually present. Maydeu-Olivares and Joe (2014) propose the use of the M_2 statistic, which combines information from multiple indices (e.g., univariate, bivariate, trivariate) in a way that allows for hypothesis tests with expected Type I error rates. However, given the sparseness of DLM data, bivariate and trivariate indices cannot be calculated for many linkage levels, making this approach unfeasible.

V.4.C. Procedure for Evaluating Model Fit

V.4.C.i. Data

The estimation of the models used data from the 2015–2016 assessment windows and the 2016–2017 instructionally embedded window. Field test testlets and retired testlets from previous years were not included. Furthermore, the data from the 2014–2015 year were excluded due to the known challenges regarding implementation in the first operational year and the observed differences in student performance when comparing results from 2015 to 2016 and 2017.

V.4.C.ii. Method

Model fit was evaluated using a k-fold cross validation procedure, also known as v-fold cross validation (see Arlot, 2010; Hastie, Tibshirani, & Friedman, 2009). The specific method was a stratified, fivefold procedure, where by the data were divided into five sections and both the fungible and nonfungible models were estimated on four of the five sections. Model fit was then evaluated using the 20% of the data excluded from calibration. This process was repeated five times so that each subsection of the data was used as the validation set once (as demonstrated in Table 25). Before creating the five samples, the data were stratified at the item level to ensure the inclusion of data from all items in each of the subsamples. This approach controlled the variation that would occur due to item exclusion, which would require a more vigorous investigation using a methodology similar to jackknife resampling (see Tukey, 1958).

Table 25. Specification of *k*-Fold Estimation Procedure

Calibration sets	Validation set
2, 3, 4, 5	1
1, 3, 4, 5	2
1, 2, 4, 5	3
1, 2, 3, 5	4
1, 2, 3, 4	5

For each validation set, both absolute and relative fit were evaluated as described above. The results were then averaged across all five validation sets. This approach has the advantages of using all of the data for both estimation and validation, while still evaluating model fit using different data than were used for estimation.

V.4.D. Results

V.4.D.i. Relative Fit

To assess relative fit, a fungible and a nonfungible model were estimated for each of the 1,275 linkage levels. For each linkage level, a likelihood ratio test was computed for the comparison of fungible (null) to nonfungible (alternative) model. For each test, if the *p* value of the likelihood ratio test was less than .05, the null model was rejected, meaning that the nonfungible model demonstrated better fit. The number of linkage levels that performed better in each model was calculated for each of the validation sets and then averaged across the five sets of results. These findings are summarized in Table 26.

Table 26. Average Number of Linkage Levels That Performed Better Over Five Validation Sets

Content area and linkage level	Fungible vs. nonfungible	
	Fungible	**Nonfungible**
English language arts		
Initial Precursor	14.2 (1.3)	133.8 (1.3)
Distal Precursor	8.8 (1.6)	139.2 (1.6)
Proximal Precursor	2.0 (1.2)	146.0 (1.2)
Target	0.8 (0.4)	147.2 (0.4)
Successor	29.2 (4.3)	118.8 (4.3)
Mathematics		
Initial Precursor	2.4 (1.1)	104.6 (1.1)
Distal Precursor	1.8 (0.4)	105.2 (0.4)
Proximal Precursor	1.0 (0.0)	106.0 (0.0)
Target	1.4 (0.5)	105.6 (0.5)
Successor	16.6 (2.1)	90.4 (2.1)

Note. Parentheses indicate the standard deviation across the five validation sets.

The results summarized in Table 26 indicate that the nonfungible model fit the data better than the fungible model for nearly all linkage levels across all subjects. Furthermore, these analyses provide evidence that, as expected, the increase in model fit provided by the extra parameters in the nonfungible model was statistically significant. This is shown by the large discrepancy in the number of linkage levels where the nonfungible model was preferred to the fungible model across content areas and linkage levels.

V.4.D.ii. Absolute Fit

When using the limited information indices of model fit, χ^2 is calculated for each item or set of items within a linkage level (see Table 22 and Table 23). To calculate fit for the entire linkage level, the χ^2 values for each item or set of items are summed (as shown in Table 24). A p value for the linkage level is then calculated for the summed χ^2 values, with degrees of freedom equal to the sum of degrees of freedom from each of the item-level tests (Lancaster & Seneta, 2005). If the p value is less than 0.05, then the expected counts are significantly different from the observed counts, indicating poor model fit. As such, nonsignificant values are desired and significant values are flagged for evidence of poor model fit. Because assumptions of the reference asymptotic distribution are likely not met, the p value may result in the identification of more misfit than is actually present and is therefore used only for flagging to give a general summary of the amount of misfit that could be present in each model.

Because results from the χ^2 test can be unreliable when cell counts are low, a minimum cell count of five was specified for each test. For example, in a linkage level measured by four items, there are six unique combinations of two items, meaning there are six possible bivariate indices. If any of the observed or expected counts for a response pattern in a given index were less than five, that index was not computed. Thus, it is possible that only four of the six possible bivariate indices would be computed. This means that when aggregating the item-level indices to the linkage level, only the four computed indices would be used. For DLM assessments, due to the sparseness of the data and the further sparseness introduced by the *k*-fold procedure, there were some linkage levels where no indices could be computed for the bivariate indices. Further, there were no linkage levels for which trivariate indices were able to be computed, and therefore they are not included in these results. The *k*-fold procedure has the benefit of using different data for the estimation and analysis of model fit. However, when using five folds, the analysis of model fit is limited to only 20% of the total data. This reduced sample size vastly limits the ability to calculate the higher order fit indices.

Table 27 shows the number of linkage levels that were flagged for having poor model fit using each of the methods (univariate and bivariate), as well as the total number of linkage levels for which the index was computed. Results were averaged across all five validation sets. For example, when looking at the bivariate fit for ELA Distal Precursor linkage levels under the fungible model, the bivariate index could be calculated for 44 linkage levels on average, and an average of 14 of those showed poor model fit.

There are several things to note from Table 27. First, as expected, the number of indices that could be computed decreases with added dimensions to the χ^2(i.e., univariate to bivariate indices). This is because with more dimensions, there are more possible response patterns, making it more difficult to obtain the sample size threshold for each. Overall, given the noted constraints, and based on the results that are calculable, the nonfungible model provides the best model fit. The nonfungible model results in the lowest rates of flags across content areas and linkage levels.

In the fungible model, a large proportion of the indices that were computed were flagged due to poor model fit, with an average percent of 77% in the univariate index and 39% in the bivariate index flagged across both content areas and linkage levels. The nonfungible model, on the other hand, showed a fairly low percentage of linkage levels flagged for misfit (18% in the univariate index and 6% in the bivariate index). It also appears that misfit substantially decreased at the higher linkage levels for the nonfungible model (range of 0% to 14% flagged across both indices for Target and Successor, compared to range of 18% to 82% for the fungible model). There is a clear trend of the percentage of indices flagged for misfit decreasing when moving from the Initial Precursor to Successor levels.

Table 27. Average Number of Flagged Linkage Levels Using Limited Information Indices Over Five Validation Sets

Content area and linkage level	Fungible				Nonfungible			
	Univariate		Bivariate		Univariate		Bivariate	
	Flags	*N*	Flags	*N*	Flags	*N*	Flags	*N*
English language arts								
Initial Precursor	104.0 (4.8)	148.0 (0.0)	14.8 (3.1)	48.4 (1.1)	40.6 (3.8)	148.0 (0.0)	4.6 (2.7)	43.4 (1.3)
Distal Precursor	118.4 (1.9)	148.0 (0.0)	14.2 (1.9)	44.0 (2.0)	35.6 (3.4)	147.8 (0.4)	1.6 (1.3)	41.8 (1.5)
Proximal Precursor	117.8 (3.4)	147.0 (0.0)	13.0 (3.4)	31.2 (1.9)	23.2 (2.8)	147.0 (0.0)	1.8 (0.8)	30.2 (2.3)
Target	115.4 (4.5)	141.2 (1.3)	4.2 (0.4)	9.8 (1.1)	19.6 (4.2)	144.8 (0.4)	0.4 (0.5)	9.4 (2.1)
Successor	68.2 (2.6)	114.4 (0.9)	5.6 (1.5)	9.8 (0.8)	12.6 (3.6)	118.4 (0.5)	0.6 (0.5)	9.4 (0.5)
Math								
Initial Precursor	94.4 (1.5)	107.0 (0.0)	44.2 (1.3)	83.6 (2.1)	48.4 (4.7)	107.0 (0.0)	17.0 (2.1)	79.0 (1.2)
Distal Precursor	96.0 (2.3)	107.0 (0.0)	33.6 (3.0)	69.8 (1.5)	12.0 (4.3)	107.0 (0.0)	3.2 (1.3)	69.0 (1.6)
Proximal Precursor	99.2 (0.8)	107.0 (0.0)	21.6 (1.8)	48.0 (0.7)	12.0 (0.7)	107.0 (0.0)	1.8 (1.9)	45.0 (1.2)
Target	80.2 (3.6)	104.8 (0.8)	4.4 (1.7)	18.2 (0.8)	12.6 (2.1)	106.8 (0.4)	0.2 (0.4)	16.2 (1.5)
Successor	44.2 (1.3)	87.0 (0.7)	0.6 (0.5)	3.4 (0.5)	6.0 (0.7)	90.6 (1.1)	0.0 (0.0)	2.4 (0.5)

Note. Parentheses indicate the standard deviation across the five validation sets. There are 148 Essential Elements for English language arts and 107 for mathematics.

V.4.E. *Operational Evaluation of Model Fit*

Statistical significance should not be the only deciding factor when evaluating appropriateness of a psychometric model. Practical significance should also be considered. Specifically, for DLM assessments, the practical significance of model fit results can be gauged by how much performance varies based on the scoring model used. Although *k*-fold cross validation is suggested for model building and evaluation, Hastie et al. (2009) suggest using all the data for the final model to be used operationally. Thus, following the best practices, all five subsets were used to create an operational fungible and nonfungible calibration to be used to assess practical significance.

One way to evaluate the impact on student results is to examine the structural parameter from each linkage level. This represents the base rate probability of mastery for the linkage level, and thus provides information about the proportion of students that are being classified as masters. If students are being classified as masters at similar rates across models, then there is preliminary evidence that student results are not overly impacted by the choice of model.

Figure 14 shows the difference in base rate mastery probabilities across models by content area and linkage level. Generally, mathematics shows the most consistency in mastery rates. On the other hand, ELA shows more variability in the mastery rates. However, across content areas, the variability is most common in the higher linkage levels. The Initial Precursor and Distal Precursor levels appear to be fairly consistent across content areas, whereas the higher levels show increasing variability, particularly at the Successor level, where the fewest number of students test. Given these results, performance in mathematics is expected to be fairly consistent across models, whereas more variability may be expected in ELA if scored with a nonfungible model.

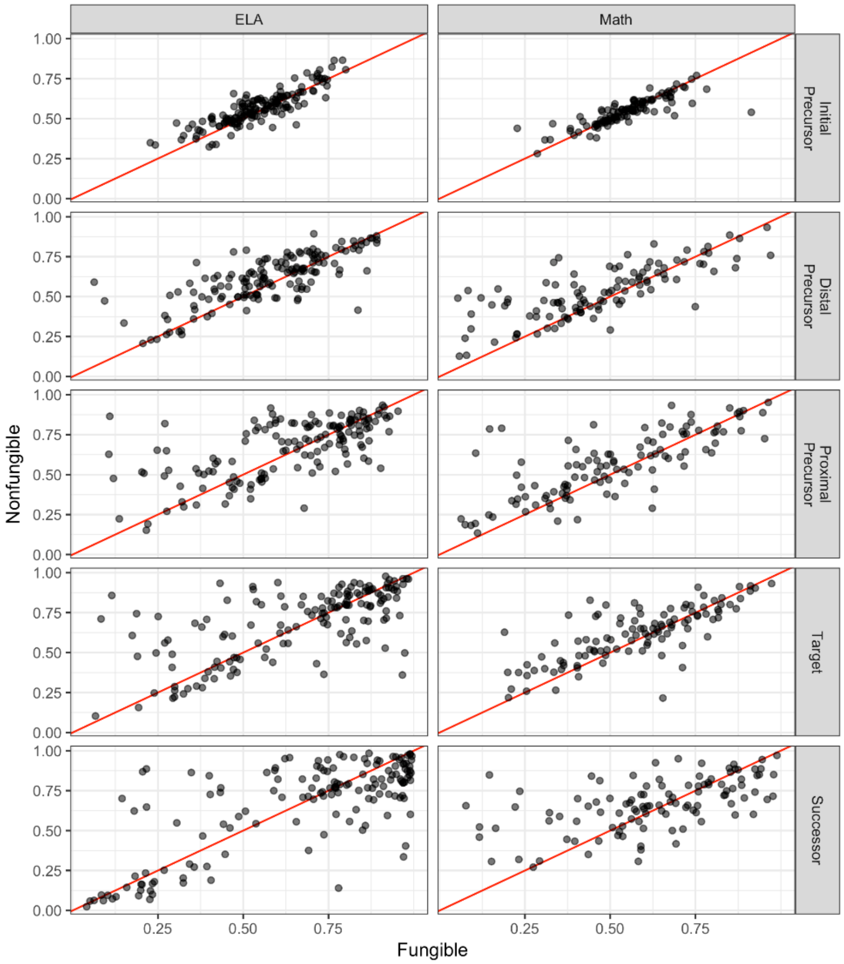


Figure 14. Comparison of base rate mastery probability in the fungible and nonfungible models.

The consistency of results can be examined by comparing the number of linkage levels that are mastered by students when the fungible or nonfungible model is used to score the assessment. As a natural extension to this analysis, the consortium-level impact data can also be compared across the two scoring models. For this analysis, each of the estimated models (i.e., fungible and nonfungible) was used to score the 2016–2017 operational assessment. For each model, the total linkage levels mastered by each student was calculated, and the percent of students at each performance level for each grade and content area was determined using current operational cut points. Figure 15 shows the comparison of total linkage levels mastered, including correlations, for year-end model ELA and mathematics assessments.

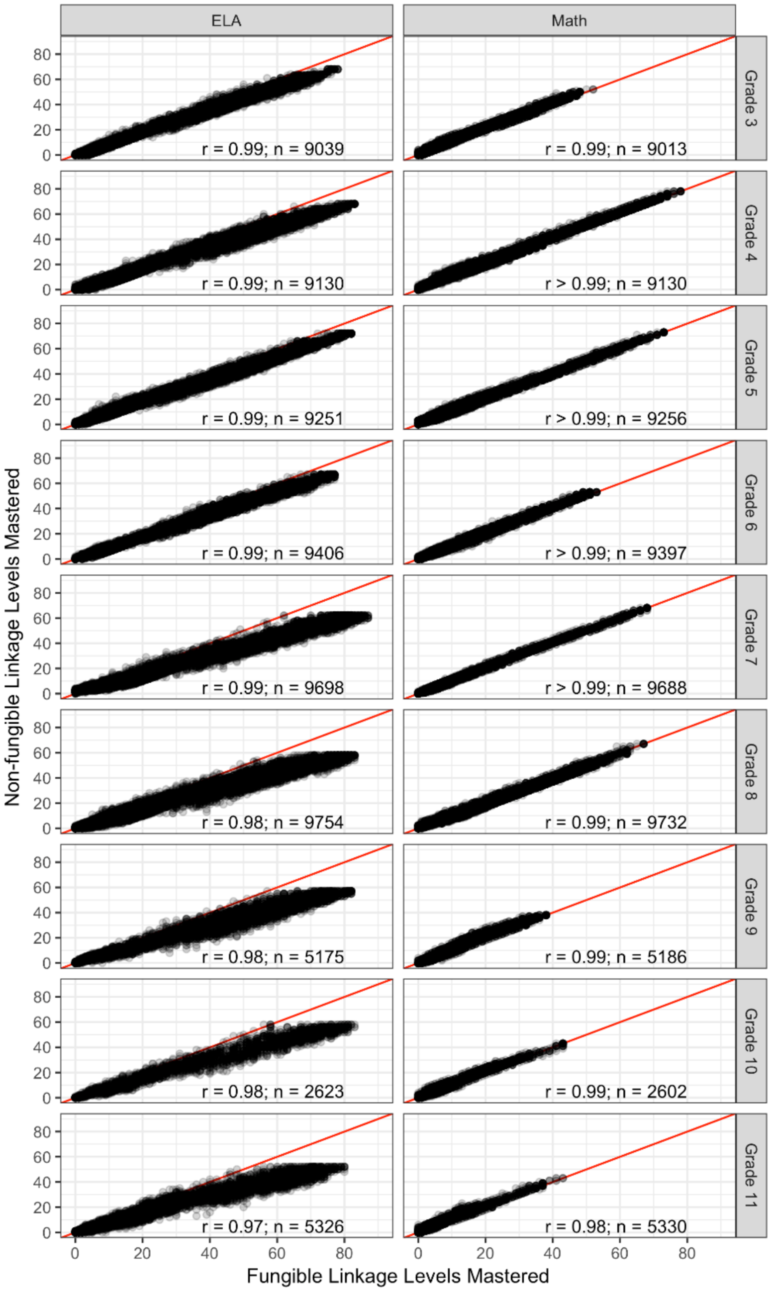
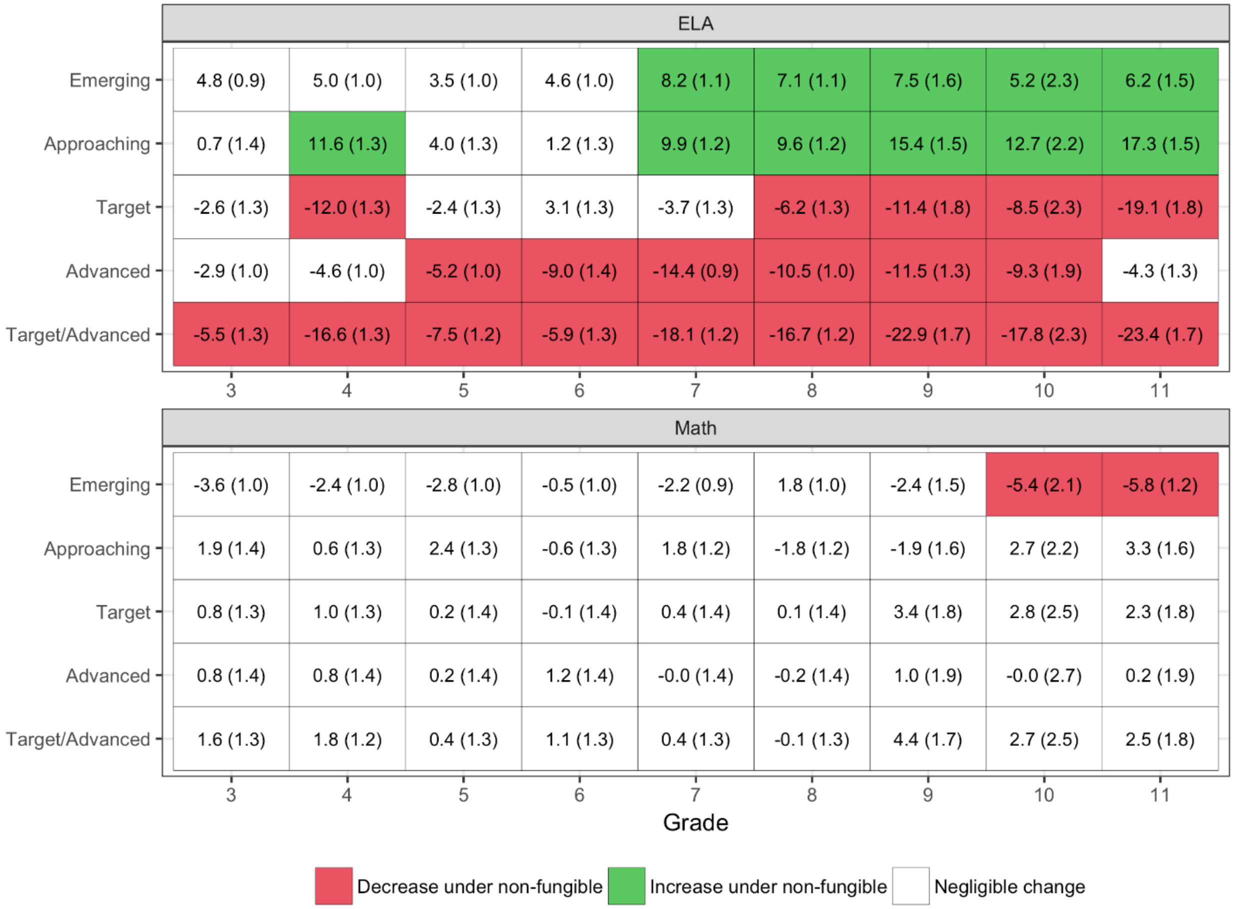


Figure 15. Total linkage levels mastered comparison.

Figure 15 demonstrates that student results are extremely consistent across scoring models, with all correlations greater than .9. The ELA results appear to be slightly less consistent than mathematics across scoring models, which may suggest that ELA overall may be more sensitive to the assumption of fungibility.

A comparison of the percent of students achieving at each performance level is also provided. Figure 16 shows the change in the percent of students at each grade and content area. The combined standard error of the difference is shown in parentheses, as calculated by $\sqrt{\sigma_1^2 + \sigma_2^2}$. For example, in third grade ELA, 58.5% of students achieved at the Emerging category with the fungible model compared to 63.3% with the nonfungible model, for a change of 4.8 percentage points, and a standard error of 0.9 that is interpreted on the scale of the percentages rather than for the difference value itself.



ELA

	3	4	5	6	7	8	9	10	11
Emerging	4.8 (0.9)	5.0 (1.0)	3.5 (1.0)	4.6 (1.0)	🟩 8.2 (1.1)	🟩 7.1 (1.1)	🟩 7.5 (1.6)	🟩 5.2 (2.3)	🟩 6.2 (1.5)
Approaching	0.7 (1.4)	🟩 11.6 (1.3)	4.0 (1.3)	1.2 (1.3)	🟩 9.9 (1.2)	🟩 9.6 (1.2)	🟩 15.4 (1.5)	🟩 12.7 (2.2)	🟩 17.3 (1.5)
Target	-2.6 (1.3)	🟥 -12.0 (1.3)	-2.4 (1.3)	3.1 (1.3)	-3.7 (1.3)	🟥 -6.2 (1.3)	🟥 -11.4 (1.8)	🟥 -8.5 (2.3)	🟥 -19.1 (1.8)
Advanced	-2.9 (1.0)	-4.6 (1.0)	🟥 -5.2 (1.0)	🟥 -9.0 (1.4)	🟥 -14.4 (0.9)	🟥 -10.5 (1.0)	🟥 -11.5 (1.3)	🟥 -9.3 (1.9)	-4.3 (1.3)
Target/Advanced	🟥 -5.5 (1.3)	🟥 -16.6 (1.3)	🟥 -7.5 (1.2)	🟥 -5.9 (1.3)	🟥 -18.1 (1.2)	🟥 -16.7 (1.2)	🟥 -22.9 (1.7)	🟥 -17.8 (2.3)	🟥 -23.4 (1.7)

Math

	3	4	5	6	7	8	9	10	11
Emerging	-3.6 (1.0)	-2.4 (1.0)	-2.8 (1.0)	-0.5 (1.0)	-2.2 (0.9)	1.8 (1.0)	-2.4 (1.5)	🟥 -5.4 (2.1)	🟥 -5.8 (1.2)
Approaching	1.9 (1.4)	0.6 (1.3)	2.4 (1.3)	-0.6 (1.3)	1.8 (1.2)	-1.8 (1.2)	-1.9 (1.6)	2.7 (2.2)	3.3 (1.6)
Target	0.8 (1.3)	1.0 (1.3)	0.2 (1.4)	-0.1 (1.4)	0.4 (1.4)	0.1 (1.4)	3.4 (1.8)	2.8 (2.5)	2.3 (1.8)
Advanced	0.8 (1.4)	0.8 (1.4)	0.2 (1.4)	1.2 (1.4)	-0.0 (1.4)	-0.2 (1.4)	1.0 (1.9)	-0.0 (2.7)	0.2 (1.9)
Target/Advanced	1.6 (1.3)	1.8 (1.2)	0.4 (1.3)	1.1 (1.3)	0.4 (1.3)	-0.1 (1.3)	4.4 (1.7)	2.7 (2.5)	2.5 (1.8)

Grade

🟥 Decrease under non-fungible 🟩 Increase under non-fungible □ Negligible change

Figure 16. Change in percentage of students achieving at each performance level.
Note. Highlighted cells indicate a change of more than 5 percentage points. The standard error of the difference is shown in parentheses.

Similar to the comparison of the structural parameters, the mathematics results are extremely consistent, with only two performance levels flagged for changes of more than five percentage points. ELA showed more change in results, with 22 performance levels being flagged for a change of five percentage points or more. Notably, results shift down in ELA across models, with more students being placed in the Emerging and Approaching levels under the nonfungible model.

V.4.F. Summary of Model Fit Analyses

This chapter presents two methods for evaluating model fit, along with comparisons of the operational impact of results obtained from the competing models. This included a relative fit analysis comparing model-to-model fit of the fungible and nonfungible models, and the absolute fit of each model summarized via univariate and bivariate indices.

Overall, the combination of relative and absolute fit from the limited information tests indicates that the data best support use of a nonfungible model. The nonfungible model showed significantly better fit on the majority of linkage levels when compared to the fungible model, and also showed the fewest number of flags in the univariate and bivariate indices. However, a number of methodological constraints were noted, including using p values to evaluate the model level χ^2 values and limited sample sizes using the k-fold validation approach that call into question their use for operational decision-making purposes. Furthermore, the operational comparison of student results showed that the choice of model had mixed impact on student results, with a substantial decrease in the percent of students at higher performance levels in ELA if scoring with a nonfungible model. Additionally, there are practical benefits to using a more parsimonious model, including simpler and faster estimation for delivering student results on the timeline that states need for accountability decision-making purposes. Finally, the recommendations of the DLM Technical Advisory Committee (TAC) have focused on exploring a Bayesian estimation procedure to help address some of the methodological issues with the current approach to assessing model fit. Specific next steps in the research agenda are to implement a Bayesian estimation technique and reevaluate model fit for both the fungible and nonfungible models. Although the current evidence suggests that the nonfungible model fits the data better than the fungible model does, methodological constraints of the current evaluation, limited and varied impact of model choice on students' results, and the practical benefits of the fungible model have led to the decision to retain the fungible model for operational scoring for the 2017–2018 academic year. Ongoing research is intended to identify an improved modeling strategy and corresponding assessment design. The plan to continue calibrating and scoring DLM assessments using a fungible model for the 2017–2018 administration was discussed with the DLM TAC during the August 2017 partner call, and they indicated support for the plan.

V.5. Conclusion

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

A review of model parameters indicates that for most linkage levels, the conditional probability of masters providing a correct response falls above .5, and for most linkage levels, the conditional probability of non-masters providing a correct response falls below .5. Beginning in spring 2018, test development teams will begin reviewing model-based flagging to identify potential areas that may be introducing construct-irrelevant variance into the calculation of student results.

Preliminary model-fit results indicated mixed support for the use of the current fungible scoring model. Because new modeling strategies have potential to provide better alternatives for the assessment of model fit, current work is focused on developing a Bayesian estimation process for the fungible, nonfungible, and a partially fungible model, whereby a partial equivalency model can be estimated. This approach would support improved methods for the assessment of model fit. Specifically, using Markov chain Monte Carlo estimation would allow for the evaluation of model fit using posterior predictive model checking (Gelman & Hill, 2006; Gelman, Meng, & Stern, 1996). The development of this procedure is underway and, upon its completion, will be disseminated to the DLM TAC modeling subcommittee, a subgroup of TAC members focused on reviewing modeling-specific topic guides, for review.

VI. STANDARD SETTING

The standard setting process for the Dynamic Learning Maps® (DLM®) Alternate Assessment System in 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 VI of the *2014–2015 Technical Manual – Year-End Model* (Dynamic Learning Maps Consortium, 2016b).

VII. ASSESSMENT RESULTS

Chapter VII of the *2014–2015 Technical Manual – Year-End Model* (Dynamic Learning Maps® [DLM®] Consortium, 2016b) 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 (a) 2016–2017 student participation data; (b) final results in terms of the percentage of students at each performance level; and (c) subgroup performance by gender, race, ethnicity, and English learner (EL) status for the 2016–2017 administration year. This chapter also reports the distribution of students by the highest linkage level mastered during 2016–2017. Finally, this chapter describes updates made to score reports, data files, and quality control procedures during the 2016–2017 operational year. For a complete description of score reports and interpretive guides, see Chapter VII of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b).

VII.1. Student Participation

The 2016–2017 assessments were administered to 69,613 students in 10 states and one Bureau of Indian Education school. Counts of students tested in each state are displayed in Table 28. The assessment sessions were administered by 18,189 educators in 9,580 schools and 3,007 school districts.

Table 28. Student Participation by State ($N = 69{,}613$)

State	Students (*n*)
Alaska	624
Colorado	5,222
Illinois	11,595
Miccosukee Indian School	8
New Hampshire	791
New Jersey	10,837
New York	21,619
Oklahoma	5,954
Utah	4,274
West Virginia	1,884
Wisconsin	6,805

Table 29 summarizes the number of students tested in each grade. In grades 3 through 8, over 9,000 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 grade-level participation can be traced to differing state-level policies about the grade in which students are assessed in high school.

Table 29. Student Participation by Grade (*N* = 69,613)

Grade	**Students (*n*)**
3	9,053
4	9,149
5	9,274
6	9,433
7	9,725
8	9,778
9	5,198
10	2,633
11	5,370

Table 30 summarizes the demographic characteristics of the students who participated in the 2016–2017 administration. The majority of participants were male (67%) and white (60%). About 6.5% of students participated in EL services.

Table 30. Demographic Characteristics of Participants

Subgroup	*n*	%
Gender		
Female	22,910	32.91
Male	46,700	67.09
Missing	3	<0.01
Race		
White	41,719	59.93
African American	13,771	19.78
Asian	3,348	4.81
American Indian	2,414	3.47
Alaska Native	244	0.35
Two or more races	7,890	11.33
Native Hawaiian or Pacific Islander	216	0.31
Missing	11	0.02
Hispanic ethnicity		
No	56,915	81.76
Yes	12,675	18.21
Missing	23	0.03
English learner (EL) participation		
Not EL eligible or monitored	65,154	93.59
EL eligible or monitored	4,459	6.41

In addition to the spring administration, instructionally embedded assessments are also made available for teachers to administer to students during the year. Results from these assessments do not contribute to final summative scoring but can be used to guide instructional decision-making. Table 31 summarizes the number of students participating in instructionally embedded testing by state. A total of 400 students took at least one instructionally embedded testlet during the 2015–2016 academic year.

Table 31. Students Completing Instructionally Embedded Testlets, by State (*N* = 400)

State	*n*
Colorado	31
Illinois	30
New York	19
Oklahoma	212
Utah	33
West Virginia	75

Table 32 and Table 33 summarize the number of instructionally embedded test sessions taken in ELA and mathematics, respectively. Across all states, students took 2,033 ELA testlets and 2,476 mathematics testlets.

Table 32. Number of Instructionally Embedded English Language Arts Test Sessions, by Grade (*N* = 2,033)

Grade	*n*
3	243
4	167
5	257
6	296
7	438
8	485
9	18
10	80
11	49

Table 33. Number of Instructionally Embedded Mathematics Test Sessions, by Grade (*N* = 2,476)

Grade	*n*
3	226
4	202
5	280
6	234
7	426
8	489
9	42
10	453
11	124

VII.2. Student Performance

Student performance on DLM assessments is interpreted using cut points, determined during standard setting (see Chapter VI in the *2014–2015 Technical Manual – Year-End Model* [DLM Consortium, 2016b]), which separate student scores into four performance levels. A student receives a performance level based on the total number of linkage levels mastered across the assessed Essential Elements (EEs).

For the 2016–2017 administration, student performance was reported using the same four performance levels approved by the DLM Consortium for the previous two 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.

VII.2.A. Overall Performance

Table 34 reports the percentage of students at each performance level from the 2016–2017 administration for ELA and mathematics. For ELA, the percentage of students who demonstrated performance at the Target or Advanced levels ranged from approximately 26% to 41%. In mathematics, the percentage of students meeting or exceeding Target expectations ranged from approximately 8% to nearly 32%.

Table 34. Percentage of Students by Grade and Performance Level

Grade	Emerging (%)	Approaching (%)	Target (%)	Advanced (%)	Target+Advanced (%)
English language arts					
3 (*n* = 9,039)	59.1	15.1	22.8	3.0	25.8
4 (*n* = 9,130)	50.3	20.3	25.0	4.4	29.4
5 (*n* = 9,251)	48.3	19.1	27.5	5.2	32.7
6 (*n* = 9,406)	47.9	24.3	17.7	10.2	27.9
7 (*n* = 9,698)	35.1	27.6	25.2	12.0	37.2
8 (*n* = 9,754)	38.2	24.8	26.9	10.1	37.0
9 (*n* = 5,175)	32.1	29.7	27.7	10.5	38.2
10 (*n* = 2,623)	27.3	31.5	34.0	7.2	41.2
11 (*n* = 5,326)	36.6	32.8	25.8	4.8	30.6
Mathematics					
3 (*n* = 9,013)	58.2	15.3	17.9	8.6	26.5
4 (*n* = 9,130)	51.4	17.0	21.5	10.1	31.6
5 (*n* = 9,256)	54.4	24.8	10.8	10.0	20.8
6 (*n* = 9,397)	55.3	26.0	10.2	8.5	18.7
7 (*n* = 9,688)	62.7	24.7	7.5	5.1	12.6
8 (*n* = 9,732)	55.5	30.7	11.1	2.8	13.9
9 (*n* = 5,186)	44.6	34.1	17.3	3.9	21.2
10 (*n* = 2,602)	47.4	34.8	16.2	1.7	17.9
11 (*n* = 5,330)	63.2	28.7	7.9	0.2	8.1

VII.2.B. Subgroup Performance

Data collection for DLM assessments includes demographic data on gender, race, ethnicity, and EL status. Table 35 and Table 36 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 35. Students at Each English Language Arts Performance Level, by Demographic Subgroup (*N* = 69,613)

	Emerging		**Approaching**		**Target**		**Advanced**	
Subgroup	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Gender								
Female	10,016	43.9	5,578	24.4	5,618	24.6	1,613	7.1
Male	20,346	43.7	10,904	23.4	11,694	25.1	3,630	7.8
Missing	n/a	n/a	3	100.0	n/a	n/a	n/a	n/a
Race								
White	17,950	43.2	9,862	23.7	10,606	25.5	3,156	7.6
African American	5,546	40.4	3,470	25.2	3,598	26.2	1,129	8.2
Asian	1,909	57.2	683	20.5	591	17.7	156	4.7
American Indian	832	34.7	577	24.0	747	31.1	245	10.2
Alaska Native	127	52.0	65	26.6	45	18.4	7	2.9
Two or more races	3,882	49.3	1,767	22.4	1,691	21.5	534	6.8
Native Hawaiian or Pacific Islander	113	52.3	57	26.4	32	14.8	14	6.5
Missing	3	27.3	4	36.4	2	18.2	2	18.2
Hispanic ethnicity								
No	24,320	42.9	13,569	23.9	14,410	25.4	4,449	7.8
Yes	6,032	47.8	2,910	23.0	2,897	22.9	793	6.3
Missing	10	45.5	6	27.3	5	22.7	1	4.5
English learner (EL) participation								
Not EL eligible or monitored	28,540	43.9	15,307	23.6	16,195	24.9	4,918	7.6
EL eligible or monitored	1,822	41.0	1,178	26.5	1,117	25.1	325	7.3

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

Table 36. Students at Each Mathematics Performance Level, by Demographic Subgroup (*N* = 69,613)

	Emerging		Approaching		Target		Advanced	
Subgroup	*n*	%	*n*	%	*n*	%	*n*	%
Gender								
Female	13,168	57.7	5,884	25.8	2,651	11.6	1,129	4.9
Male	25,377	54.6	11,371	24.5	6,431	13.8	3,321	7.1
Missing	n/a	n/a	2	100.0	n/a	n/a	n/a	n/a
Race								
White	23,134	55.7	10,674	25.7	5,329	12.8	2,410	5.8
African American	7,234	52.7	3,399	24.8	1,969	14.4	1,117	8.1
Asian	2,142	64.3	625	18.8	364	10.9	199	6.0
American Indian	1,035	43.0	638	26.5	461	19.1	274	11.4
Alaska Native	150	62.0	67	27.7	21	8.7	4	1.7
Two or more races	4,708	59.9	1,799	22.9	912	11.6	443	5.6
Native Hawaiian or Pacific Islander	140	64.8	50	23.1	23	10.6	3	1.4
Missing	2	20.0	5	50.0	3	30.0	n/a	n/a
Hispanic ethnicity								
No	31,150	54.9	14,216	25.1	7,543	13.3	3,786	6.7
Yes	7,380	58.5	3,036	24.1	1,537	12.2	664	5.3
Missing	15	68.2	5	22.7	2	9.1	n/a	n/a
English learner (EL) participation								
Not EL eligible or monitored	36,379	56.1	16,084	24.8	8,324	12.8	4,101	6.3
EL eligible or monitored	2,166	48.7	1,173	26.4	758	17.0	349	7.8

Note. Students were not assessed on any mathematics Essential Elements. n/a = not applicable.

VII.2.C. Linkage Level Mastery

As described earlier in the chapter, overall performance in each content area is calculated based on the number of linkage levels mastered across all EEs. Results indicate the highest linkage level the student mastered based on the scoring method for each EE. This means that a student can be classified as a master of zero, one (Initial Precursor), two (Initial Precursor and Distal Precursor), three (Initial Precursor, Distal Precursor, and Proximal Precursor), four (Initial Precursor, Distal Precursor, Proximal Precursor, and Target), or five (Initial Precursor, Distal Precursor, Proximal Precursor, Target, and Successor) linkage levels. This section summarizes the distribution of students by highest linkage level mastered across all EEs. For each grade, the number of students who showed no evidence of mastery, mastery of the Initial Precursor level, mastery of the Distal Precursor level, mastery of the Proximal Precursor level, mastery of the Target level, and mastery of the Successor level (as the highest level of mastery) was summed across all EEs and divided by the total (number of students assessed times total EEs) to obtain the proportion of students who mastered each linkage level.

Table 37 and Table 38 report the percentage of students who mastered each linkage level as the highest linkage level across all EEs for ELA and mathematics, respectively. For example, across all third grade ELA EEs, the Initial Precursor level was the highest level that students mastered 19% 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 20% in grade 6 to 28% in grade 7. For mathematics, the average percentage of students who mastered the Target or Successor linkage level across all EEs ranged from approximately 4% in grade 11 to 12% in grade 9.

Table 37. Percentage of Students Demonstrating Highest Level Mastered Across English Language Arts Essential Elements, by Grade

	Linkage level					
Grade	**No evidence (%)**	**IP (%)**	**DP (%)**	**PP (%)**	**T (%)**	**S (%)**
3 (*n* = 9,039)	25.2	18.9	18.5	16.3	12.0	9.1
4 (*n* = 9,130)	24.3	16.8	15.5	16.7	12.5	14.3
5 (*n* = 9,251)	21.0	17.6	17.6	17.8	12.0	14.0
6 (*n* = 9,406)	23.5	21.0	19.3	16.4	10.0	9.8
7 (*n* = 9,698)	20.0	19.9	17.6	14.6	11.9	16.1
8 (*n* = 9,754)	24.4	17.7	15.6	15.4	13.4	13.4
9 (*n* = 5,175)	22.8	18.2	13.6	21.3	13.8	10.4
10 (*n* = 2,623)	21.1	18.7	14.4	18.8	13.8	13.3
11 (*n* = 5,326)	27.0	20.6	17.2	14.2	12.4	8.6

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

Table 38. Percentage of Students Demonstrating Highest Level Mastered Across Mathematics Essential Elements, by Grade

	Linkage level					
Grade	**No evidence (%)**	**IP (%)**	**DP (%)**	**PP (%)**	**T (%)**	**S (%)**
3 (*n* = 9,013)	38.4	27.3	14.6	10.6	5.4	3.6
4 (*n* = 9,130)	31.6	25.7	18.0	14.2	5.5	5.0
5 (*n* = 9,256)	35.6	30.3	16.3	9.2	5.1	3.5
6 (*n* = 9,397)	37.1	24.5	15.8	12.8	5.5	4.3
7 (*n* = 9,688)	34.7	30.6	15.1	11.0	5.4	3.3
8 (*n* = 9,732)	33.7	27.1	18.4	11.4	6.7	2.7
9 (*n* = 5,186)	27.8	26.8	20.2	13.4	6.5	5.3
10 (*n* = 2,602)	35.3	30.8	17.9	8.9	4.7	2.4
11 (*n* = 5,330)	46.7	31.0	15.3	3.5	2.8	0.7

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

VII.3. Data Files

Four data files, made available to DLM state partners, summarized results from the 2016–2017 year. Similar to the previous two 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. During the 2016–2017 year, the GRF was restructured to include one row per student per subject, with a corresponding EE crosswalk provided to identify the EE reported in each column, with columns generically named EE1 – EE26.

The DLM Consortium delivered several supplemental files to state partners in addition to the GRF. Consistent with prior years, the Incident File listed students affected by one of the known administration incidents (see Chapter IV of this manual) using the same structure as the prior two years. Similarly, the Special Circumstances File was retained in 2016–2017, which 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 new supplemental file to identify exited students who did not reenroll for the remainder of the window.

Consistent with 2015–2016, state partners were provided with a 2-week review window following delivery of the GRF to invalidate student records. Once state partners submitted final GRFs back to DLM staff, the final GRF was uploaded to Educator Portal.

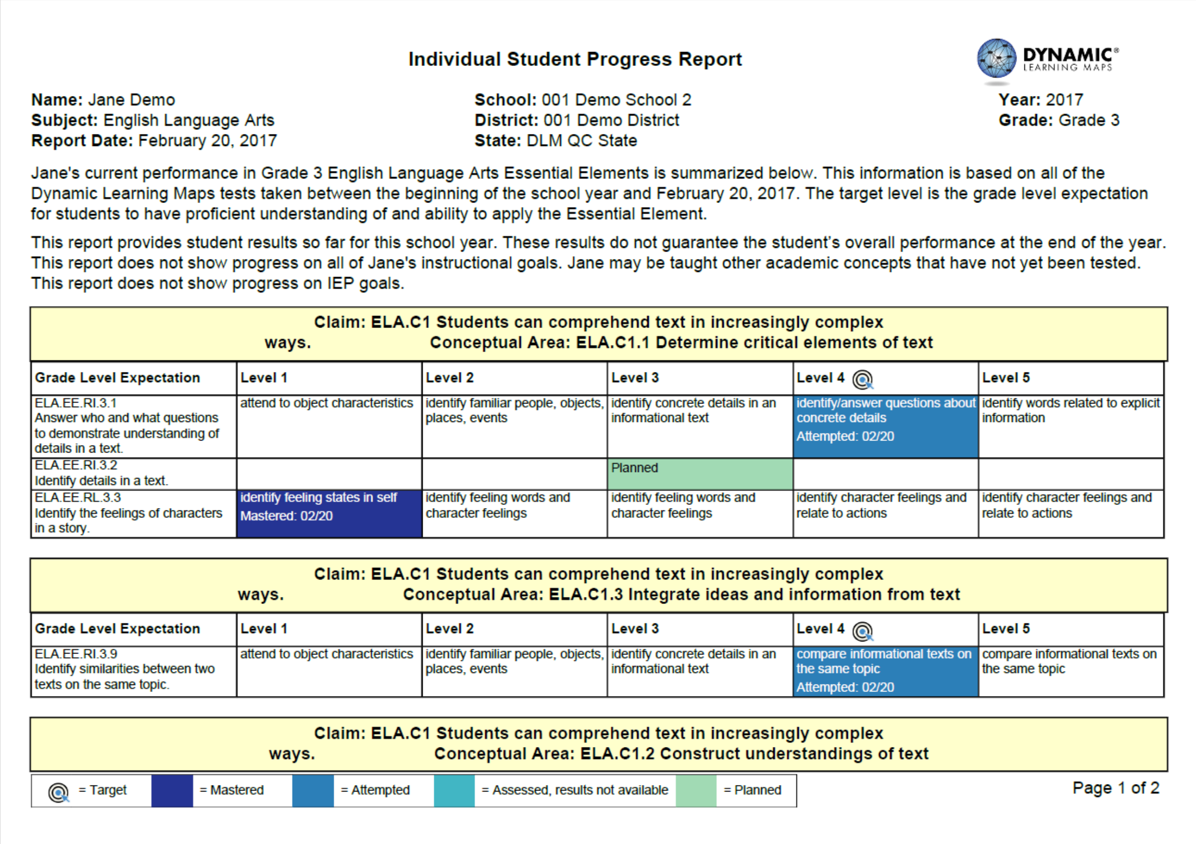
VII.4. Score Reports

The DLM Consortium provides assessment results to all member states to report to parents/guardians and to educators at state and local education agencies. Individual student score reports were provided to educators and parents/guardians. Several aggregated reports were provided to state and local education agencies, including reports for the classroom, school, district, and state. No changes were made to the individual student or aggregated report reports during 2017; however, district and state aggregated reports were generated in Educator Portal following final GRF upload rather than being generated outside the system by the score report program. For a complete description of score reports, including aggregated reports, see Chapter VII of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b).

VII.4.A. Progress Reports

Progress reports are available on demand in Educator Portal for each subject during the instructionally embedded window. An example progress report for a demo student is provided in Figure 17. The report indicates the levels mastered (based on percent correct), levels attempted, and levels assessed but for which results are not yet available (for writing, which is scored external to the system). The progress report also notes the standards and levels for which instructional plans were created, but for which the student has not yet been assessed.

Progress reports are intended to be useful for teachers during instructional planning and for IEP team use in evaluating student progress throughout the year. However, the progress report cautions that results may not be reflective of final summative performance at the end of the year, may not cover all academic concepts taught in the classroom, and does not show progress on specific IEP goals.



Individual Student Progress Report

DYNAMIC LEARNING MAPS

Name: Jane Demo
Subject: English Language Arts
Report Date: February 20, 2017

School: 001 Demo School 2
District: 001 Demo District
State: DLM QC State

Year: 2017
Grade: Grade 3

Jane's current performance in Grade 3 English Language Arts Essential Elements is summarized below. This information is based on all of the Dynamic Learning Maps tests taken between the beginning of the school year and February 20, 2017. The target level is the grade level expectation for students to have proficient understanding of and ability to apply the Essential Element.

This report provides student results so far for this school year. These results do not guarantee the student's overall performance at the end of the year. This report does not show progress on all of Jane's instructional goals. Jane may be taught other academic concepts that have not yet been tested. This report does not show progress on IEP goals.

Claim: ELA.C1 Students can comprehend text in increasingly complex ways. Conceptual Area: ELA.C1.1 Determine critical elements of text

Grade Level Expectation	Level 1	Level 2	Level 3	Level 4 ◎	Level 5
ELA.EE.RI.3.1 Answer who and what questions to demonstrate understanding of details in a text.	attend to object characteristics	identify familiar people, objects, places, events	identify concrete details in an informational text	identify/answer questions about concrete details Attempted: 02/20	identify words related to explicit information
ELA.EE.RI.3.2 Identify details in a text.			Planned		
ELA.EE.RL.3.3 Identify the feelings of characters in a story.	identify feeling states in self Mastered: 02/20	identify feeling words and character feelings	identify feeling words and character feelings	identify character feelings and relate to actions	identify character feelings and relate to actions

Claim: ELA.C1 Students can comprehend text in increasingly complex ways. Conceptual Area: ELA.C1.3 Integrate ideas and information from text

Grade Level Expectation	Level 1	Level 2	Level 3	Level 4 ◎	Level 5
ELA.EE.RI.3.9 Identify similarities between two texts on the same topic.	attend to object characteristics	identify familiar people, objects, places, events	identify concrete details in an informational text	compare informational texts on the same topic Attempted: 02/20	compare informational texts on the same topic

Claim: ELA.C1 Students can comprehend text in increasingly complex ways. Conceptual Area: ELA.C1.2 Construct understandings of text

◎	= Target	■	= Mastered	■	= Attempted	■	= Assessed, results not available	■	= Planned

Page 1 of 2

Figure 17. Example progress report created for a demo student.

VII.5. Quality Control Procedures for Data Files and Score Reports

Quality control procedures were updated in 2017 to include a manual quality control program and to reflect the updated GRF structure of one row, per student, per subject. No changes were made to the manual or automated quality control checks for 2017. For a complete description of quality control procedures, see Chapter VII in the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b) and *2015–2016 Technical Manual – Year-End Model* (DLM Consortium, 2017c).

VII.5.A. Manual Quality Control Program

A PDF viewer tool was developed for 2017 score reporting to increase the speed and efficiency of the manual quality control process. Based on the data file read into the program, reports were selected randomly and individually from a relevant subset (model, grade, and content area). When a report was selected, its data row from the GRF was displayed and the tool automatically opened the report, allowing both to be compared quickly without manually navigating through folders in which reports were stored. After a Quality Control person reviewed the selected report, he or she clicked through; the tool then selected the next report and its corresponding data row for review. This process was repeated until a minimum threshold for number of reports checked was met in the relevant subset.

VIII. RELIABILITY

Chapter VIII of the *2014–2015 Technical Manual – Year-End Model* (Dynamic Learning Maps® [DLM®] Consortium, 2016b) describes the methods used to calculate reliability for the DLM assessment system along with results at three reporting levels. The *2015–2016 Technical Manual Update – Year-End Model* (DLM Consortium, 2017c) 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 2016–2017 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 and a detailed description of the methods used, see the *2015–2016 Technical Manual Update – Year-End Model* (DLM Consortium, 2017c).

VIII.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 [AERA], American Psychological Association, & National Council on Measurement in Education, 2014). Simulation studies were conducted to assemble reliability evidence according to the *Standards*' assertion that "the general notion of reliability/precision is defined in terms of consistency over replications of the testing procedure" (AERA et al., 2014, p. 35). The DLM reliability evidence reported here supports "interpretation for each intended score use," as Standard 2.0 dictates (AERA et al., 2014, p. 42). The "appropriate evidence of reliability/precision" (AERA et al., 2014, p. 42) was assembled using a nontraditional methodology that aligns to the design of the assessment and interpretations of results.

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 VII of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b). 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 content area (content area reliability; provided for ELA and mathematics); (c) the number of linkage levels mastered within each conceptual area for ELA 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.

VIII.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 reliabilities for DLM versions of scores and classifications takes into consideration the unique design of DLM assessments. The use of simulation is necessitated by two factors: the assessment blueprint and the classification-based results that such administrations give. Due to the limited number of items students complete to cover the blueprint, students take only minimal items per EE. The reliability simulation replicates DLM versions of scores from actual examinees based upon the actual set of items each examinee took. Therefore, this simulation provides a replication of the administered items for the examinees. Because the simulation is based on a replication of the exact same items that were administered to examinees, the two administrations are perfectly parallel.

VIII.2.A. Reliability Sampling Procedure

The simulation design that was used to obtain the reliability estimates developed a resampling design to mirror the trends existing in the DLM assessment data. In accordance with Standard 2.1, the sampling design used the entire set of operational testing data to generate simulated examinees. Using this process guarantees that the simulation takes on characteristics of the DLM operational test data that are likely to affect the reliability results. For one simulated examinee, the process was as follows:

1. Draw with replacement the student record of one student from the operational testing data (spring window). Use the student's originally scored pattern of linkage level mastery and 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[4] 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 (see Chapter V of 2015–2016 Technical Manual Update – Year-End Model [DLM Consortium, 2017c] for more information),[5] producing estimates of linkage level mastery or non-mastery for the simulated student.
4. Compare the estimated linkage level mastery or non-mastery to the known values from Step 2 for all linkage levels for which the student was administered items.
5. Repeat Steps 1 through 4 for 2,000,000 simulated students.

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

[5] 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 V of this manual.

Figure 18 shows Steps 1 through 4 of the simulation process as a flow chart.

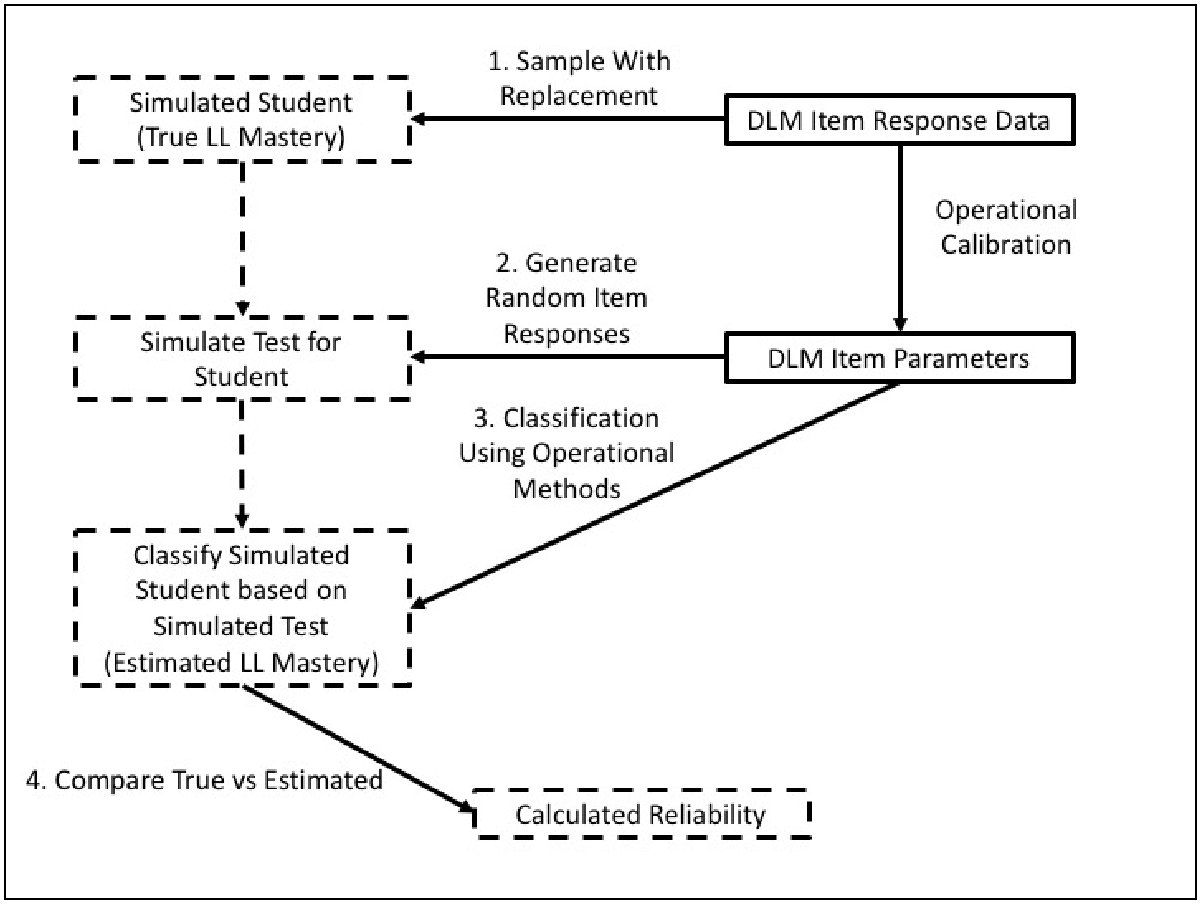

Simulated Student (True LL Mastery)
1. Sample With Replacement
DLM Item Response Data
Operational Calibration
2. Generate Random Item Responses
Simulate Test for Student
DLM Item Parameters
3. Classification Using Operational Methods
Classify Simulated Student based on Simulated Test (Estimated LL Mastery)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

Figure 18. Simulation process for creating reliability evidence.
Note. LL = linkage level.

VIII.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) content area reliability, (c) conceptual area reliability, (d) EE reliability, (e) linkage level reliability, and (f) conditional reliability by linkage level. With 242 EEs, each with 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 provides a full report of reliability evidence for all 1,210 linkage levels and 242 EEs (http://dynamiclearningmaps.org/reliabevid). The full set of evidence is furnished in accordance with Standard 2.12.

Reporting reliability at six levels ensures that the simulation and resulting reliability evidence were conducted in accordance with Standard 2.2. Additionally, providing reliability evidence for each of the six levels ensures that these reliability estimation procedures meet Standard 2.5.

VIII.3.A. PERFORMANCE LEVEL RELIABILITY EVIDENCE

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

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

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

Table 39 presents this information across all grades and content areas. Polychoric correlations between true and estimated performance levels range from .960 to .991. Correct classification rates range from .831 to .925 and Cohen's kappa values are between .852 and .960. These results indicate that the DLM scoring procedure of assigning and reporting performance levels based on total linkage levels mastered results in reliable classification of students to performance level categories.

Table 39. Summary of Performance Level Reliability Evidence

Grade	Content area	Polychoric correlation	Correct classification rate	Cohen's kappa
3	English language arts	.982	.921	.952
3	Mathematics	.985	.884	.932
4	English language arts	.985	.914	.951
4	Mathematics	.991	.897	.953
5	English language arts	.987	.925	.960
5	Mathematics	.985	.874	.933
6	English language arts	.989	.883	.941
6	Mathematics	.988	.884	.937
7	English language arts	.987	.880	.941
7	Mathematics	.985	.894	.924
8	English language arts	.988	.883	.941
8	Mathematics	.988	.917	.933
9	English language arts	.987	.879	.935
9	Mathematics	.986	.880	.915
10	English language arts	.964	.914	.924
10	Mathematics	.960	.831	.854
11	English language arts	.985	.906	.931
11	Mathematics	.964	.884	.852

VIII.3.B. CONTENT AREA RELIABILITY EVIDENCE

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

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

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

The correct classification kappa for which linkage levels were mastered as averaged across all simulated students Table 40 shows the three summary values for each grade and content area. Classification rate information is provided in accordance with Standard 2.16. The two summary statistics included in Table 40 also meet Standard 2.19. The correlation between true and estimated number of linkage levels mastered, ranges from .965 to .993. Average student correct classification rates range from .949 to .989 and average student Cohen’s kappa values ranges from .838 to .973. These values indicate the DLM scoring procedure of reporting the number of linkage levels mastered provides reliable results of student performance.

Table 40. Summary of Content Area Reliability Evidence

Grade	Content area	Linkage levels mastered correlation	Average student correct classification	Average student Cohen's kappa
3	English language arts	.991	.966	.896
3	Mathematics	.979	.974	.916
4	English language arts	.992	.959	.870
4	Mathematics	.989	.965	.889
5	English language arts	.993	.965	.895
5	Mathematics	.987	.966	.879
6	English language arts	.989	.957	.872
6	Mathematics	.983	.971	.911
7	English language arts	.989	.951	.857
7	Mathematics	.985	.968	.897
8	English language arts	.989	.949	.838
8	Mathematics	.985	.968	.904
9	English language arts	.988	.953	.860
9	Mathematics	.978	.985	.969
10	English language arts	.985	.957	.867
10	Mathematics	.972	.985	.966
11	English language arts	.987	.958	.876
11	Mathematics	.965	.989	.973

VIII.3.C. CONCEPTUAL AREA RELIABILITY EVIDENCE

Within each content area, students are assessed on multiple content strands. These strands of related EEs describe the overarching sections of the learning map model upon which DLM assessments are developed (see Chapter II in *2014-2015 Technical Manual – Year-End Model* [DLM Consortium, 2016b] for more information). 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 VII of this manual for more information), reliability evidence is also provided for each conceptual area.

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

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

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

Table 41 and Table 42 show the three summary values for each conceptual area, by grade, for ELA and mathematics, respectively. Values range from .705 to .999 in ELA and from .587 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 41. Summary of English Language Arts Conceptual Area Reliability Evidence

Grade	Conceptual area	Linkage levels mastered correlation	Average student correct classification	Average student Cohen's kappa
3	ELA.C1.1	.978	.983	.965
3	ELA.C1.2	.973	.986	.971
3	ELA.C1.3	.932	.996	.995
3	ELA.C2.1	.913	.995	.993
4	ELA.C1.1	.980	.981	.955
4	ELA.C1.2	.974	.974	.935
4	ELA.C1.3	.927	.999	.999
4	ELA.C2.1	.971	.996	.995
5	ELA.C1.1	.960	.994	.992
5	ELA.C1.2	.985	.979	.949
5	ELA.C1.3	.963	.991	.985
5	ELA.C2.1	.932	.997	.996
6	ELA.C1.1	.705	.998	.998
6	ELA.C1.2	.981	.964	.903
6	ELA.C1.3	.956	.994	.990
6	ELA.C2.1	.910	.997	.996
7	ELA.C1.1	.779	.998	.997

Grade	Conceptual area	Linkage levels mastered correlation	Average student correct classification	Average student Cohen's kappa
7	ELA.C1.2	.983	.977	.945
7	ELA.C1.3	.968	.990	.982
7	ELA.C2.1	.879	.978	.958
8	ELA.C1.2	.980	.958	.874
8	ELA.C1.3	.944	.991	.986
8	ELA.C2.1	.954	.986	.973
9	ELA.C1.2	.982	.970	.921
9	ELA.C1.3	.937	.990	.983
9	ELA.C2.1	.884	.985	.974
9	ELA.C2.2	.899	.996	.995
10	ELA.C1.2	.984	.967	.907
10	ELA.C1.3	.931	.989	.982
10	ELA.C2.1	.893	.991	.986
10	ELA.C2.2	.916	.997	.997
11	ELA.C1.2	.975	.974	.939
11	ELA.C1.3	.957	.986	.973
11	ELA.C2.1	.938	.989	.980
11	ELA.C2.2	.871	.995	.994

Table 42. 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	.925	.995	.993
3	M.C1.3	.874	.998	.998
3	M.C2.2	.834	.998	.998
3	M.C3.1	.909	.995	.993
3	M.C3.2	.838	.998	.998
3	M.C4.1	.931	.996	.994
3	M.C4.2	.692	.998	.998
4	M.C1.1	.874	.997	.996

Grade	Conceptual area	Linkage levels mastered correlation	Average student correct classification	Average student Cohen's kappa
4	M.C1.2	.837	.994	.992
4	M.C1.3	.904	.999	.998
4	M.C2.1	.941	.993	.990
4	M.C2.2	.897	.999	.999
4	M.C3.1	.955	.996	.994
4	M.C3.2	.779	.998	.998
4	M.C4.1	.897	.995	.993
4	M.C4.2	.587	.997	.997
5	M.C1.1	.795	.994	.992
5	M.C1.2	.942	.993	.989
5	M.C1.3	.938	.997	.996
5	M.C2.1	.941	.997	.996
5	M.C2.2	.911	.999	.999
5	M.C3.1	.936	.993	.989
5	M.C3.2	.870	.998	.998
5	M.C4.2	.724	.998	.997
6	M.C1.1	.865	.999	.998
6	M.C1.2	.890	.995	.993
6	M.C1.3	.932	.996	.995
6	M.C2.2	.935	.997	.996
6	M.C3.2	.845	.998	.998
6	M.C4.1	.883	.991	.985
7	M.C1.1	.905	.995	.994
7	M.C1.2	.825	.998	.998
7	M.C1.3	.925	.993	.990
7	M.C2.1	.950	.996	.995
7	M.C2.2	.852	.998	.998

Grade	Conceptual area	Linkage levels mastered correlation	Average student correct classification	Average student Cohen's kappa
7	M.C3.2	.918	.997	.997
7	M.C4.1	.798	.998	.998
7	M.C4.2	.784	.998	.998
8	M.C1.1	.661	.996	.996
8	M.C1.2	.851	.998	.998
8	M.C1.3	.956	.998	.997
8	M.C2.1	.906	.987	.976
8	M.C2.2	.900	.999	.999
8	M.C3.2	.890	.998	.998
8	M.C4.1	.936	.999	.999
8	M.C4.2	.932	.991	.985
9	M.C1.3	.941	.994	.991
9	M.C2.1	.919	.996	.995
9	M.C2.2	.850	.999	.998
9	M.C4.1	.790	.996	.995
10	M.C1.3	.860	.999	.998
10	M.C2.1	.891	.999	.999
10	M.C3.1	.785	.998	.998
10	M.C3.2	.902	.997	.996
10	M.C4.1	.840	.997	.996
10	M.C4.2	.872	.996	.996
11	M.C1.3	.899	.997	.996
11	M.C1.3	.775	.998	.998
11	M.C2.1	.860	.999	.999
11	M.C3.2	.942	.995	.991

VIII.3.D. ESSENTIAL ELEMENT RELIABILITY EVIDENCE

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

Three statistics are used to summarize reliability evidence for EEs.

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

Because there are 242 EEs, the summaries reported herein are based on the number and proportion of EEs that fall within a given range of an index value. Results are given in both tabular and graphical forms. Table 43 and Figure 19 provide proportions and the number of EEs, respectively, falling within prespecified ranges of values for the three reliability summary statistics (i.e., correct classification rate, kappa, and correlation). In general, the reliability summaries for number of linkage levels mastered within EEs show strong evidence of reliability.

Table 43. Reliability Summaries Across All Essential Elements: Proportion of Essential Elements Falling Within a Specified Index Range

Reliability index	**Index range**								
	<.60	**.60–.64**	**.65–.69**	**.70–.74**	**.75–.79**	**.80–.84**	**.85–.89**	**.90–.94**	**.95–1.0**
Polychoric correlation	.000	.000	.004	.000	.021	.041	.223	.504	.207
Correct classification rate	.000	.000	.000	.041	.132	.351	.413	.054	.008
Kappa	.000	.004	.004	.033	.045	.174	.318	.384	.037

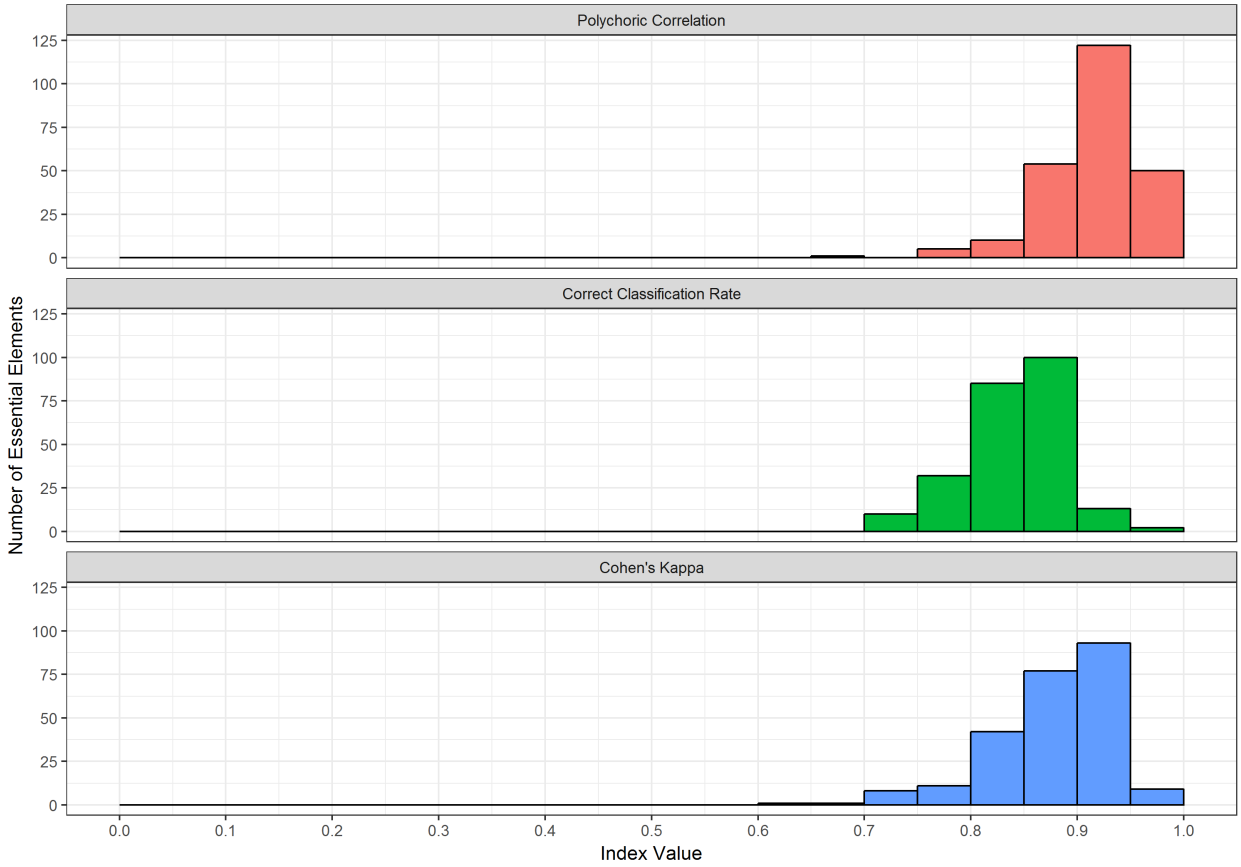


Figure 19. Number of linkage levels mastered within Essential Element reliability summaries.

VIII.3.E. LINKAGE LEVEL RELIABILITY EVIDENCE

Evidence at the linkage level comes from the comparison of true and estimated mastery statuses for each of the 1,210 linkage levels in the operational DLM assessment.[6] This level of reliability reporting is even more fine-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 where mastery classifications are made for DLM assessments. As one example, Table 44 shows a simulated table from one linkage level of an EE.

[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 III – Pilot Administration: Initialization and Chapter IV – Adaptive Delivery in the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b).

Table 44. Example of True and Estimated Mastery Status From Reliability Simulation

		Estimated mastery status	
		Non-master	**Master**
True mastery status	**Non-master**	574	235
	Master	83	592

The reported summary statistics are all based on tables like this one: the comparison of true and estimated mastery statuses across all simulated examinees. As with any contingency table, a number of summary statistics are possible.

For each statistic, figures are given comparing the results of all 1,210 linkage levels. Three summary statistics are presented:

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

As there are 1,210 total linkage levels across all 242 EEs, the summaries reported herein are based on the proportion and number of linkage levels that fall within a given range of an index value. Results are given in both tabular and graphical form. Table 45 and Figure 20 provide proportions and number of linkage levels, respectively, that fall within prespecified ranges of values for the three reliability summary statistics (i.e., correct classification rate, kappa, and correlation). The kappa value for 37 linkage levels and the tetrachoric correlation for 39 linkage levels 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 50 had a kappa value below .6.

Table 45. Reliability Summaries Across All Linkage Levels: Proportion of Linkage Levels Falling Within a Specified Index Range

Reliability index	Index range								
	<.60	**.60–.64**	**.65–.69**	**.70–.74**	**.75–.79**	**.80–.84**	**.85–.89**	**.90–.94**	**.95–1.0**
Tetrachoric correlation	.003	.003	.001	.000	.007	.015	.051	.158	.762
Correct classification rate	.000	.000	.000	.000	.002	.018	.108	.385	.486
Kappa	.043	.032	.048	.079	.123	.205	.215	.150	.106

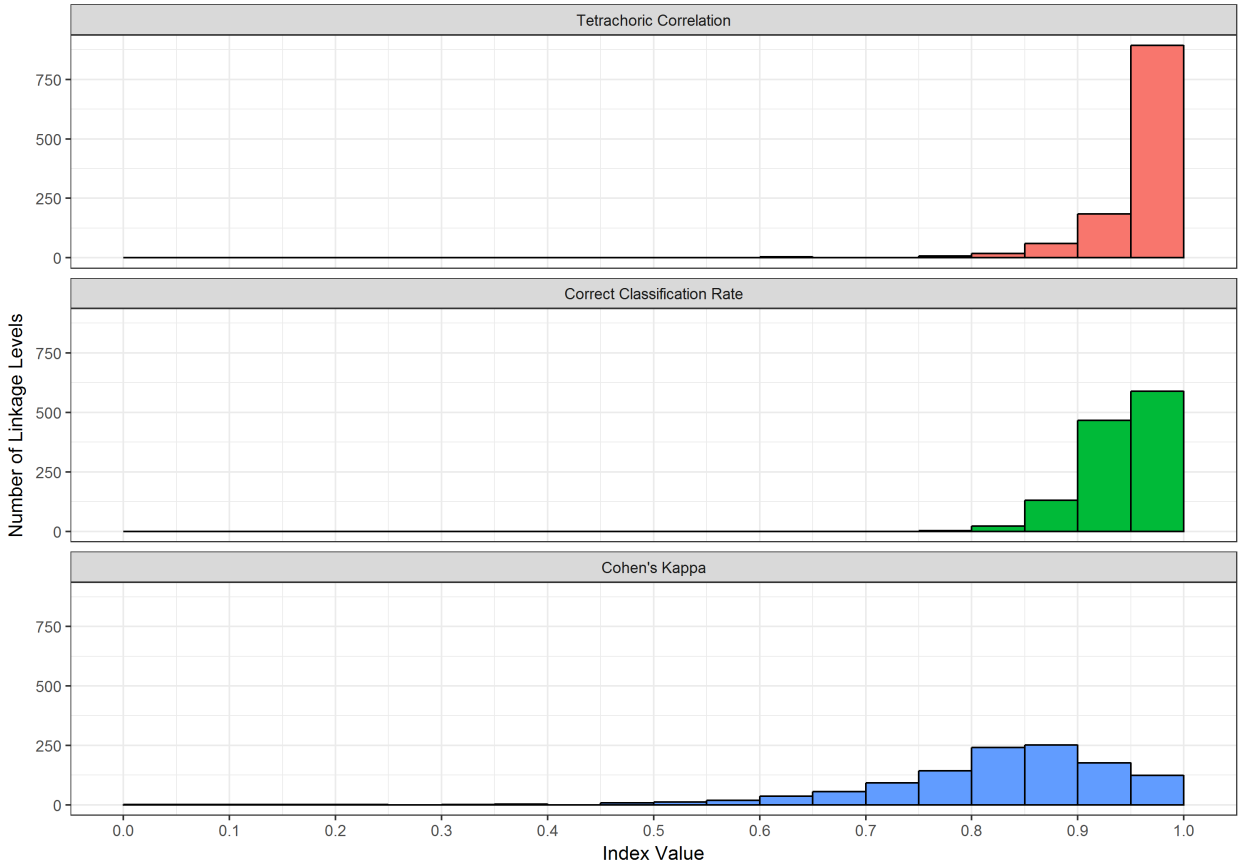


Figure 20. Linkage level reliability summaries.

VIII.3.F. CONDITIONAL RELIABILITY EVIDENCE BY LINKAGE LEVEL

Traditional assessment programs often report conditional standard errors of measurement to indicate how the precision of measurement differs along the score continuum. The DLM assessment system does not report total or scale score values. However, because DLM assessments were designed to span the continuum of students' varying skills and abilities as defined by the five linkage levels, evidence of reliability can be summarized for each linkage level to approximate conditional evidence over all EEs, similar to a conditional standard error of measurement for a total score.

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

Figure 21 provides the number of linkage levels that fall within prespecified ranges of values for the three reliability summary statistics (i.e., tetrachoric correlation, correct classification rate, and kappa). The correlations and correct classification rates generally indicate that all five linkage levels provide reliable classifications of student mastery, with results being fairly consistent across all linkage levels for each of the three statistics reported.

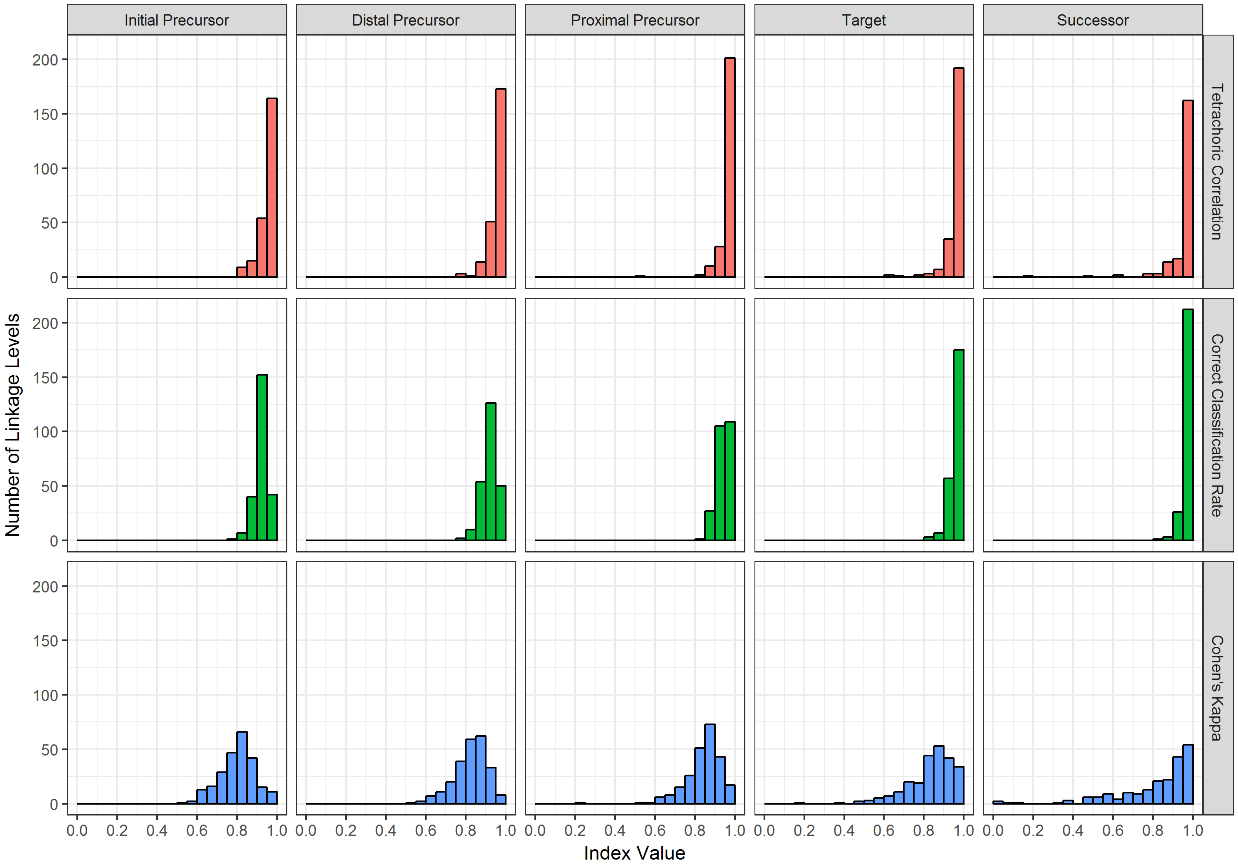


Figure 21. Conditional reliability evidence summarized by linkage level.

VIII.4. CONCLUSION

In summary, reliability measures for the DLM assessment system addresses the standards set forth by AERA et al., 2014. The DLM methods are consistent with assumptions of diagnostic classification modeling and yielded evidence to support the argument for internal consistency of the program for each level of reporting. Because the reliability results are dependent upon the model used to calibrate and score the assessment, any changes to the model or evidence obtained when evaluating model fit would also 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 exact same test items (perfectly parallel forms) which theoretically reduces the amount of variance that may be found in test scores across administrations. Furthermore, while results in general may be higher than those observed for some traditionally scored assessments, research suggests that DCMs have higher reliability with fewer items (e.g., Templin & Bradshaw, 2013), suggesting the results are expected.

IX. VALIDITY STUDIES

The preceding chapters and the *2014–2015 Technical Manual – Year-End Model* (Dynamic Learning Maps® [DLM®] Consortium, 2016b) provide evidence in support of the overall validity argument for results produced by the Dynamic Learning Maps (DLM) Alternate Assessment System. Chapter IX presents additional evidence collected during 2016–2017 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 IX of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b) and the *2015–2016 Technical Manual – Year-End Model* (DLM Consortium, 2017c).

IX.1. EVIDENCE BASED ON TEST CONTENT

Evidence based on test content relates to the evidence "obtained from an analysis of the relationship between the content of the test and the construct it is intended to measure" (AERA et al., 2014, p. 14). The validity study presented in this section summarizes data collected during 2016–2017 regarding student opportunity to learn the assessed content. For additional evidence based on test content, including the alignment of test content to content standards via the DLM maps (which underlie the assessment system), see Chapter IX of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b).

IX.1.A. OPPORTUNITY TO LEARN

After completing administration of the spring 2017 operational assessments, teachers were invited to complete a survey about the assessment administration process (see Chapter IV 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 survey served several purposes.[7] One item provided preliminary 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 46 reports the results. Approximately 68% of teachers (n = 17,185) reported that most or all reading testlets matched instruction, compared to 41% (n = 10,133) for writing and 55% (n = 13,868) for mathematics. More specific measures of instructional alignment are planned.

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

Table 46. Teacher Ratings of Portion of Testlets That Matched Instruction

	None		Some (< half)		Most (> half)		All		Did not administer	
Subject	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Reading	1,612	6.4	5,679	22.6	10,450	41.6	6,735	26.8	656	2.6
Writing	2,921	11.8	5,469	22.1	6,414	26.0	3,719	15.0	6,191	25.1
Mathematics	2,479	9.9	7,935	31.7	9,120	36.4	4,748	19.0	751	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*, *5–10 hours*, *10–15 hours*, *15–20 hours*, or *more than 20 hours*. Table 47 and Table 48 indicate the amount of instructional time spent on conceptual areas, for ELA and mathematics, respectively. For all ELA conceptual areas and most mathematics conceptual areas, the most commonly selected response was *more than 20 hours*. Using 10 or more hours per conceptual area as a criterion for instruction, about 70% of the teachers provided this amount of instruction to their students in ELA, although only about 60% or less did so in mathematics.

Table 47. Instruction Time Spent on English Language Arts Conceptual Areas, in Hours

	Number of hours									
	0–5		**5–10**		**10–15**		**15–20**		**>20**	
Conceptual area	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Determine critical elements of text	741	16.5	505	11.2	582	13.0	744	16.6	1,919	42.7
Construct understandings of text	526	11.7	444	9.9	540	12.0	756	16.9	2,217	49.5
Integrate ideas and information from text	607	13.6	508	11.4	645	14.4	842	18.9	1,863	41.7
Use writing to communicate	833	18.6	570	12.7	662	14.8	751	16.7	1,671	37.2
Integrate ideas and information in writing	991	22.1	608	13.6	655	14.6	767	17.1	1,457	32.5
Use language to communicate with others	280	6.2	286	6.4	376	8.4	585	13.0	2,967	66.0
Clarify and contribute in discussion	500	11.2	476	10.6	544	12.1	811	18.1	2,151	48.0
Use sources and information	991	22.1	680	15.2	693	15.4	810	18.1	1,313	29.3
Collaborate and present ideas	985	22.0	689	15.4	707	15.8	789	17.6	1,317	29.4

Note. Only the first five conceptual areas listed in this table are measured by the DLM assessment. For more information on the English language arts blueprint, see Chapter III of *2014-2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b).

Table 48. Instruction Time Spent on Mathematics Conceptual Areas, in Hours

	Number of hours									
	0–5		5–10		10–15		15–20		>20	
Conceptual area	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Understand number structures (counting, place value, fraction)	493	10.9	433	9.6	526	11.6	697	15.4	2,380	52.6
Compare, compose, and decompose numbers and steps	957	21.2	664	14.7	713	15.8	820	18.2	1,354	30.0
Calculate accurately and efficiently using simple arithmetic operations	841	18.7	419	9.3	530	11.8	671	14.9	2,044	45.4
Understand and use geometric properties of two- and three-dimensional shapes	1,306	28.9	962	21.3	825	18.3	751	16.6	671	14.9
Solve problems involving area, perimeter, and volume	2,278	50.4	732	16.2	527	11.7	498	11.0	485	10.7
Understand and use measurement principles and units of measure	1,356	29.9	1,004	22.1	774	17.1	737	16.2	668	14.7
Represent and interpret data displays	1,380	30.6	911	20.2	771	17.1	761	16.9	689	15.3
Use operations and models to solve problems	1,075	23.8	738	16.3	704	15.6	770	17.1	1,228	27.2
Understand patterns and functional thinking	843	18.7	876	19.4	882	19.5	849	18.8	1,065	23.6

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 EE, it is generally expected that students who mastered more linkage levels in the area would also have spent more instructional time in the area.

Table 49 and Table 50 summarize the Pearson correlations between conceptual area instructional time and linkage levels mastered in the conceptual area for ELA and mathematics, respectively. Based on guidelines from Cohen (1988), the observed correlations fell in the small to medium range, with the strongest correlation observed for writing conceptual areas in ELA.

Table 49. Correlation Between Instruction Time in English Language Arts Conceptual Area and Linkage Levels Mastered in That Conceptual Area

Conceptual area	Correlation with instruction time
Determine critical elements of text	.19
Construct understandings of text	.27
Integrate ideas and information from text	.26
Use writing to communicate	.33
Integrate ideas and information in writing	.35

Table 50. Correlation Between Mathematics Conceptual Area Instruction Time and Linkage Levels Mastered in That Conceptual Area

Conceptual area	Correlation with instruction time
Understand number structures (counting, place value, fraction)	.13
Compare, compose, and decompose numbers and steps	.24
Calculate accurately and efficiently using simple arithmetic operations	.31
Understand and use geometric properties of two- and three- dimensional shapes	.19
Solve problems involving area, perimeter, and volume	.27
Understand and use measurement principles and units of measure	.20
Represent and interpret data displays	.25
Use operations and models to solve problems	.27
Understand patterns and functional thinking	.12

IX.2. Evidence Based on Response Processes

The study of the response processes of test takers 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 2017 regarding students' abilities to respond to testlets, test administration observation data collected during 2016–2017, and a study of interrater agreement on the scoring of teacher-administered writing products. For additional evidence based on response process, including studies on student and teacher behaviors during testlet administration and evidence of fidelity of administration, see Chapter IX of the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b).

IX.2.A. Evaluation of Test Administration

After administering spring operational assessments in 2017, 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 2017 teacher survey included three items about students' ability to respond. Teachers were asked to use a 4-point scale (*strongly disagree*, *disagree*, *agree*, or *strongly agree*). Results were combined in the summary presented in Table 51. The majority of teachers 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.

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

Table 51. Teacher Perceptions of Student Experience With Testlets

	Strongly disagree		Disagree		Agree		Strongly agree		Agree or strongly agree	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
The student responded to items to the best of their knowledge and ability.	891	3.5	1,923	7.7	13,327	53.0	8,986	35.8	22,313	88.8
The student was able to respond regardless of disability, behavior, or health concerns.	1,662	6.6	2,251	9.0	13,509	53.7	7,718	30.7	21,227	84.4
The student had access to all supports necessary to participate.	614	2.4	764	3.0	13,022	51.8	10,753	42.8	23,775	94.6

IX.2.B. Test Administration Observations

Test administration observations were conducted in multiple states during 2016–2017 to further understand student response processes. Students' typical test administration process with their actual test administrator was observed. Administrations were observed for the full range of students eligible for DLM assessments (i.e., students with the most significant cognitive disabilities). Test administration observations were collected by DLM project staff, as well as 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.

Test administration observations were collected in five states during the 2016–2017 academic year. Table 52 shows the number of observations collected by state.

Table 52. Teacher Observations by State ($N = 172$)

State	*n*	%
Kansas	4	2.3
Missouri	52	30.2
New York	18	10.5
North Dakota	11	6.4
West Virginia	87	50.6

Of the 172 test administration observations collected, 102 (59.3%) were of computer-delivered assessments and 70 (40.7%) were of teacher-administered testlets. Of these 172 observations, 88 (51.2%) were of ELA reading testlets, 15 (8.7%) were of ELA writing testlets, and 80 (46.5%) were of mathematics testlets; nine observations were made for multiple subjects within a single observation.

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 53; behaviors were identified as supporting, neutral, or nonsupporting. For example, clarifying directions (50% 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) clearly indicates that the teacher directly influenced the student’s answer choice.

Table 53. Test Administrator Actions During Computer Administered Testlets (*n* = 102)

Evidence	Action	*n*	%
Supporting	Read one or more screens aloud to the student	51	50.0
	Clarified directions or expectations for the student	51	50.0
	Navigated one or more screens for the student	30	29.4
	Repeated question(s) before student responded	40	39.2
Neutral	Asked the student to clarify or confirm one or more responses	14	13.7
	Repeated question(s) after student responded (i.e., gave a second trial at the same item)	5	4.9
	Allowed student to take a break during the testlet	17	16.7
	Used verbal prompts to direct the student's attention or engagement (e.g., "look at this")	44	43.1
	Used pointing or gestures to direct student attention or engagement	29	28.4
	Used materials or manipulatives during the administration process	12	11.8
Nonsupporting	Reduced the number of answer choices available to the student	1	1.0
	Physically guided the student's hand to an answer choice	2	2.0

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 29% 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 from watching.

A related assumption is that students are able to interact with the system as intended. Evidence for this assumption was gathered by observing students taking computer-delivered testlets, as shown in Table 54. Independent response selection was observed in 76% 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 6% 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 54. Student Actions During Computer Administered Testlets (*n* = 102)

Action	*n*	%
Selected answers independently	77	75.5
Navigated the screens independently	62	60.8
Selected answers with verbal prompts	32	31.4
Navigated the screens with verbal prompts	25	24.5
Navigated screens after test administrator pointed or gestured	15	14.7
Independently revisited a question after answering it	9	8.8
Used materials outside of KITE Client to indicate responses to testlet items	6	5.9
Skipped one or more items	5	4.9
Revisited one or more questions after verbal prompt(s)	3	2.9

Note. Respondents could select multiple responses to this question.

Another assumption in the validity argument is that students are able to respond to tasks irrespective of sensory, mobility, health, communication, or behavioral constraints. This assumption was evaluated by having observers note whether there was difficulty with accessibility supports (including lack of appropriate available supports) during observations of teacher-administered testlets. Of the 70 observations of teacher-administered testlets, observers noted difficulty in five cases (7.1%). For computer-delivered testlets, evidence to evaluate this assumption was collected by noting students' indicating responses to items using multiple response modes such as eye gaze (2.0%) and using manipulatives or materials outside of KITE (5.9%). Additional evidence for this assumption was gathered by observing whether students were able to complete testlets. Of the 172 test administration observations collected, students completed the testlet in 162 cases (94.2%).

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 55 summarizes students' response modes for teacher-administered testlets. The most frequently observed behavior was the student verbally indicated response to test administrator who selected answers.

Table 55. Primary Response Mode for Teacher-Administered Testlets (*n* = 70)

Response mode	***n***	**%**
Verbally indicated response to test administrator who selected answers	29	41.4
Gestured to indicate response to test administrator who selected answers	27	38.6
Used computer/device to respond independently	17	24.3
Eye-gaze system indication to test administrator who selected answers	5	7.1
Used switch system to respond independently	0	0.0
No response	8	11.4

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 & 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 25 of 102 (25%) observations of computer-delivered testlets, the test administrator entered responses on the student's behalf. In 20 (80.0%) of those cases, observers indicated that the entered response matched the student's response, while three observers could not tell, and two left the item blank. This evidence supports the assumption that test administrators entered student responses with fidelity.

IX.2.C. Interrater Agreement of Writing Product 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 products. Evaluation of students' writing products does not use a high-inference process common in large-scale assessment, such as applying analytic or holistic rubrics. Instead, writing products 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 product. Only test administrators rate writing products, 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 products. For a complete description of writing-testlet design and scoring, including example items, see Chapter III of this manual.

The number of items that evaluated the writing product per grade-level testlet is summarized in Table 56. Testlets included one to six items evaluating the product, 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 56. Number of Items That Evaluate the Writing Product per Testlet, by Grade

	Number of items that evaluated writing product	
Grade	**Emergent product**	**Conventional product**
3	*	2
4	1	5
5	*	1
6**	*	3
7	1	4
8	*	4
9**	3	6
10**	3	6
11**	2	6
12**	2	6

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

IX.2.C.i. Recruitment

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

Products

During the spring 2017 administration, state partners were asked to recruit district coordinators to submit 10 samples of student writing products. Samples requested included papers that students used during testlet administration, copies of student writing products, or printed photographs of student writing products. To allow the product to be matched with test administrator response data from the spring 2017 administration, each product was submitted with a cover sheet that indicated the state, district, school, teacher, and student identifier.

A total of 177 student writing products were submitted from districts in six states. In several grades, the emergent writing testlet does not include any tasks that evaluate the writing product (see Table 56 above); therefore, products submitted for these grades were not included in the interrater reliability analysis (e.g., grade 3 emergent writing products). Additionally, writing products 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 131 writing products to raters for evaluation of interrater agreement.

Raters

Beginning in September 2017, state partners recruited teachers to participate in the rating of student writing products. Recruited teachers were required to have experience administering and rating DLM writing testlets to ensure they had already completed training and were familiar with how to score the writing products. A total of 150 teachers from across nine consortium states volunteered via an online Qualtrics survey. Final raters were selected from the available pool of volunteers, balancing the distribution of expertise, years of experience, demographic groups, and states. To support the study's design, including rater overlap and size of assignment, and with a contingency for rater attrition, 53 raters were selected.

Raters had a range of teaching experience, as indicated in Table 57. Most had taught ELA and/or students with the most significant cognitive disabilities for 11 or more years. Furthermore, four raters (7.5%) reported prior experience as DLM item writers, and nine (17.0%) reported experience as DLM external reviewers. Teachers were also asked to indicate how many years they had administered DLM writing testlets; two people (3.8%) reported 1 year of experience, eight (15.1%) had 2 years of experience, and 43 (8.1%) had 3 or more years of experience.

Table 57. Raters' Teaching Experience (*N* = 53)

	1–5 years		**6–10 years**		**≥11 years**	
Teaching experience	***n***	**%**	***n***	**%**	***n***	**%**
English language arts	14	26.4	12	22.6	27	50.9
Students with significant cognitive disabilities	11	20.8	14	26.4	28	52.8

Demographic information was collected as part of the volunteer survey administered in Qualtrics and is summarized in Table 58. Selected raters were mostly female (89%), white (89%), and non-Hispanic/Latino (94%), 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 58. Raters' Demographic Information (*N* = 53)

Subgroup	*n*	%
Gender		
Female	47	88.7
Male	6	11.3
Race		
White	47	88.7
African American	4	7.5
Asian	1	1.9
American Indian/Alaskan Native	1	1.9
Hispanic ethnicity		
No	50	94.3
Yes	2	3.8
Missing	1	1.9
Teaching setting		
Urban	20	37.7
Suburban	17	32.1
Rural	16	30.2

IX.2.C.ii. Product Ratings

Ratings were completed independently and asynchronously. Raters were instructed how to access the products and the Qualtrics survey where they entered their ratings. After completing a security agreement, raters received de-identified student writing products via a secure site that allowed separate assignments for each rater. The site also included 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 product. Raters submitted all ratings online and were given two-and-half weeks to complete all assigned ratings.

Writing products were assigned to raters in batches of five, using a partially crossed matrix design to assign each product a pair of raters. Some raters who completed all ratings for the first batch of five writing products were assigned a second batch to help address potential attrition on other batches. A total of 34 teachers rated at least one writing product, reflecting an attrition rate of 35.8%. Teachers rated between one and 10 writing products; however, 20 writing products were unrated by the close of the rating period. The unrated products were spread randomly across all grades. For the remaining 111 rated writing products, 53 were rated by one rater, 56 were rated by two raters, and two were rated by three raters. Table 59 summarizes the number of rated products per grade.

Table 59. Student Writing Products With Ratings, by Grade (*N* = 111)

	Number of writing products		Total number of
Grade	Emergent	Conventional	products
3	*	8	8
4	10	13	23
5	*	8	8
6	*	5	5
7	7	10	17
8	*	10	10
9	2	9	11
10	1	6	7
11	9	9	18
12	2	2	4
Total	31	80	111

Note. *The emergent testlet at this grade included only tasks evaluating the writing process, with no evaluation of writing product.

Product ratings submitted in Qualtrics were combined with the original student data from spring 2017, when the writing product was rated by the student's teacher, resulting in two to four ratings for each of the 111 student writing products.

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 (see Chapter III of this manual for more information). The script applied nested scoring rules (in instances where selection of the option reflecting the highest-level skill also indicates student also did 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").

IX.2.C.iii. Interrater Reliability

Because each product 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

$$k = \frac{\overline{P - \overline{P}_e}}{1 - \overline{P}_e},$$

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 both 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 product 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 60. To summarize global agreement across all student writing products, teachers' original ratings (from spring 2017 operational administration) were compared against either the one additional rating or one randomly selected rating from the additional raters (when more than one rating was collected). Results are also provided separately for emergent and conventional testlets and are summarized within rating-specific groups to indicate rater agreement for writing products with two ratings and three ratings. Because of decreasing sample sizes, these results are not disaggregated by group.

Agreement for the ICC tended to fall in the *excellent* range (≥ .75), and Fleiss's kappa tended to fall in the *good* range (.60–.74), as identified by Cichetti (1994). Moreover, findings suggest that conventional tasks may be rated more consistently. However, because more products were available for the conventional testlets and the testlets tended to have more tasks available to evaluate the products, this finding is not surprising.

Table 60. Interrater Agreement for Writing Products (*N* = 111)

Data	Group	*n*	ICC	ICC lower bound	ICC upper bound	Fleiss's kappa
Teacher + 1 random rater	Overall	111	.80	.78	.82	.67
Teacher + 1 random rater	Emergent	31	.63	.53	.72	.47
Teacher + 1 random rater	Conventional	80	.81	.79	.83	.69
Products with only two ratings	Overall	53	.78	.75	.81	.64
Products with three ratings	Overall	56	.88	.87	.89	.71

Note. Because only two writing products had four raters, a category for four raters was not included. ICC = intraclass correlation.

The results presented here reflect a first analysis of interrater agreement for teacher-administered writing testlets. Teacher-administered testlets measuring reading and mathematics were not included in the study. Also, although student writing products were evaluated, the student writing process was not. Additional data collection related to teacher fidelity, including fidelity in teacher-administered testlets in each content area, is provided in the Test Administration Observations section of this chapter.

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

A discussion of next steps is included in Chapter XI of this manual.

IX.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.

IX.3.A. Evaluation of Item Level Bias

Differential item functioning (DIF) addresses the broad problem created when some test items are "asked in such a way that certain groups of examinees who are knowledgeable about the intended concepts are prevented from showing what they know" (Camilli & Shepard, 1994, p. 1). 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, thereby contributing to an overall argument for validity and fairness.

IX.3.A.i. Method

DIF analyses for 2017 followed the same procedure used in the previous 2 years, including data from 2014–2015 through 2016–2017 to flag items for evidence of DIF. As additional data are collected in subsequent operational years, the scope of DIF analyses will be expanded to include additional items, subgroups, and approaches to detecting DIF.

Items were selected for inclusion in the DIF analyses based on minimum sample size requirements for the two gender subgroups: male and female. Within the DLM population, the number of female students responding to items is smaller than the number of male students by a ratio of approximately 1:2; therefore, a threshold for item inclusion was retained from the previous 2 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 are scored at the option level rather than the item level, and they include nonindependent response options (see Chapter III in this manual for more information).

Consistent with 2016, additional criteria were included to prevent estimation errors. Items with an overall *p* value (or proportion correct) greater than .95 were removed from the analyses. Items for which the *p* value for one gender group was greater than .97 were also removed from the analyses.

Using the above criteria for inclusion, 3,477 (54%) items on multi-EE testlets were selected. The number of items evaluated by grade level and content area ranged from 131 items in grade 8 ELA to 265 items in grade 9 mathematics. Item sample sizes ranged from 224 to 13,483.

For each item, logistic regression was used to predict the probability of a correct response, given group membership and total linkage levels mastered by the student in the content area. The logistic regression equation for each item included a matching variable comprised of the student's total linkage levels mastered in the content area of the item and a group membership variable, with females coded 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 nonuniform DIF indicates that the item functions differently because of the interaction between total linkage levels mastered and gender. When nonuniform DIF is present, the gender group with the highest probability of a correct response to the item differs along the range of total linkage levels mastered, in which 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:

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

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

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

where π_i is the probability of a correct response to the item for group i, X is the matching criterion, α is the intercept, β is the slope, γ_I is the group-specific parameter, and $\delta_I X$ is the interaction term.

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_2 to M_1 or M_0, 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 0.13 and 0.26; values less than 0.13 have a negligible effect, values between 0.13 and 0.26 have a moderate effect, and values of 0.26 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 classified with an A, moderate effect with a B, and large effect with a C for both methods.

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. As with equivalent sample size groups, the authors found that power increased and Type I error rates decreased as sample size increased for the unequal sample size groups. Decreased power to detect DIF items was observed when sample size discrepancies reached a ratio of 1:4.

IX.3.A.ii. Results

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

A total of 487 items were flagged for evidence of uniform DIF when comparing M_1 to M_2. Table 61 summarizes the total number of items flagged for evidence of uniform DIF by content area and grade for each model. The percentage of items flagged for uniform DIF for each grade and content area ranged from 7.6% to 18.4%.

Table 61. Items Flagged for Evidence of Uniform Differential Item Functioning

Content area	Grade	Items flagged (*n*)	Total items (*N*)	Items flagged (%)	Items with moderate or large effect size (*n*)
English language arts	3	25	169	14.8	0
	4	24	187	12.8	0
	5	19	188	10.1	0
	6	25	154	16.2	0
	7	17	133	12.8	0
	8	24	131	18.3	0
	9	24	145	16.6	0
	10	18	172	10.5	0
	11	30	163	18.4	0
Mathematics	3	31	185	16.8	0
	4	40	223	17.9	0
	5	38	234	16.2	0
	6	26	238	10.9	0
	7	39	225	17.3	0
	8	36	211	17.1	0
	9	27	265	10.2	0
	10	17	225	7.6	0
	11	27	229	11.8	0

Using the Zumbo and Thomas (1997) effect-size classification criteria, all 487 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 487 items were found to have a negligible effect-size change after the gender term was added to the regression equation.

IX.3.A.ii.b Combined Model

A total of 600 items were flagged for evidence of DIF when both the gender and interaction terms were included in the regression equation. Table 62 summarizes the number of items flagged by content area and grade. The percentage of items flagged for each grade and content area ranged from 8.9% to 24.6%.

Table 62. Items Flagged for Evidence of Differential Item Functioning for the Combined Model

Content area	**Grade**	**Items flagged (*n*)**	**Total items (*N*)**	**Items flagged (%)**	**Items with moderate or large effect size (*n*)**
English language arts	3	31	169	18.3	0
	4	27	187	14.4	0
	5	25	188	13.3	0
	6	33	154	21.4	1
	7	18	133	13.5	0
	8	26	131	19.8	0
	9	20	145	13.8	0
	10	18	172	10.5	1
	11	36	163	22.1	0
Mathematics	3	43	185	23.2	0
	4	46	223	20.6	0
	5	40	234	17.1	0
	6	36	238	15.1	1
	7	47	225	20.9	0
	8	52	211	24.6	0
	9	38	265	14.3	0
	10	20	225	8.9	0
	11	44	229	19.2	0

Using the Zumbo and Thomas (1997) effect-size classification criteria, one item had a large change in effect size, and the remaining 599 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, one item had a large change in effect size, two items had a moderate change in effect size, and the remaining 597 items were found to have a negligible change in effect size after adding the gender and interaction terms to the regression equation.

Information about the flagged items with a moderate or large change in effect size is summarized in Table 63 and Table 64 for ELA and mathematics, respectively. Two ELA items had a moderate change in effect-size values, as represented by a value of B. One mathematics item had a large change in effect-size value, as represented by a value of C. All three items favored the male group (as indicated by a negative γ value).

Table 63. English Language Arts Items Flagged for Differential Item Functioning With Moderate Effect Size

Grade	Item ID	EE	χ^2	*p* value	γ	$\delta_i X$	R^2	Z&T*	J&G*
6	37509	RL.6.3	11.96	<.01	-1.69	0.29	0.04	A	B
10	28628	RI.9-10.4	8.62	.01	-2.78	0.13	0.04	A	B

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

Table 64. Mathematics Item Flagged for Differential Item Functioning With Large Effect Size

Grade	Item ID	EE	χ^2	*p* value	γ	$\delta_i X$	R^2	Z&T*	J&G*
6	24139	6.EE.1-2	16.69	<.01	-0.37	0.12	0.95	C	C

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

A comparison of results from 2016 to 2017, after the collection of an additional year of data, indicates none of the ELA or mathematics items flagged in 2016 was flagged in 2017.

Appendix A includes plots labeled by the item ID, which 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.

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

The test development teams for each content area were provided with data files that listed all items flagged with a moderate or large effect size. 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 Chapter III of *2014–2015 Technical Manual – Year-End Model* [DLM Consortium, 2016b]) 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 and the engagement activity in mathematics, content 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 content 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. For items with a decision code of 1: Accept, no evidence could be found in any of the items indicating the content favored one gender group over the other.

IX.3.B. Internal Structure Across 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 V 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 V of this manual, for 1,190 (98.3%) linkage levels, masters had a greater than .5 chance of providing a correct response to items. Similarly, for 916 (75.7%) linkage levels, non-masters had a less than .5 chance of providing a correct response to items. This finding provides support for how well the linkage levels measured the construct and for the overall validity of inferences that can be made from mastery classifications for the linkage levels.

Chapter III of this manual included additional evidence of internal consistency in the form of standardized difference figures. Standardized difference values are calculated for operational and field test items to indicate how far from the linkage level mean each item’s p value falls.

Across all linkage levels, 3,906 (93%) of items fell within two standard deviations of the mean for the linkage level.

These sources of evidence indicate that overall, the linkage levels provide consistent measures of what students know and can do. 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.

IX.4. Evidence Based on Relation to Other Variables

According to *Standards for Educational and Psychological Testing*, "analyses of the relationship of test scores to variables external to the test provide another important source of validity evidence" (AERA et al., 2014, p. 16). Results from the assessment should be related to other external sources of evidence measuring the same construct.

IX.4.A. Teacher Ratings on First Contact Survey

One source of external evidence for DLM assessments comes from teacher ratings of students' academic knowledge, skills, and understanding via the First Contact survey. Before administering testlets, educators complete (or annually update) the First Contact survey, which is a survey of learner characteristics.[9] Because ratings on the First Contact survey are distinct from the DLM assessment (which uses only a subset of items to calculate the student complexity band for each subject), they can serve as one source of external evidence regarding the construct being measured. The First Contact survey includes academic skill items: nine in the reading section, seven in the writing section, and 13 in the mathematics section.

For each academic item on the First Contact survey, test development teams reviewed the learning map model to identify tested nodes that measured the same skill. Not all First Contact items directly corresponded to nodes in the map. Tested nodes were identified for two of the reading First Contact items, three of the writing First Contact items, and nine of the mathematics First Contact items. A summary of the First Contact academic items and the number of nodes identified in the learning map model is provided in Table 65. Appendix A includes the complete list of nodes identified for each First Contact item.

[9] More information on the First Contact survey, including calculation of complexity band, can be found in Chapter III of DLM Consortium (2016).

Table 65. First Contact Items With Nodes Identified in Learning Map Model

First Contact item	Number of tested nodes	Number of linkage levels measuring the nodes
Reading		
Recognizes single symbols presented visually or tactually	1	1
Identifies individual words without symbol support	1	11
Writing		
Writes sentences or complete ideas without copying, using spelling (with or without word prediction)	1	2
Writes words or simple phrases without copying, using spelling (with or without word prediction)	2	8
Writes words using letters to accurately reflect some of the sounds	3	9
Mathematics		
Creates or matches patterns of objects or images	2	6
Identifies simple shapes in two or three dimensions (e.g., square, circle, triangle, cube, sphere)	8	4
Sorts objects by common properties (e.g., color, size, shape)	1	17
Adds or subtracts by joining or separating groups of objects	2	10
Adds and/or subtracts using numerals	15	13
Forms groups of objects for multiplication or division	2	12
Multiplies and/or divides using numerals	19	9
Tells time using an analog or digital clock	4	5
Uses common measuring tools (e.g., ruler, measuring cup)	5	3

IX.4.A.i. Relationship Between Mastery and First Contact Ratings

For each tested node identified by the test development teams, all EEs and linkage levels measuring the node were identified. A dataset was created that included student mastery of the EE and linkage level measuring the node, as well as First Contact survey responses.[10] Reading and mathematics First Contact items asked teachers to use a 4-point scale to indicate how consistently students demonstrated each skill: *almost never* (0%–20% of the time), *occasionally* (21%–50% of the time), *frequently* (51%–80% of the time), or *consistently* (81%–100% of the time). For writing, teachers indicated the highest level that described a student's writing skill.

Tetrachoric correlations for writing and polychoric correlations for reading and mathematics were calculated to determine the relationship between the teacher's First Contact rating and the student's mastery of the linkage level measuring nodes associated with the First Contact items.

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

Correlations for First Contact items with linkage level mastery are summarized in Table 66. Across all three First Contact sections of academic items, most correlations (70-100%) differed significantly from 0. Writing First Contact items showed the strongest relationship with linkage level mastery; this result was likely influenced by the blueprint requirement that all students complete writing testlets and by the testlet design whereby multiple linkage levels are measured, ensuring large sample sizes at all levels.

Table 66. Correlations of First Contact Item Response to Linkage Level Mastery

First Contact section	Linkage levels (*n*)	*r*			*SE*			% significant
		Min	Max	Median	Min	Max	Median	
Reading	12	-.13	.62	.43	0.01	0.04	0.03	83
Writing	14	.47	.75	.60	0.01	0.03	0.02	100
Mathematics	71	-.24	.60	.18	0.01	0.12	0.04	69

[10] Students who demonstrated mastery via the two-down rule were not included. See Chapter V in this manual for a complete description of the scoring rules.

Mathematics First Contact items varied most in their relationship to linkage level mastery. Because mathematics nodes represent finer-grained skills, and test development teams identified more nodes in mathematics, more correlations were calculated (*n* = 71) than for reading (*n* = 12) and writing (*n* = 14). Mathematics results were also likely affected by sample size. As few as 250 student data points were available for some linkage levels, compared to at least 967 in reading and 2,116 in writing. The decreased sample size is likely attributable to fewer students testing at the Target and Successor linkage levels. Small sample size is associated with increased standard errors (Moinester & Gottfried, 2014). Furthermore, a negative relationship between mathematics First Contact rating and linkage level mastery was observed in six instances. An example is seen in the relationship between the Proximal Precursor level of the high school mathematics EE M.G-CO.1 and the First Contact item "Identifies simple shapes in 2 or 3 dimensions." The linkage level statement for this EE and level is "Recognize circle, perpendicular/parallel lines." Although the linkage level measures the node "Recognize circles," it also measures other nodes; this combination likely contributed to the negative relationship observed.

Overall, 93% (*n* = 90) of the correlations were positive and 75% (*n* = 73) were significantly different from 0, indicating generally positive associations between linkage level mastery and First Contact ratings. Results for all correlations are summarized in Figure 22.

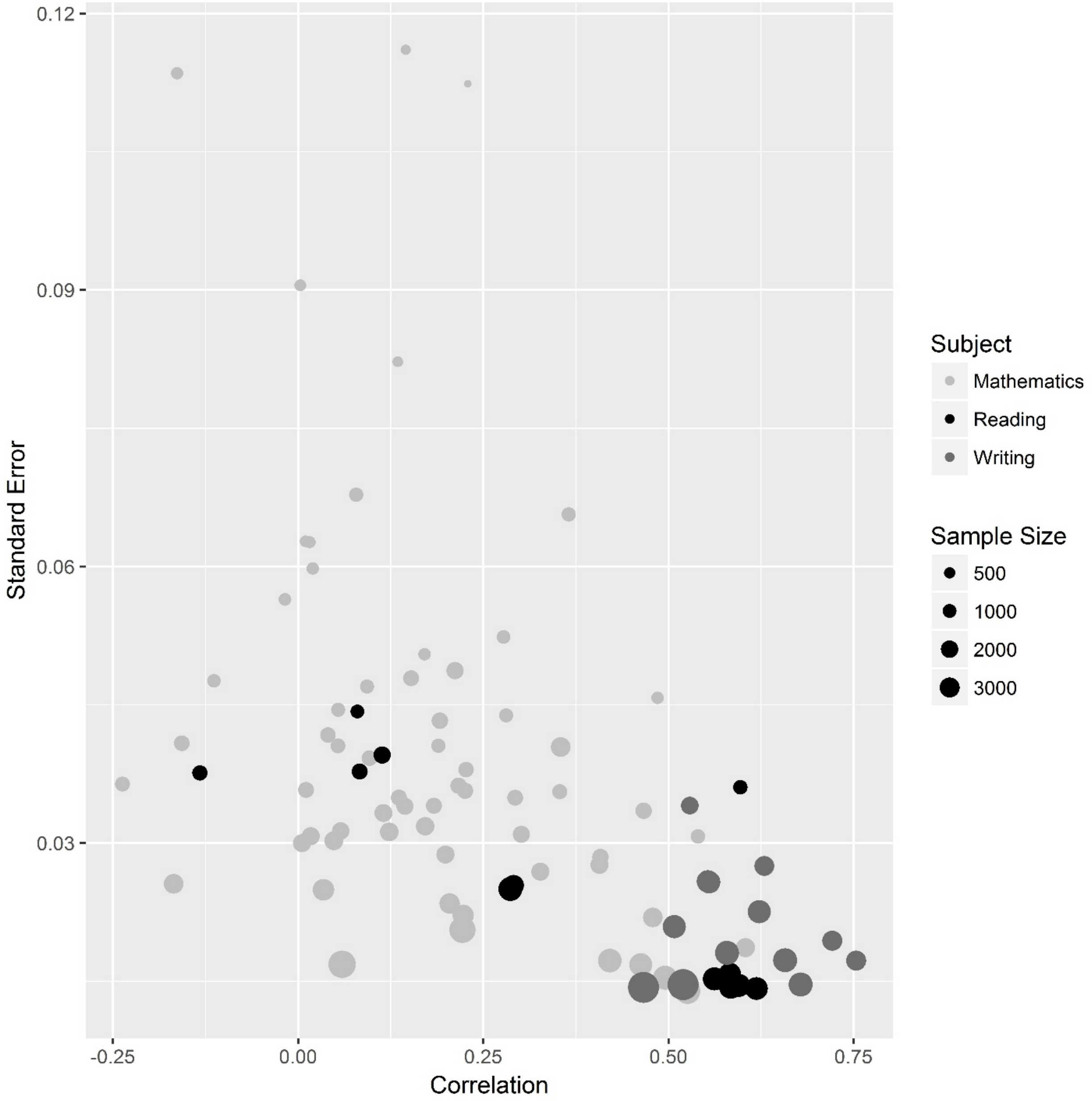


Figure 22. Relationship of First Contact responses to linkage level mastery for mathematics, reading, and writing.

This study provides preliminary evidence of the relationship between a portion of the ELA and mathematics blueprints with external variables, as indicated by teacher ratings on First Contact academic items. Because nodes can be measured by multiple EEs and linkage levels and because the grain size of linkage level statements also varies by level and grade, the relationship of linkage level mastery to First Contact rating was expected to be stronger in some areas than in others.

While the First Contact survey has separate sections for reading and writing, the EEs for these two areas sometimes overlap, particularly for foundational skills (e.g., the node "Can identify words that describe familiar persons, places, things, and events" is assessed in both reading and writing EEs, under different contexts). Therefore, the strength of the relationship between linkage level mastery and the First Contact item may also vary due to these differences.

Because this study examined only the subset of nodes and the corresponding EEs and linkage levels linked to items in the First Contact survey, evidence of relation to external variables is available for a portion of the blueprint in ELA and mathematics. An additional study is planned for spring 2018 to obtain evidence regarding the relationship between performance and external data for the complete blueprint. See Chapter XI of this manual for more information.

IX.5. Evidence Based on Consequences of Testing

Validity evidence must include the evaluation of the overall "soundness of these proposed interpretations for their intended uses" (AERA et al., 2014, p. 19). 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 2017 via teacher survey responses regarding teacher perceptions of assessment content. An additional study was conducted based on a score report tutorial to evaluate teachers' interpretation of report contents. Additional consequential evidence, including teacher focus groups on using score report contents in the subsequent academic year, will be collected in subsequent years.

IX.5.A. Teacher Perception of Assessment Contents

On the spring 2017 survey,[11] teachers were asked three questions about their perceptions of the assessment contents; Table 67 summarizes their responses. Teachers generally responded that content reflected high expectations for their students (82% agreed or strongly agreed), measured important academic skills (72% agreed or strongly agreed), and was similar to instructional activities used in the classroom (69% agreed or strongly agreed). While the majority of teachers agreed with these statements, approximately 20%–30% 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.

[11] Recruitment and sampling are described in Chapter IV of this manual.

Table 67. Teacher Perceptions of Assessment Content

	Strongly disagree		Disagree		Agree		Strongly agree	
Statement	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Content measured important academic skills and knowledge for this student.	2,592	10.3	4,470	17.7	14,755	58.5	3,420	13.6
Content reflected high expectations for this student.	1,313	5.2	3,241	12.9	15,201	60.4	5,392	21.4
Activities in testlets were similar to instructional activities used in the classroom.	2,275	9.1	5,565	22.2	13,973	55.7	3,278	13.1

IX.5.B. Score Report Interpretation Tutorial

To evaluate teacher interpretation and use of DLM score reports, a study was conducted based on an online tutorial created to support teacher interpretation of score report contents (Karvonen, Swinburne Romine, Clark, Brussow, & Kingston, 2017). The tutorial included an informed consent portion, followed by pre-test items, the training video, evaluation questions, and a post-test. The video incorporated concepts from the interpretation guide and addressed misconceptions identified in score report interpretation interviews with teachers. Researchers and DLM item writers familiar with DLM score reports wrote the pre- and post-test questions in the tutorial. Researchers wrote the evaluation questions, which included four Likert-scale items and two open-ended items.

Participating teachers reported a range of confidence in their ability to interpret and use DLM score reports before completing the tutorial; Table 68 summarizes the results. The greatest number of teachers reported being somewhat confident, while the fewest reported being not at all confident.

Table 68. Teacher Confidence in Ability to Interpret and Use DLM Score Reports Prior to Tutorial (*N* = 92)

Level of confidence	*n*	%
Very confident	11	12.0
Somewhat confident	33	35.9
Neither confident nor unconfident	25	27.2
Somewhat unconfident	13	14.1
Not at all confident	10	10.9

Following the training video, evaluation questions were presented to the participants; 55 participants responded to these questions. All respondents either strongly agreed (40%) or agreed (60%) that the tutorial covered important information. Most respondents strongly agreed (25%) or agreed (64%) that explanations provided in the tutorial were clear. Most respondents also reported that they felt prepared to explain DLM score report information to parents (87% agreed or strongly agreed) and to use DLM score reports to inform instruction (80% agreed or strongly agreed).

The evaluation included two open-ended items. The first asked teachers if they had remaining questions about interpreting DLM score reports. The second asked teachers to indicate additional resources that would help with interpretation and use of DLM score reports. Most teachers reported that they did not have remaining questions about the score reports. Additional feedback included requests for local training and supplemental materials to support instructional planning and decision-making. One participant requested a repository of training videos on different aspects of DLM, which is already available; this request indicates a need to better inform teachers about the resources available. Several participants also requested transcripts and hard copies of the sample reports used in the video, which will be made available online.

Post-test items were included following the evaluation section of the tutorial to prevent performance on the quiz from influencing participant evaluation of the tutorial. Forty-two participants took the post test. Of those, 18 participants (42.9%) passed (at least 80% accuracy) on their first try. If participants did not respond correctly to 80% of the items, the tutorial was presented again for retaking. Twenty-four participants (57.1%) completed the post-test a second time, two of whom reached the 80% threshold on their second attempt. Ten participants (23.8%) completed the tutorial a third time, but none achieved the passing threshold.

IX.6. Conclusion

This chapter presents additional studies as evidence to support the overall validity argument for the DLM Alternate Assessment System. The studies are organized into categories (content, response process, internal structure, 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 XI, references evidence presented through the technical manual, including Chapter IX, and expands the discussion of the overall validity argument. Chapter XI 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, 2016b) and the *2015–2016 Technical Manual – Year-End Model* (DLM Consortium, 2017c), in support of the assessment's validity argument.

X. TRAINING AND PROFESSIONAL DEVELOPMENT

Chapter X of the *2015–2016 Technical Manual – Year-End Model* (Dynamic Learning Maps® [DLM®] Consortium, 2017c) 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 2016–2017 instructional professional development. No changes were made to training in 2016–2017.

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

X.1. INSTRUCTIONAL PROFESSIONAL DEVELOPMENT

The DLM Professional Development System includes approximately 50 modules, including 20 focused on ELA instruction, 25 focused on mathematics instruction, and five others that address individual education programs, the DLM claims and conceptual areas, Universal Design for Learning, DLM Essential Elements (EEs), and the Common Core State Standards. The complete list of module titles is included in Table 70 below. The modules are available in two formats: self-directed and facilitated and are accessed at http://dlmpd.com. No new modules were added in 2016–2017.

The self-directed modules were designed to meet the needs of all educators, especially those in rural and remote areas, to offer educators with 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, as well as the opportunity to reflect upon and apply what they are learning. Each module ends with a posttest, and educators who achieve a score of 80% or higher on the posttest receive a certificate via email.

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

X.1.A. Professional Development Participation and Evaluation

As reported in Table 69, 102,917 modules were completed in the self-directed format from the fall of 2012, when the first module was launched, until September 30, 2017. This is an increase of 10,053 modules since September 30, 2016. Data are not available for the number of educators who have completed the modules in the facilitated format, but it is known that several states (e.g., Iowa, Missouri, West Virginia) use the facilitated modules extensively.

Table 69. Number of Self-Directed Modules Completed by Educators in Dynamic Learning Maps States and Other Localities, Through September 2017 (*N* = 102,917)

State	Self-directed modules completed (*n*)
Missouri	21,746
Kansas	19,904
New Jersey	9,407
Colorado	6,635
Wisconsin	5,414
Utah	2,635
Illinois	2,527
Oklahoma	1,863
Iowa	1,168
New Hampshire	708
Alaska	627
New York	630
North Dakota	449
West Virginia	166
Maryland	85
Non-DLM states and other locations	28,533

To evaluate educator perceptions of the utility and applicability of the modules, DLM staff asked educators to respond to a series of evaluation questions upon completion of each self-directed module. Through September 2017, on average, educators completed the evaluation questions 77% of the time. The responses are consistently positive, as illustrated in Table 70.

Table 70. Teacher Response Rates and Average Ratings on Self-Directed Module Evaluation Questions

Module name	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.
0: Who are Students with Significant Cognitive Disabilities?	12,583	.45	3.46	3.15	3.28	2.72
1: Common Core Overview	6,457	.39	3.18	2.97	3.12	2.63
2: Dynamic Learning Maps Essential Elements	10,230	.44	3.34	3.23	3.20	2.70
3: Universal Design for Learning	6,159	.44	3.36	3.27	3.28	2.71
4: Principles of Instruction in English Language Arts	5,471	.49	3.31	3.22	3.22	2.72
5: Standards of Mathematics Practice	7,776	.26	3.25	3.22	3.22	2.68
6: Counting and Cardinality	4,077	.52	3.38	3.31	3.31	2.72
7: IEPs Linked to the DLM Essential Elements	4,956	.45	3.28	3.21	3.22	2.67
8: Symbols	3,457	.29	3.37	3.30	3.32	2.71
9: Shared Reading	5,023	.56	3.45	3.37	3.32	2.75

Module name	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.
10: DLM Claims and Conceptual Areas	2,937	.71	3.27	3.13	3.14	2.63
11: Speaking and Listening	2,977	.52	3.34	3.26	3.26	2.71
12: Writing: Text Types and Purposes	3,057	.62	3.25	3.19	3.15	2.67
13: Writing: Production and Distribution	1,414	.92	3.26	3.20	3.19	2.69
14: Writing: Research and Range of Writing	1,778	.71	3.26	3.22	3.20	2.70
15: The Power of Ten-Frames	1,311	.92	3.26	3.24	3.20	2.66
16: Writing with Alternate Pencils	1,810	.91	3.39	3.34	3.32	2.66
17: DLM Core Vocabulary and Communication	1,960	.90	3.45	3.38	3.42	2.73
18: Unitizing	906	.88	3.19	3.12	3.13	2.60
19: Forms of Number	1,159	.85	3.16	3.13	3.11	2.57
20: Units and Operations	848	.89	3.14	3.10	3.07	2.56
21: Place Value	884	.87	3.14	3.10	3.06	2.51

Module name	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.
22: Fraction Concepts and Models Part I	744	.89	3.17	3.13	3.11	2.54
23: Fraction Concepts and Models Part II	634	.90	3.18	3.14	3.12	2.57
24: Composing, Decomposing, and Comparing Numbers	1,025	.84	3.27	3.23	3.22	2.55
25: Basic Geometric Shapes and Their Attributes	972	.86	3.23	3.19	3.16	2.54
26: Writing Information and Explanation Texts	622	.91	3.18	3.17	3.16	2.61
27: Calculating Accurately with Addition	831	.88	3.24	3.20	3.16	2.52
28: Measuring and Comparing Lengths	468	.88	3.26	3.22	3.19	2.53
29: Emergent Writing	1,504	.89	3.43	3.38	3.38	2.69
30: Predictable Chart Writing	597	.93	3.40	3.34	3.37	2.72
31: Calculating Accurately with Subtraction	492	.88	3.24	3.22	3.19	2.52

Module name	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.
32: Teaching Text Comprehension: Anchor-Read-Apply	808	.87	3.38	3.34	3.33	2.64
33: Generating Purposes for Reading	586	.88	3.30	3.27	3.27	2.62
34: Exponents and Probability	260	.87	3.16	3.16	3.13	2.49
35: Beginning Communicators	1,498	.88	3.49	3.36	3.40	2.68
36: Time and Money	430	.92	3.30	3.23	3.21	2.64
37: DR-TA and Other Text Comprehension Approaches	500	.88	3.33	3.31	3.29	2.64
38: Supporting Participation in Discussions	540	.83	3.38	3.33	3.29	2.63
39: Algebraic Thinking	613	.90	3.31	3.25	3.24	2.54
40: Composing and Decomposing Shapes and Areas	416	.88	3.32	3.27	3.26	2.53
41: Writing: Getting Started with Writing Arguments	190	.89	3.11	3.13	3.08	2.49
42: Calculating Accurately with Multiplication	368	.85	3.35	3.28	3.27	2.50

Module name	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.
43: Perimeter, Volume, and Mass	220	.89	3.12	3.09	3.07	2.43
44: Writing: Getting Started in Narrative Writing	174	.91	3.25	3.23	3.20	2.57
45: Patterns and Sequence	195	.89	3.16	3.10	3.08	2.42
46: Functions and Rates	123	.81	3.12	3.13	3.10	2.39
47: Calculating Accurately with Division	298	.85	3.37	3.34	3.31	2.52
48: Organizing and Using Data to Answer Questions	223	.81	3.43	3.39	3.35	2.54
49: Strategies and Formats for Presenting Ideas	327	.80	3.42	3.39	3.39	2.55
50: Properties of Lines and Angles	29	.72	3.48	3.43	3.38	2.43
Total	**102,917**					
Average		**.77**	**3.29**	**3.23**	**3.23**	**2.60**

Note. The first three questions used a 4-point scale. The final question had three response options: No, Maybe, and Yes. SWSCDs = students with significant cognitive disabilities.

As part of the spring 2017 teacher survey, teachers were asked to indicate how many professional development modules they had completed in the last two years. Results are summarized in Table 71 below. Most respondents indicated they had completed between one and five modules in the last two years.

Table 71. Number of Dynamic Learning Maps Professional Development Modules Completed in the Last Two Years (*N* = 36, 792)

Number of Modules	*n*	%
0	4,472	12.2
1–5	11,459	31.1
6–10	5,625	15.3
11–15	3,193	8.7
16–20	2,041	5.5
≥21	3,030	8.2
Missing	6,972	18.9

In addition to the modules, the DLM instructional professional development system includes a variety of other instructional 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 a virtual community of practice that is open to educators, related service providers, families, and others who are seeking support in teaching students with the most significant cognitive disabilities in achieving academic standards. The virtual community of practice allows registered users to create and join groups, ask and answer questions, and share instructional resources and materials. The DLM professional development team at the University of North Carolina at Chapel Hill continues to work to seed and support the development of the virtual community of practice and is working to identify a new format that may attract more active users

XI. CONCLUSION AND DISCUSSION

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

The DLM system completed its third operational administration year in 2016–2017. This technical manual update provides updated evidence from the 2016–2017 year intended to support the propositions and assumptions that undergird the assessment system as described at the onset of its design in the DLM theory of action. The contents of this manual address the information summarized in Table 72. Evidence summarized in the *2015–2016 Technical Manual – Year-End Model* (Dynamic Learning Maps Consortium, 2017c) builds on the original evidence included in the *2014–2015 Technical Manual – Year-End Model* (DLM Consortium, 2016b). Together, the three documents summarize the validity evidence collected to date.

Table 72. Review of Technical Manual Update Contents

Chapter(s)	Contents
I	Provides an overview of information updated for the 2016–2017 year
II	Not updated for 2016–2017
III, IV, X	Provides procedural evidence collected during 2016–2017 of test content development and administration, including field test information, teacher survey results, and professional development module use
V	Describes the statistical model used to produce results based on student responses, along with evidence of model fit
VI	Not updated for 2016–2017
VII, VIII	Describes results and analysis of the third operational administration's data, evaluating how students performed on the assessment, the distributions of those results, including aggregated and disaggregated results, and analysis of the internal consistency of student responses
IX	Provides additional studies from 2016–2017 focused on specific topics related to validity and in support of the score propositions and assessment purposes

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.

XI.1. Validity Evidence Summary

The accumulated evidence available by the end of the 2016–2017 year provides additional support for the validity argument. Each proposition is addressed by evidence in one or more of the categories of validity evidence, as summarized in Table 73. While many sources of evidence contribute to multiple propositions, Table 73 lists the primary associations. For example, Proposition 4 is indirectly supported by content-related evidence described for Propositions 1 through 3. Table 74 shows the titles and sections for the chapters cited in Table 73.

Table 73. Dynamic Learning Maps Alternate Assessment System Propositions and Sources of Updated Evidence for 2016–2017

	Sources of evidence*				
Proposition	**Test content**	**Response processes**	**Internal structure**	**Relations with other variables**	**Consequences of testing**
1. Scores represent what students know and can do.	1, 2, 3, 4, 5, 6, 8, 9, 10, 12	6, 13	3, 4, 7, 11, 14	15	8, 9, 16
2. Achievement level descriptors provide useful information about student achievement.	8, 9		11		8, 9, 16
3. Inferences regarding student achievement, progress, and growth can be drawn at the conceptual area level.	9, 12		11	12	9, 16
4. Assessment scores provide useful information to guide instructional decisions.					16

Note. *See Table 74 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 74. Evidence Sources Cited in Previous Table

Evidence no.	Chapter	Section
1	III	English Language Arts Writing Testlets
2	III	External Reviews
3	III	Operational Assessment Items for 2015–2016
4	III	Field Testing
5	IV	Administration Incidents
6	IV	User Experience with DLM System
7	V	All
8	VII	Student Performance
9	VII	Score Reports
10	VII	Quality Control Procedures for Data Files and Score Reports
11	VIII	All
12	IX	Evidence Based on Test Content
13	IX	Evidence Based on Response Process
14	IX	Evidence Based on Internal Structure
15	IX	Evidence Based on Relation to Other Variables
16	IX	Evidence Based on Consequences of Testing

XI.2. Continuous Improvement

XI.2.A. Operational Assessment

As noted previously in this manual, 2016–2017 was the third year the DLM Alternate Assessment System was operational. While the 2016–2017 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 2017–2018. This section describes significant changes from the second to third year of operational administration, as well as examples of improvements to be made during the 2017–2018 year.

Overall, there were no significant changes to the learning map models, item writing procedures, item flagging outcomes, test administration, or the modeling procedure used to calibrate and score assessments from the two previous years to 2016–2017.

Results from the 2015–2016 administration indicated the percentage of students classified to the At Target or Advanced performance levels decreased from 2014–2015 to 2015–2016 in some grades and subjects. Results from the 2016–2017 administration were compared to results from 2015–2016 to determine if a similar pattern was evident. However, results indicated patterns of performance remained consistent from 2015–2016 to 2016–2017 in both content areas. Results will be examined again following the 2017–2018 administration.

Based on an ongoing effort to improve KITE® system functionality, several changes are being implemented during 2017–2018. States will be able to set state-specific instructionally embedded testing windows. The spring 2018 administration also will expand availability of braille forms—which currently includes English Braille American Edition (EBAE)—by adding Unified English Braille (UEB). Educator Portal will also be enhanced to support creation and delivery of data files and score reports to maintain faster delivery timelines. This includes automated creation of all aggregated reports provided at the class, school, district, and state levels; delivery and 2-week review of General Research File in the interface; on-demand Special Circumstance supplemental files; system-generated exited student files; and, in the event of administration incidents, Incident Files indicating actual student impact, rather than students potentially impacted by the incident.

The validity evidence collected in 2016–2017 expands upon the data compiled in the first two operational years for each 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, relation to other variables, and consequences of testing. Specifically, analysis of blueprint coverage and 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. An analysis of the relationship of First Contact survey responses to linkage level mastery provided evidence based on relation to other variables. Teacher survey responses also provided evidence based on consequences of testing, as well as a score report interpretation tutorial. Studies planned for 2017–2018 to provide additional validity evidence are summarized in the following section.

XI.2.B. Future Research

The continuous improvement process also leads to future directions for research to inform and improve the DLM Alternate Assessment System in 2017–2018 and beyond. The manual identifies some areas for further investigation.

DLM staff members are planning several studies for spring 2018 to collect data from teachers in the DLM Consortium states. The consortium plans to form a set of score report interpretation focus groups to collect information about how teachers use the 2017 summative score reports to inform instruction in the subsequent academic year. DLM staff will conduct interviews with teachers of students with the most significant cognitive disabilities who are also English learners to determine how teachers identify students needing services and how to support those students during instruction. Teachers will also 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 2018 to obtain the second year of data for longitudinal survey items as further validity evidence.

Teachers will compile and rate student writing products to expand the collection and evaluation of interrater agreement of writing products. 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. 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. Studies on the comparability of results for students who use various combinations of accessibility supports are also dependent upon the availability of larger data sets. This line of research is expected to begin in 2018. Additional evaluation of interrater agreement on writing products will be considered.

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

XII. REFERENCES

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.

Arlot, C. (2010). "A survey of cross-validation procedures for model selection." *Statistics Surveys*. 4, 40-79.

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

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

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

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Routledge.

Dynamic Learning Maps Consortium. (2016a). *Test Administration Manual 2016–2017.* Lawrence, KS: University of Kansas.

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

Dynamic Learning Maps Consortium. (2017a). *Educator Portal User Guide*. Lawrence, KS: University of Kansas.

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

Dynamic Learning Maps Consortium. (2017c). *2015-2016 Technical Manual Update – Year-End Model*. Lawrence, KS: University of Kansas.

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.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: SAGE.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, New York: Springer-Verlag New York.

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

Karvonen, M., Swinburne Romine, R., Clark, A. K., Brussow, J., & Kingston, N. (2017, April). Promoting accurate score report interpretation and use for instructional planning. Paper presented at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

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

Lancaster, H. O., & Seneta, E. (2005). "Chi-square distribution." *Encyclopedia of Biostatistics*. New York, New York: John Wiley & Sons, LTD. doi: 10.1002/0470011815.b2a15018

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

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

Maydeu-Olivares, A. & Joe, H. (2006). "Limited information goodness-of-fit testing in multidimensional contingency tables." *Psychometrika*. 71 (713). doi: 10.1007/s11336-005-1295-9

Maydeu-Olivares, A. & Joe, H. (2014). "Assessing approximate fit in categorical data analysis." *Multivariate Behavioral Research*. 49(4), 305-28. doi: 10.1080/00273171.2014.911075

Neyman, J. & Pearson, E.S. "On the problem of the most efficient tests of statistical hypothesis." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*. 231, 289-337.

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

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

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

Tukey, J.W. (1958). "Bias and confidence in not-quite large samples." *Annals of Mathematical Statistics*, 29, 614-623.

Wells-Moreaux, S., Bechard, S. & Karvonen, M. (2016). *Accessibility Manual for the Dynamic Learning Maps® Alternate Assessment, 2016-2017*. Lawrence, KS: The University of Kansas Center for Educational Testing and Evaluation.

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