DYNAMIC™
LEARNING MAPS

Summary of Results from the 2014 and 2015 Field Test Administrations of the Dynamic Learning Maps® Alternate Assessment System

Technical Report #15-04

4/14/2016

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.

The present publication was developed under grant 84.373X100001 from the U.S. Department of Education, Office of Special Education Programs. The views expressed herein are solely those of the authors, and no official endorsement by the U.S. Department of Education should be inferred.

Contents

Introduction

The Dynamic Learning Maps® (DLM) Alternate Assessment Consortium conducted six field test administrations during 2014 and 2015. The field test assessments were available to educators and students in states belonging to the DLM® Consortium. The purpose of the field tests was to evaluate the quality of items and testlets, which measure Essential Elements (EEs), prior to making them operational. In addition, the field tests collected information on system accessibility and student classification to complexity bands.

Field testing of DLM content was necessary to support two different blueprint testing models: the integrated model (IM) and the year-end model (YE). To support the IM blueprint, pools of single-EE testlets must be available for administration to students during instructionally embedded and spring testing windows. To support the YE blueprint, multi-EE testlets must be available for the spring operational testing window. Educators in YE states also have the option of administering the single-EE testlets during the instructionally embedded testing windows, although these testlets do not count toward operational test results or state accountability.

The report that follows includes a summary of findings from the 2014 and 2015 field test events for English language arts (ELA) and mathematics. The first section of the report provides a high-level overview of the DLM structure. The second section summarizes field test implementation, including content, item analysis, and final decisions regarding operational use of items and testlets. The third section details student initialization into the assessment based on the First Contact survey completed by educators. The last section highlights system accessibility for the field test events.

Overview of Dynamic Learning Maps

Dynamic Learning Maps assessments are designed for students with the most significant cognitive disabilities. The DLM Alternate Assessment System is based on large, fine-grained learning map models. These learning map models are highly connected representations of how students acquire academic skills, as reflected in research literature. Nodes in the maps represent discrete knowledge, skills, and understandings in either ELA or mathematics, as well as important foundational skills that provide an understructure for the academic skills. The maps go beyond traditional learning progressions to include multiple and alternate pathways by which students may develop content knowledge. As of May 2016, there were 1,919 nodes in the ELA map, 2,399 nodes in the mathematics map, and 150 foundational nodes associated with both content-area maps.

The DLM EEs are specific statements of knowledge and skills linked to the grade-level expectations identified in college and career readiness standards. The purpose of the EEs is to build a bridge from those content standards to academic expectations for students with the most significant cognitive disabilities.

The EEs specify academic targets while the learning map model clarifies how students can reach those targets. For each EE, small collections of nodes are identified earlier in the map that represent critical junctures on the path toward the standard. These small collections of nodes are called linkage levels. The fourth level, called the Target, reflects the grade-level expectation in the EE. There are three levels below the Target (Initial Precursor, Distal Precursor, and Proximal Precursor) and one level beyond the Target (Successor).

Prior to taking a DLM assessment, a student’s teacher completes the First Contact survey, which is a survey of learner characteristics. The results of this survey inform the student’s assignment to one of the five linkage levels by determining the student’s complexity band for each content area. The complexity band a student is placed in is used to assign the student’s first testlet during spring testing.

Once the student’s linkage level is determined, DLM assessments are delivered as a series of testlets, each of which contains a nonscored engagement activity and three to eight items. Assessment items are aligned to nodes at one of the five linkage levels, as illustrated in Figure 1.

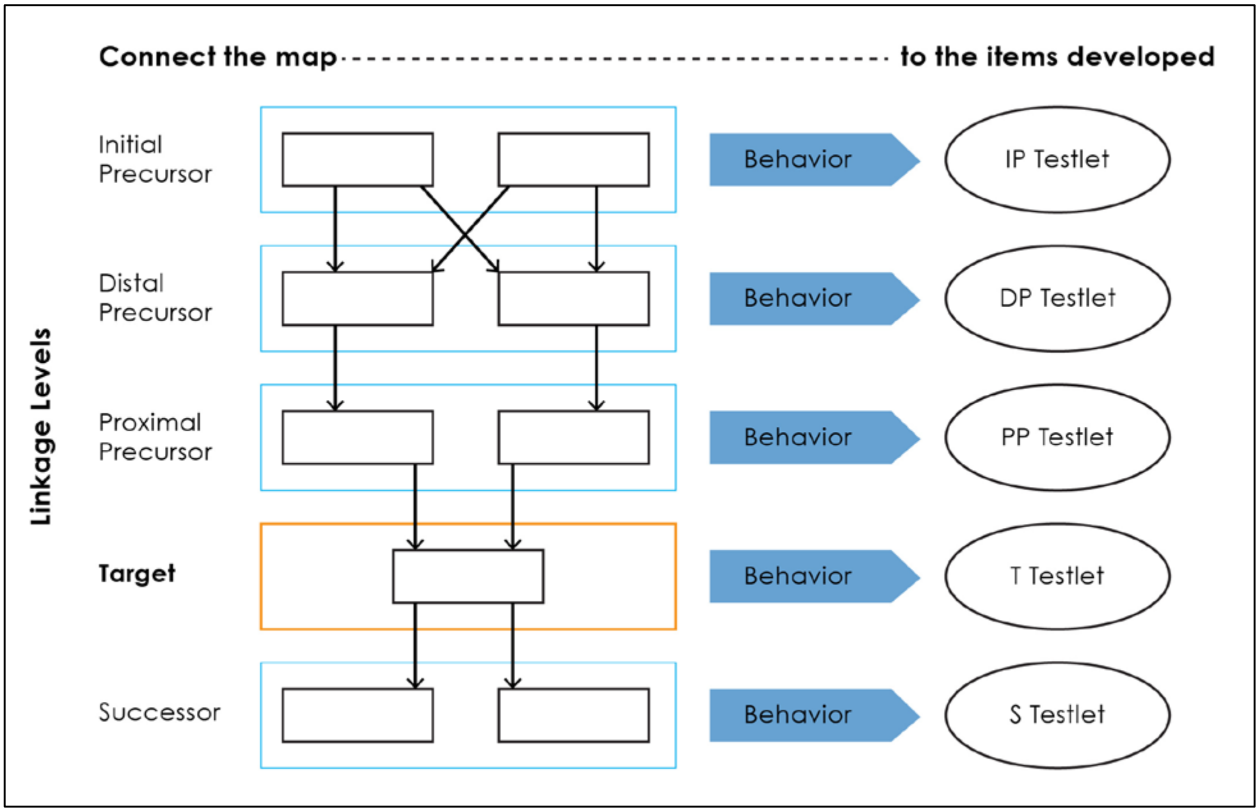

Figure 1. Relationship between testlet items and the five linkage levels in the learning map model.

All items are field-tested prior to being added to the operational pool of available testlets that cover the blueprints. The report that follows summarizes the results from the field tests conducted during 2014 and 2015.

Field Test Implementation

Purpose

The 2014 and 2015 DLM field tests were administered to evaluate the quality of items assessing EEs at each grade level and for each End-of-Instruction course in mathematics and ELA. In addition to evaluating item quality, the field tests also evaluated student initialization into the assessment system and student performance on items at the assessed linkage level.

Six field test windows were implemented during 2014 and 2015. Table 1 summarizes the dates of each field test window. The length of each field test window ranged from 10 business days to nine weeks.

Table 1

Field Test Windows

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

Field Test Design

For each of the six field test windows, the mathematics and ELA content teams selected testlets to cover the blueprints for grades 3–12, making testlets available at all five linkage levels for each EE. The initialization process determined the level(s) each student was assessed on during the field test.

Prior to being field-tested, all items went through an internal and external review process. For more information on this process, please see Clark, Karvonen, & Swinburne Romine (2014) and Clark, Swinburne Romine, Bell, & Karvonen (2015).

Field Tests 1 and 2. Field Tests 1 and 2 occurred before the blueprints were defined. As a result, all testlets included in Field Tests 1 and 2 were the first type of testlet designed: single-EE testlets. During Field Tests 1 and 2, two EEs were assessed at each grade and content area.

In order to evaluate student performance at more than one linkage level, Field Tests 1 and 2 used matrix sampling to combine three testlets within a single test form. Students received single-EE testlets at two or three adjacent linkage levels. The testlet(s) at the lowest linkage level were administered first, followed by testlet(s) at a higher linkage level.

Figure 2 presents a single example of the matrix sampling available at each complexity band. Each row identifies the testlet levels assigned to a student for that complexity band. As an example, a student classified into the Foundational band based on the educator's First Contact responses received three testlets covering the two tested EEs. The first testlet administered to the student was at the Initial Precursor level. The second testlet administered to the student was also at the Initial Precursor level but assessed the other EE. The third and final testlet also assessed the second EE but at the next highest linkage level, Distal Precursor. Additional combinations of testlets were available at each complexity band beyond the examples shown in the figure, always confined to two or three adjacent linkage levels.

Complexity Band	**Essential Element 1**					**Essential Element 2**				
Foundational	IP					IP	DP			
Band 1	IP	DP						PP		
Band 2		DP						PP	T	
Band 3				T	S					S

Figure 2. Matrix sampling examples for each complexity band in Field Tests 1 and 2. IP = Initial Precursor; DP = Distal Precursor; PP = Proximal Precursor; T = Target; S = Successor.

Using this matrix-sampling approach, 199 testlets were administered during Field Test 1. Table 2 indicates the number of Field Test 1 testlets by grade level and content area.

Table 2

Number of Testlets in Field Test 1

Grade	ELA	Math
3	15	29
4	7	13
5	11	14
6	12	11
7	11	13
8	10	13
9–10	13	14
11–12	13	N/A
Total	92	107

Note. For mathematics, high school testlets are administered in a single band of grades 9–12, as opposed to the two grade bands used in ELA. This difference is based on the high school grade banding for the EEs in the two content areas.

Field Test 2 covered the same two EEs that were tested in Field Test 1. In order to cover all linkage level combinations for the matrix-sampling approach in Field Test 2, some testlets from Field Test 1 were re-administered during Field Test 2. This also increased the sample size for those re-administered testlets. Table 3 provides the number of testlets administered in Field Test 2 by grade and content area. A total of 296 testlets were administered during Field Test 2. Of those testlets, 44 were administered in Field Test 1 as well.

Table 3

Number of Testlets in Field Test 2

	ELA		Math	
Grade	FT2 Only	Also in FT1	FT2 Only	Also in FT1
3	12	5	26	0
4	15	3	20	2
5	7	5	17	1
6	9	3	8	5
7	21	3	38	1
8	14	2	12	6
9–10	11	2	20	2
11–12	22	4	N/A	N/A
Total	111	27	141	17

Note. For mathematics, high school testlets are administered in a single band of grades 9–12, as opposed to the two grade bands used in ELA. This difference is based on the high school grade banding for the EEs in the two content areas.

Field Test 3. In contrast to Field Tests 1 and 2, Field Test 3 was designed to more closely reflect the operational assessments that would be available in the 2014–2015 year. During Field Test 3, students received three testlets, all at the same linkage level, based on initialization from responses to the First Contact survey. Each testlet was on a separate form and assessed a different EE out of the five available for each grade and content area.

During Field Test 3, a total of 738 single-EE testlets were administered. Table 4 provides a breakdown of the number of testlets administered at each grade and content area. No testlets were re-administered from Field Tests 1 or 2 during Field Test 3.

Table 4

Number of Testlets in Field Test 3

Grade	ELA	Math
3	67	43
4	47	77
5	30	63
6	42	70
7	36	62
8	28	42
9–10	40	50
11–12	41	N/A
Total	331	407

Note. For mathematics, high school testlets are administered in a single band of grades 9–12, as opposed to the two grade bands used in ELA. This difference is based on the high school grade banding for the EEs in the two content areas.

Phase A. In preparation for operational testing, the Phase A field test was structured similarly to Field Test 3. Students were assigned three or four testlets per content area at a single linkage level, based on their First Contact survey results.

Because blueprints were developed and approved by states in spring 2015, the Phase A window was the first field test to support two testing models with different blueprints. For the first time, multi-EE testlets were included to cover the YE blueprint. The number of multi-EE testlets administered by grade is included in Table 5. However, the multi-EE testlets available at the high school level are for individual grades rather than grade bands because of grade-specific blueprints for YE states. The YE blueprint does not include content requirements for grade 12.

Single-EE testlets were also field-tested during Phase A. This included a mixture of previously field tested testlets and new testlets. Some testlets from Field Test 3 were field-tested again after edits were made or to obtain a larger sample prior to evaluating the items for operational use. A total of 157 ELA single-EE testlets and 215 math single-EE testlets from Field Test 3 were retested in Phase A. Table 5 gives the total number of single-EE testlets administered during Phase A by grade and content area. The Phase A single-EE testlets covered a varied number of EEs at each grade level, approaching complete blueprint coverage. For ELA, between 8 and 13 EEs were assessed per grade. For mathematics, between 10 and 21 EEs were assessed per grade.

Table 5

Number of Phase A Testlets by Grade, Content Area, and Model

	Single-EE Testlets		Multi-EE Testlets	
Grade	ELA	Math	ELA	Math
3	52	53	19	28
4	64	64	16	26
5	36	62	15	23
6	29	70	14	18
7	27	61	15	22
8	20	43	17	16
9-10	27	111	36	35
11-12	28	N/A	16	15
Total	283	464	148	183

Note. For mathematics, single-EE high school testlets are administered in a single band of grades 9–11, as opposed to the two grade bands used in ELA. This difference is based on the high school grade banding for the EEs in the two content areas.

Phase B. Phase B was the first field test window to include complete coverage of all EEs and all linkage levels for both content areas and blueprint testing models. Another change introduced during the Phase B window was the shift to delivering all single-EE testlets through the Instructional Tools Interface (ITI) in Educator Portal. The ITI is where educators manage instructionally embedded assessments. In Phase B, educators logged into Educator Portal and created instructional plans for the EEs and linkage levels of their choosing. The system recommended a linkage level based on the student's First Contact survey results, but educators had the option to assess the student at a different linkage level if they wanted. Each instructional plan was associated with a single linkage level of a single EE. Educators could create separate instructional plans for different linkage levels of a single EE if they chose. Within ITI, a mixture of both operational and field test testlets were available. Multi-EE testlets were delivered during Phase B with the same method used in Phase A: an enrollment process automatically assigned up to four testlets, all at a single linkage level.

For both testing models, some testlets from Phase A were field-tested again. Reasons for retesting included the need for larger samples for item analysis and the need to have complete coverage at every EE and linkage level for instructionally embedded assessments. A total of 254 single-EE testlets and 97 multi-EE testlets were retested from Phase A. Table 6 includes a summary of the number of testlets available during Phase B by grade, content area, and model, including field test testlets specifically written for students who are blind or visually impaired (BVI). A total of 1,739 testlets were available during Phase B.

Table 6

Number of Phase B Testlets by Grade, Content Area, and Model

	Single-EE Testlets		Multi-EE Testlets	
Grade	ELA	Math	ELA	Math
3	89	83	59	63
4	46	66	43	57
5	71	65	41	51
6	31	81	36	47
7	29	89	43	50
8	33	74	38	38
9-10	25	120	74	93
11-12	29	N/A	31	44
Total	353	578	365	443

Note. For mathematics, high school single-EE testlets are administered in a single band of grades 9–11, as opposed to the two grade bands used in ELA. High school multi-EE testlets are by grade level rather than grade band. This difference is based on the high school grade banding for the EEs in the two content areas.

For both testing models, states provided their users with guidance on the number of field test testlets to complete. In most states, participation was voluntary; only four states (all IM) required participation during the Phase B window.

Phase C. During Phase C, single-EE testlets were administered through ITI, with a mix of field test and operational testlets available. Educators followed the same process from Phase B to select EEs and linkage levels on which to assess students on.

Multi-EE testlets were delivered by following the sequencing and adaptive algorithm rules planned for the spring operational testing window. Testlets were available to cover the complete blueprint, with students receiving between four and seven testlets.

The linkage level of the first multi-EE testlet assigned to a student was based on the results from the First Contact survey. After the first testlet, the linkage level of subsequent testlets was based on a student's performance on the previous testlet. Students were routed to the next highest or next lowest linkage level based on their proportion of correct responses for the EE in which they answered the lowest proportion of items correctly. If the lowest proportion of items answered correctly for that EE was above .79, the student advanced to the next highest linkage level. If the lowest proportion of items answered correctly for that EE was below .35, the student was assigned the next lowest linkage level. If the lowest proportion of items answered correctly for that EE fell between .35 and .79, the student received the same linkage level for the next testlet.

Table 7 provides the number of field test testlets available by grade, content area, and model. Because single-EE testlets must be available for every EE and linkage level to support educator choice of EEs as allowed by the IM blueprint, and because both instructionally embedded and spring windows require complete coverage, the total number of single-EE testlets needed for blueprint coverage is much higher than the number of multi-EE testlets required for coverage of every EE and linkage level.

Table 7

Count of Phase C Field Test Testlets by Grade, Content Area, and Model

	Single-EE Testlets		Multi-EE Testlets	
Grade	ELA	Math	ELA	Math
3	84	108	10	46
4	92	213	20	67
5	96	147	26	59
6	120	74	19	52
7	92	111	13	49
8	104	135	10	61
9-10	100	282	44	100
11-12	103	N/A	15	61
EOI	N/A	N/A	38	120
Total	791	1,071	195	615

Note. For mathematics, high school single-EE testlets are administered in a single band of grades 9–11, as opposed to the two grade bands used in ELA. High school multi-EE testlets are by grade level rather than grade band. Students were routed to the next highest or next lowest linkage level based on their proportion of correct responses for the EE in which they answered the lowest proportion of items correctly. End-of-Instruction (EOI) testlets are available to students rostered to any high school grade in which the course is available.

For both testing models, states provided their users with guidance on the number of field test testlets to complete during Phase C. In most states, participation was voluntary; only four states (all IM) required participation during the Phase C window.

Field Test Participation Counts

Students and educators were recruited for participation in each of the field test events by state and district education agencies within the DLM Consortium. In most states, participation was voluntary. Students and educators participated in anywhere from one to all six of the field test events during the 2014 and 2015 years. A summary of student, educator, district, and state participation during each of the field test windows is presented in Table 8. The counts include students with at least one testlet completed or in progress during the window dates.

Table 8

Field Test Participation

Group	Field Test 1	Field Test 2	Field Test 3	Phase A	Phase B	Phase C
Students	9,615	10,445	9,731	10,181	14,617	17,997
Educators	3,288	3,673	3,375	3,490	4,895	5,870
Districts	608	648	654	936	1,087	1,470
States	14	16	17	8	12	17

For Phase A through Phase C, states were asked to supply projected participation numbers to assist with planning for the field test windows. Across all three phases, participation was lower than states originally projected. The discrepancies between projected and actual participation were especially pronounced for states in the YE model. As a result, sample size per testlet was smaller than anticipated and fewer testlets met the sample size threshold for item analysis.

Figures 3–8 show the projected and actual numbers of students in each state participating in each phase, A through C, during the 2014–2015 academic year.

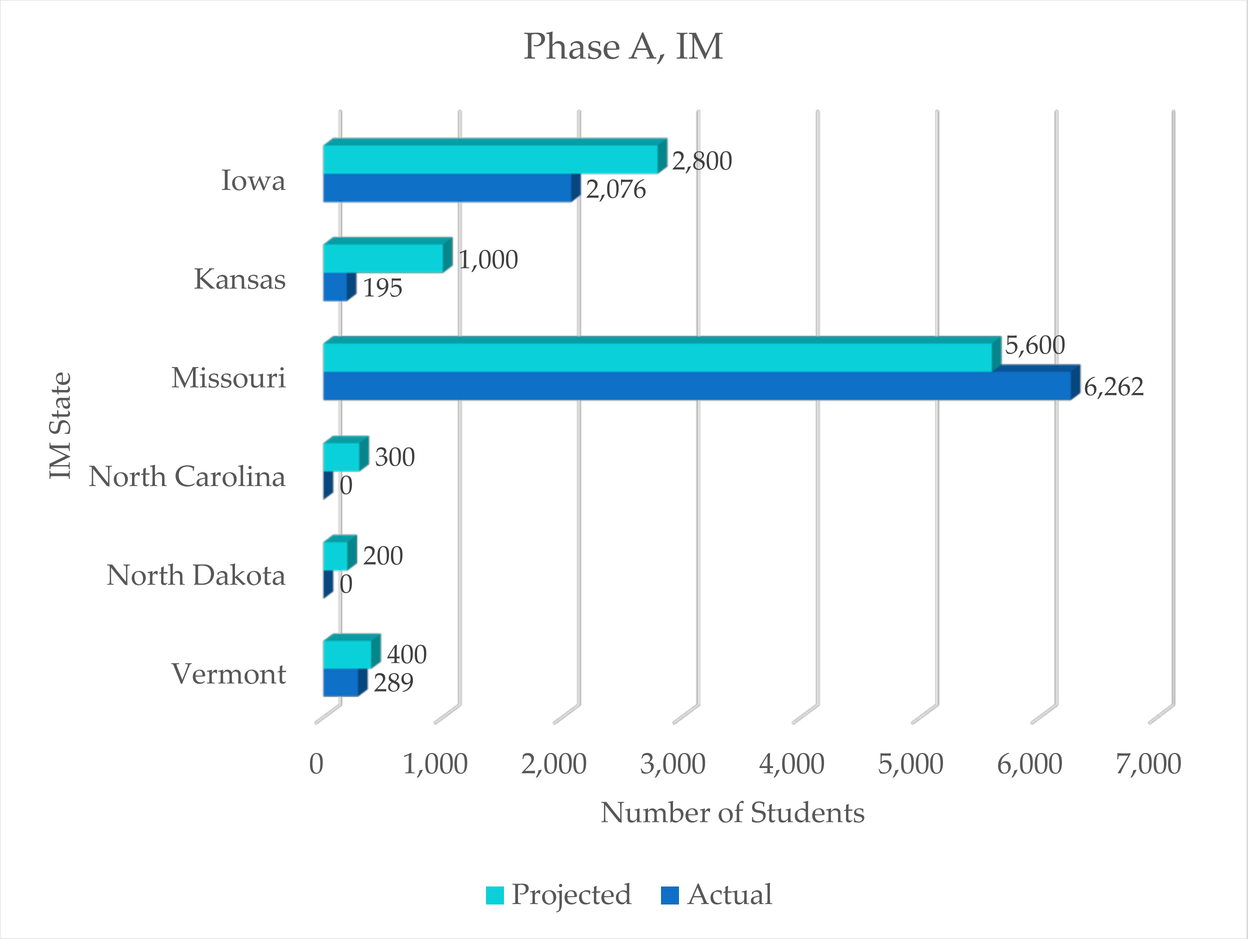

Figure 3. Phase A integrated model projected and actual participation.

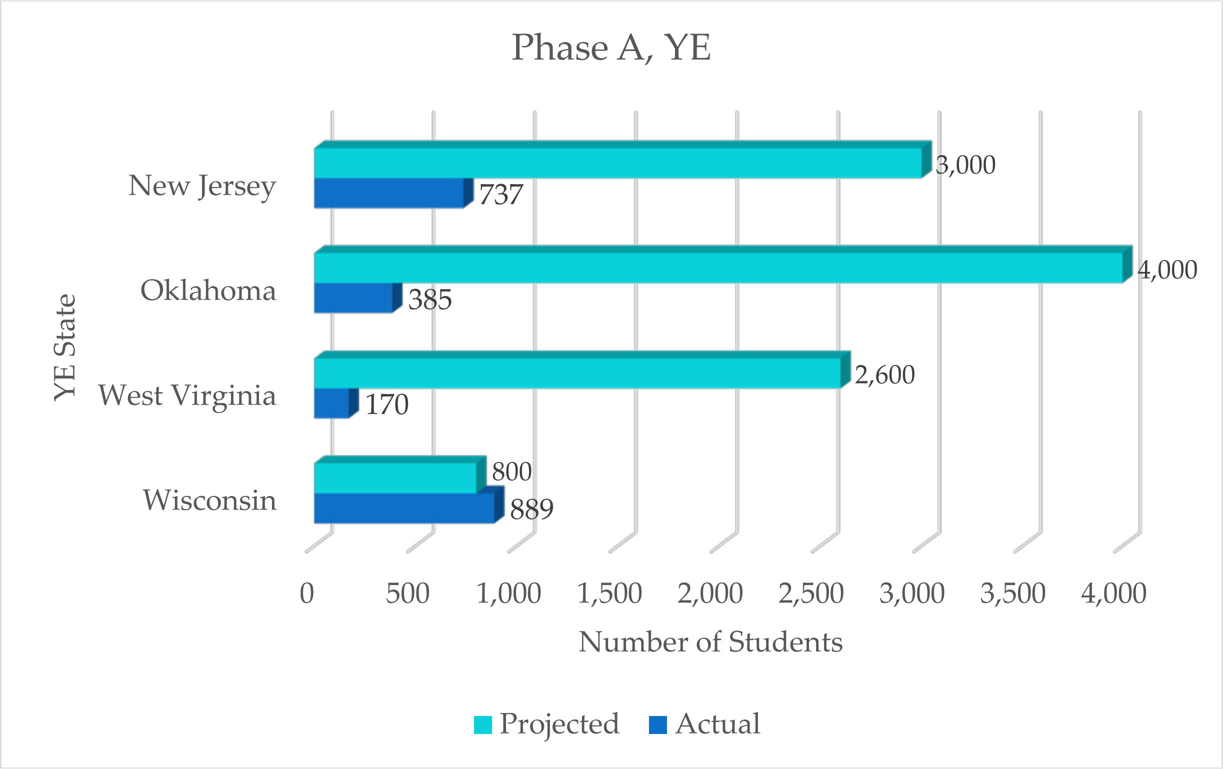

Figure 4. Phase A year-end model projected and actual participation. No projections or results for Utah.

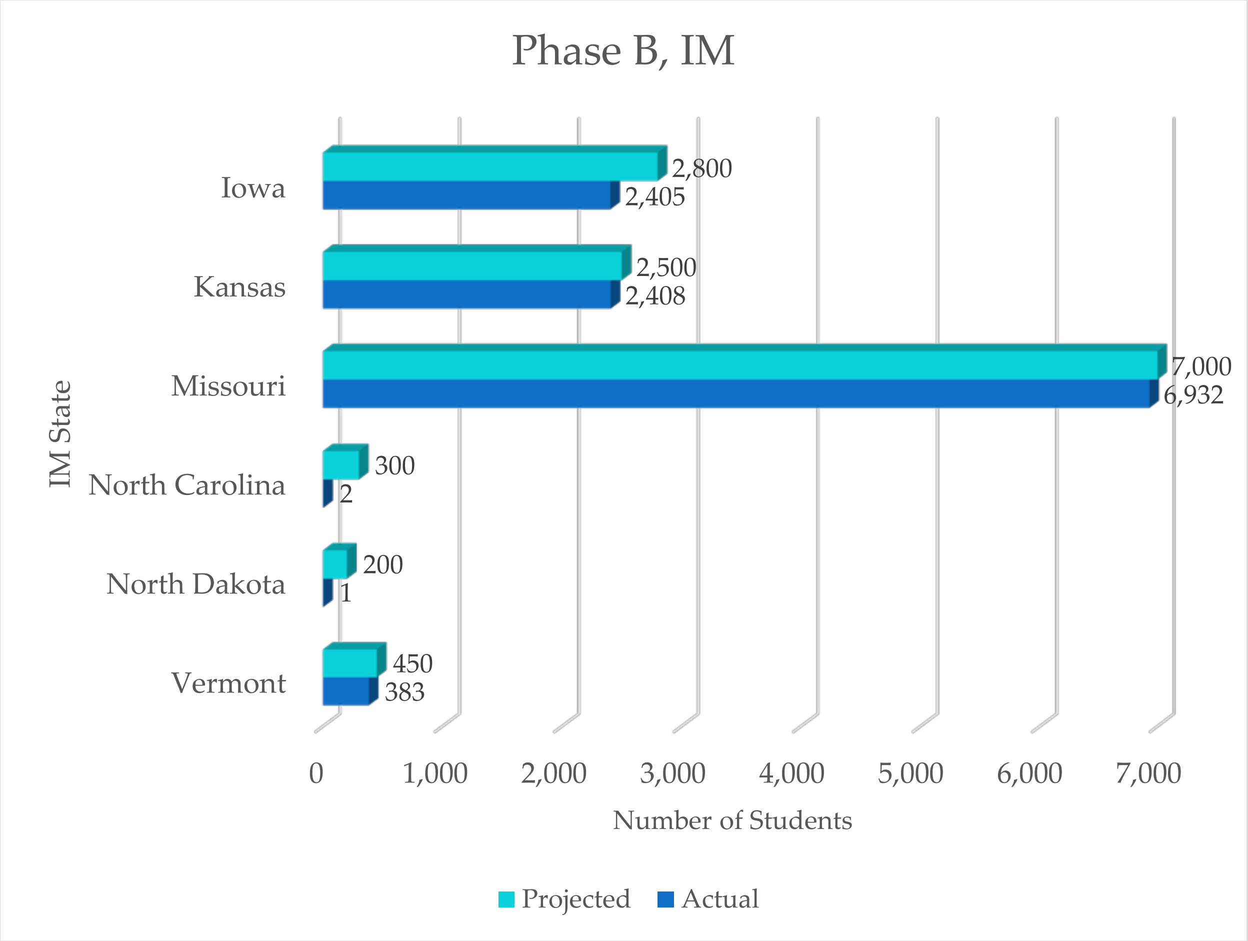

Figure 5. Phase B integrated model projected and actual participation.

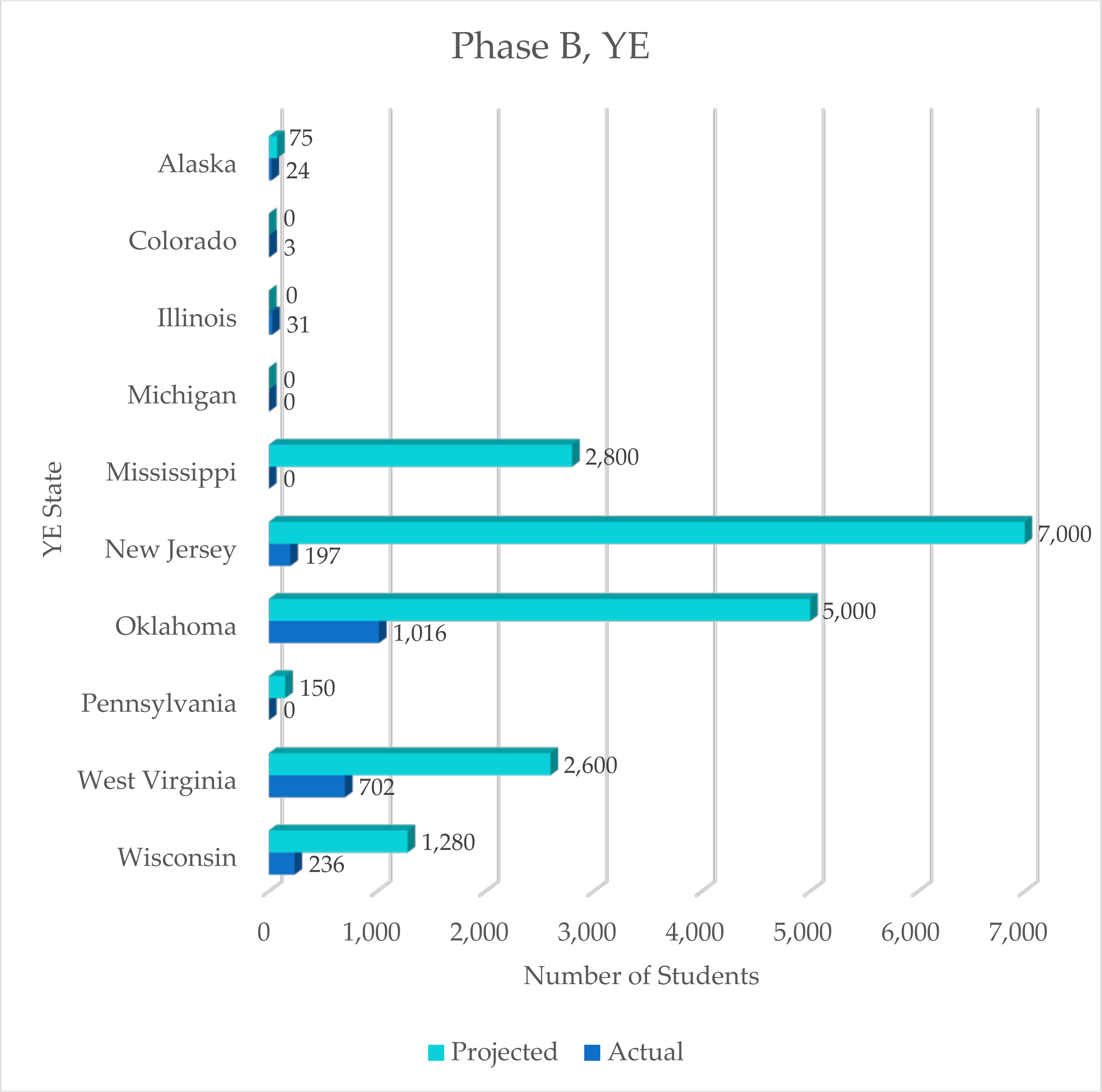

Figure 6. Phase B year-end model projected and actual participation. No projections or results for Utah.

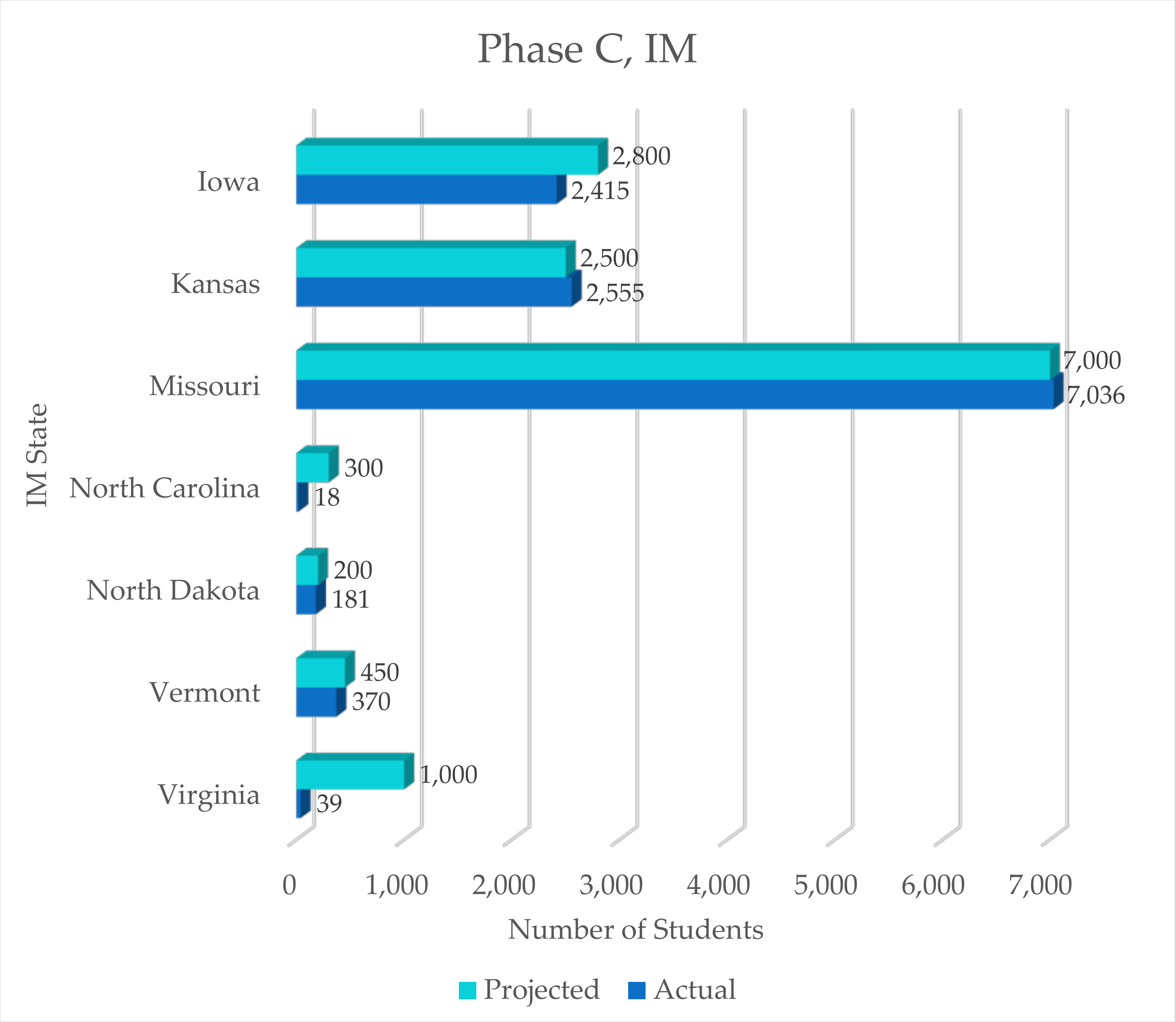

Figure 7. Phase C integrated model projected and actual participation.

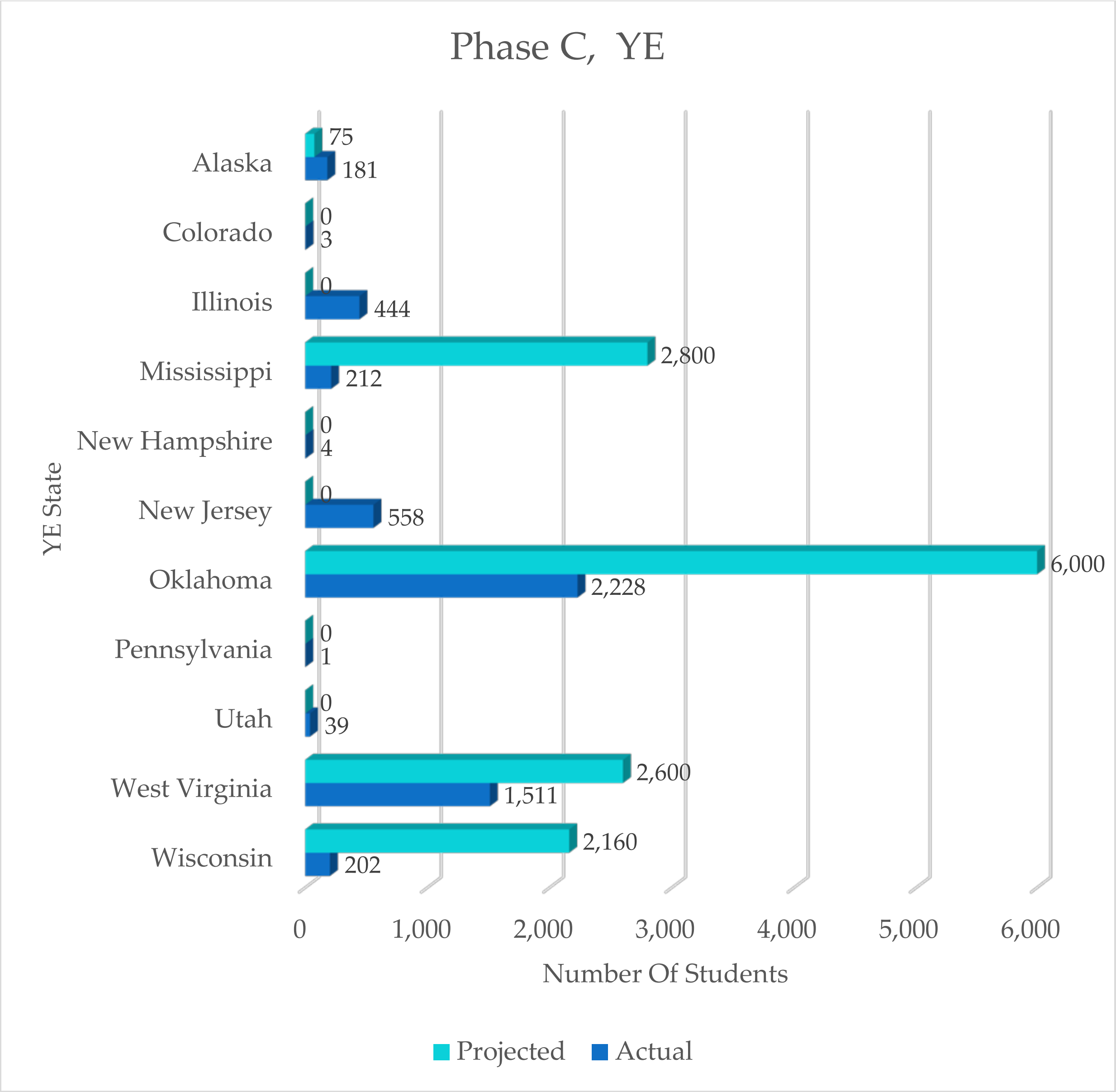

Figure 8. Phase C year-end model projected and actual participation.

Item Review

Following the conclusion of each field test window, student responses were analyzed to evaluate item quality. Items had to have at least 20 student responses in order to be evaluated for quality. Items and testlets with fewer than 20 student responses were scheduled for retesting in subsequent windows to collect additional student response data.

Due to time constraints between testing windows and the number of testlets in each field test, not every item and testlet could be closely reviewed by content teams following each field testing window. In order to focus the content teams' review of field test items, flagging criteria were developed to identify items in need of review.

Items were flagged for content team review if they met any of the following statistical criteria:

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

For Field Tests 1 and 2, in which matrix sampling was employed, analyses for item flagging were conducted within a single complexity band rather than across bands when examining item-level performance. As a result of this test development structure, an additional criterion was also applied for flagging items for review during Field Tests 1 and 2. Items were flagged in instances in which the item appeared more challenging as the complexity band increased, as indicated by the *p*-value at a lower complexity band being greater than the *p*-value at a higher complexity band.

Once the flagging process was complete, members of each content team met to review flagged items. However, DLM content teams did not make item- or testlet-level decisions based on statistical evidence alone. Rather, the content teams examined the statistical evidence along with the item content and its context in the testlet to determine if edits were necessary.

Figure 9 and Figure 10 provide histograms of item *p*-values from all field test windows for items with sample sizes of at least 20 for ELA and mathematics, respectively. In both content areas, most items fell above the *p*-value flagging threshold of 35%. In general, items field-tested in ELA appeared to be easier than items field-tested in mathematics. This could be due to differences in the difficulty of the nodes selected for assessment, or this could be the result of differences in the students' opportunity to learn the content being assessed.

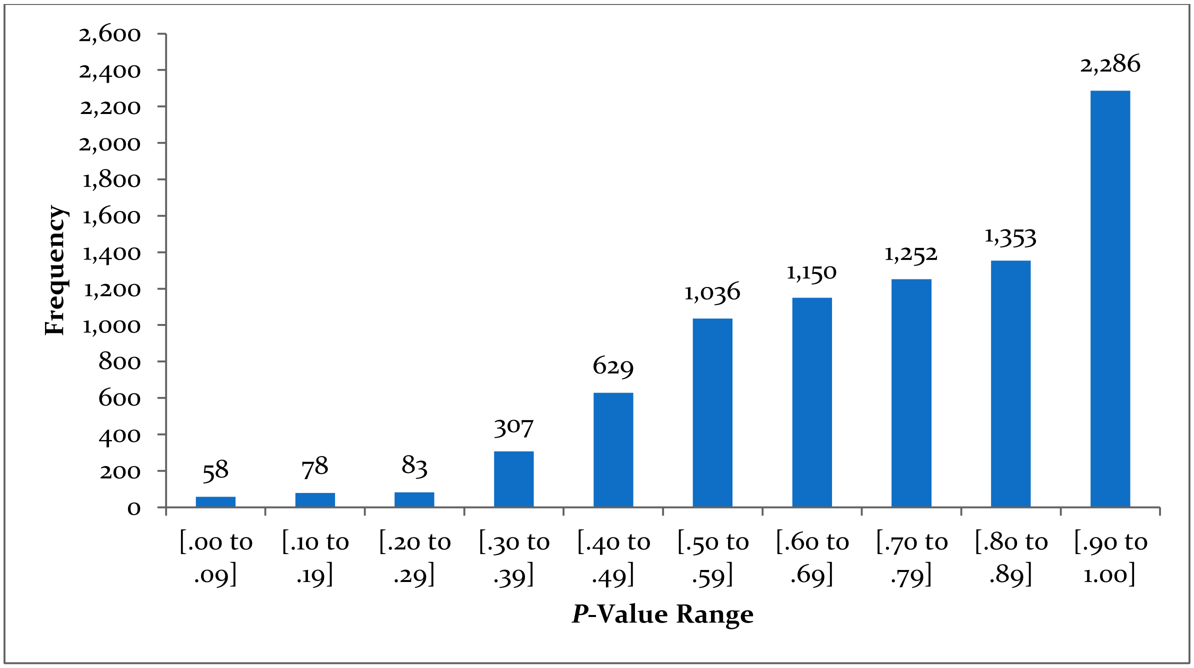

Figure 9. ELA *p*-values for field test items meeting sample size thresholds during all 2014 and 2015 field testing windows.

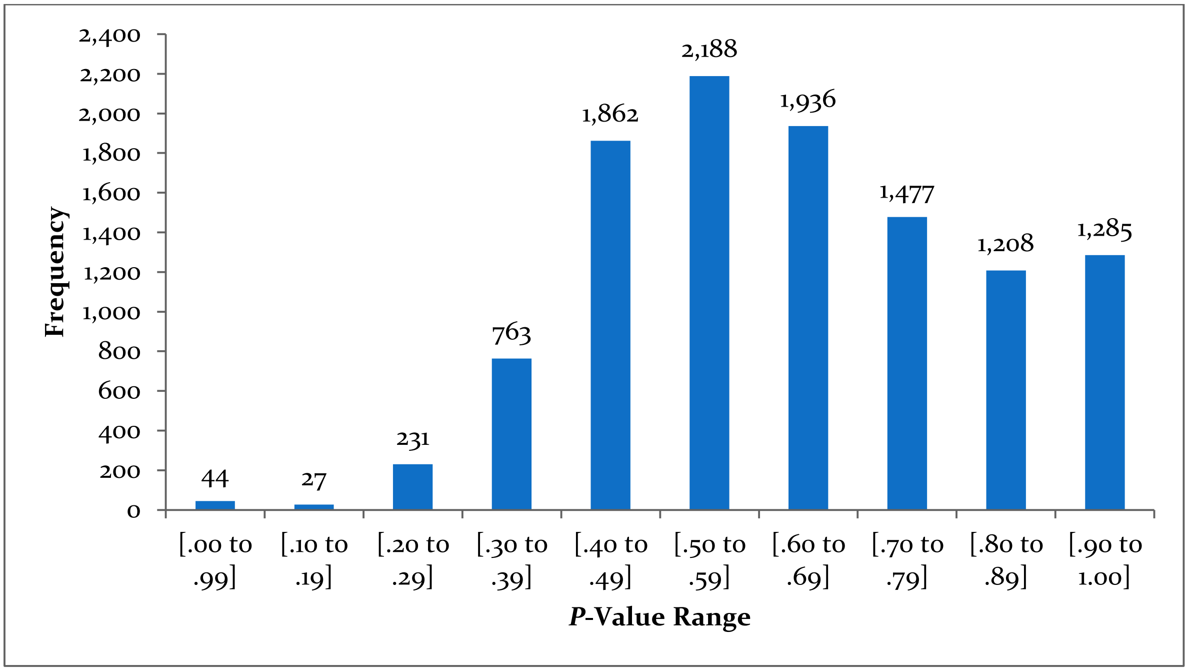

Figure 10. Mathematics *p*-values for field test items meeting sample size thresholds during all 2014 and 2015 field testing windows.

Figure 11 and Figure 12 provide standardized difference values for items in all field test windows in which sample size was at least 20 for ELA and mathematics, respectively. The vast majority of items fall within two standard deviations of the mean *p* value by node. Items falling beyond that threshold were flagged for review by content teams.

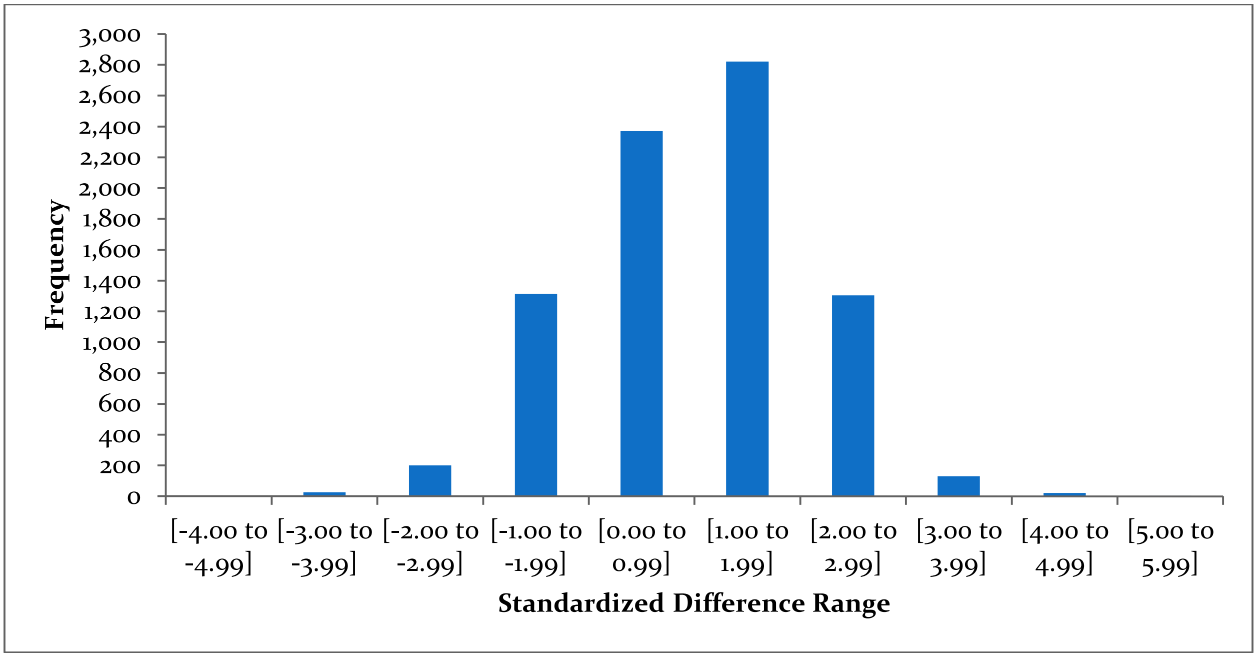

Figure 11. ELA standardized difference z-scores for items meeting sample size thresholds during all 2014 and 2015 field test windows.

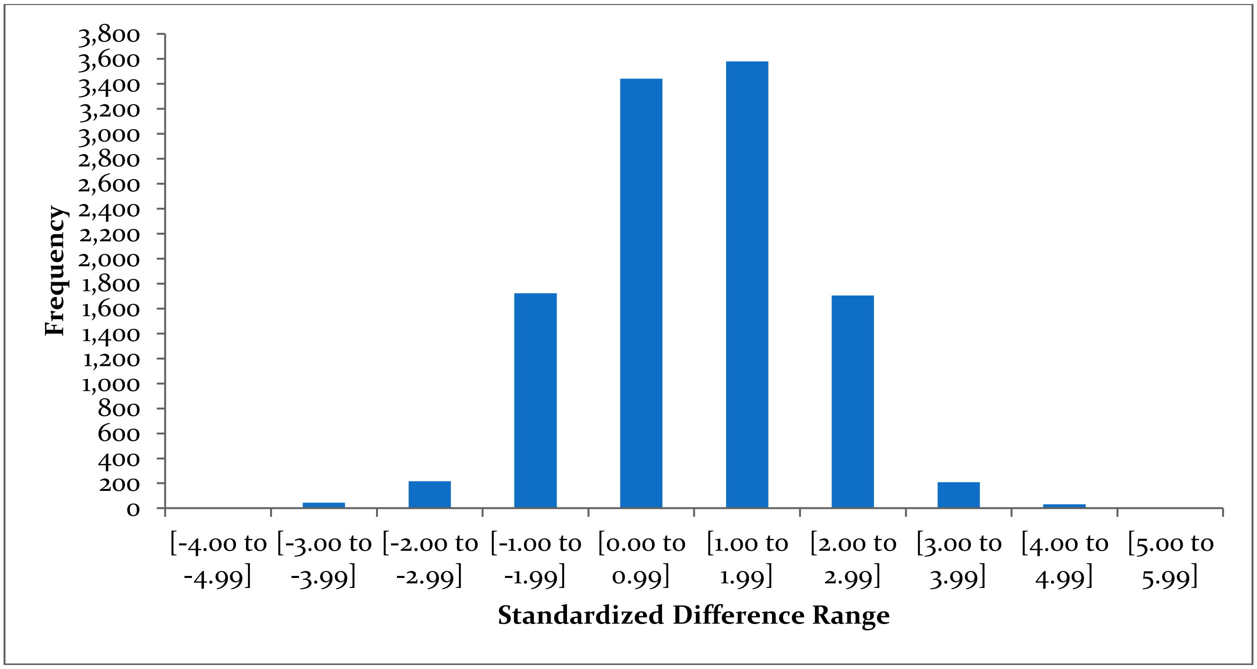

Figure 12. Mathematics standardized difference z-scores for items meeting sample size thresholds during all 2014 and 2015 field test windows.

Table 9, Table 10, and Table 11 summarize the number of items flagged, the total number of items field-tested, and the percent of items flagged that also met sampl -size thresholds for single-EE items in Field Tests 1–3, single-EE items in Phases A–C, and multi-EE items in Phases A–C. Items were included in the count of flagged items if they were flagged for one or more criteria. Across both content areas, 515 items (12.2%) were flagged in Field Tests 1–3 and 2,875 items (22.5%) were flagged during Phases A–C as needing review by content teams. While Phases B and C contained a mix of operational and field test content, this report only includes data for field test items, because operational content had previously been reviewed by content teams prior to becoming operational.

Table 9

Item Flags for Single-EE Testlets Administered During Field Test 1 Through Field Test 3

	ELA			Mathematics		
Grade	Flagged Items	Total Items	Percent Flagged	Flagged Items	Total Items	Percent Flagged
3	41	241	17.0	34	308	11.0
4	15	218	6.9	27	319	8.5
5	8	230	3.5	50	306	16.3
6	9	216	4.2	61	302	20.2
7	41	278	14.7	63	424	14.9
8	25	226	11.1	35	299	11.7
9–10	25	233	10.7	68	353	19.3
11–12	13	283	4.6	N/A	N/A	N/A
Total	177	1,925	9.2	338	2,311	14.6

Note. For mathematics, high school testlets are administered in a single band of grades 9–11, as opposed to the two grade bands used in ELA. Students were routed to the next highest or next lowest linkage level based on their proportion of correct responses for the EE in which they answered the lowest proportion of items correctly.

Table 10

Item Flags for Single-EE Testlets Administered During Phase A Through Phase C

	ELA			Mathematics		
Grade	Flagged Items	Total Items	Percent Flagged	Flagged Items	Total Items	Percent Flagged
3	94	565	16.6	122	601	20.3
4	39	542	7.2	125	863	14.5
5	40	633	6.3	116	732	15.8
6	77	564	13.7	122	698	17.5
7	85	525	16.2	142	835	17.0
8	71	491	14.5	152	798	19.0
9–10	94	520	18.1	466	1,678	27.8
11–12	131	519	25.2	N/A	N/A	N/A
Total	631	4,359	14.5	1,245	6,205	20.1

Note. For mathematics, high school testlets are administered in a single band of grades 9–11, as opposed to the two grade bands used in ELA. Students were routed to the next highest or next lowest linkage level based on their proportion of correct responses for the EE in which they answered the lowest proportion of items correctly.

Table 11

Item Flags for Multi-EE Testlets Administered During Phase A Through Phase C

	ELA			Mathematics		
Grade	Flagged Items	Total Items	Percent Flagged	Flagged Items	Total Items	Percent Flagged
3	44	253	17.4	71	398	17.8
4	34	276	12.3	55	362	15.2
5	22	313	7.0	76	430	17.7
6	39	329	11.9	91	482	18.9
7	46	285	16.1	119	459	25.9
8	37	311	11.9	84	418	20.1
9	26	297	8.8	85	384	22.1
10	29	189	15.3	33	161	20.5
11	34	278	12.2	74	236	31.4
Total	311	2,531	12.3	688	3,330	20.7

Content teams reviewed all flagged items to determine possible reasons for the flag and whether an edit was likely to resolve the issue. Upon examining an item's content, the team made one of three decisions: accept as is, revise the content, or reject outright. For an item to be accepted as is, the content team had to have determined that the item was consistent with DLM item-writing guidelines and that the item was aligned to the node. An item or testlet was rejected completely if it was inconsistent with DLM item-writing guidelines, the EE and linkage level were covered by other testlets that had better performing items, or there was not a clear content-based revision to improve the item. In some instances, a decision to reject an item resulted in the rejection of the testlet as well.

Common reasons for editing an item included item mis-keys (i.e., no correct response indicated or an incorrect response option was labeled as the correct option), item misalignment to the node, distractors that could be argued as partially correct options, or unnecessary complexity in the language of the stem.

Table 12 provides the counts for content team acceptances, revisions, and rejections for ELA items field-tested during Field Tests 1–3. In ELA, 76 items and 38 testlets were rejected. The ELA content team elected to reject some items outright when the testlet already had four or five items, rather than make edits to one poorly performing item.

Table 12

ELA Team Responses to Item Flags From Field Test 1 Through Field Test 3

Grade	Flagged Items	Accept		Revise		Reject	
		n	%	*n*	%	*n*	%
3	41	34	82.9	1	2.4	6	14.6
4	15	6	40.0	1	6.7	8	53.3
5	8	3	37.5	2	25.0	3	37.5
6	9	5	55.6	1	11.1	3	33.3
7	41	18	43.9	3	7.3	20	48.8
8	25	3	12.0	1	4.0	21	84.0
9–10	25	8	32.0	6	24.0	11	44.0
11–12	13	6	46.2	3	23.1	4	30.8
Total	177	83	46.9	18	10.2	76	42.9

Table 13 provides the counts for the content team acceptances, revisions, and rejections for mathematics items field-tested during Field Tests 1–3. In mathematics, 26 items and 8 testlets were rejected. The higher acceptance rate (65%) in mathematics can partially be explained by the prevalence of flags for Initial Precursor testlets. These testlets generally contain five-option multiple-choice items rather than the typical three-option multiple-choice items found on testlets at higher linkage levels and were flagged due to the *p*-value falling below the threshold of .35. Upon closer evaluation by the content team, the response most commonly chosen for these flagged items was “no response.” The content team determined the predominance of selecting this response option did not relate to the item’s quality, but instead likely resulted from the testing situation or the student’s lack of opportunity to learn the field tested content. In these instances, the decision was made to retain the items in their field-tested format and evaluate the items again after additional data collection.

Table 13

Mathematics Team Responses to Item Flags From Field Test 1 Through Field Test 3

Grade	Flagged Items	Accept		Revise		Reject	
		n	%	*n*	%	*n*	%
3	34	28	82.4	6	17.6	0	0.0
4	27	21	77.8	6	22.2	0	0.0
5	50	42	84.0	8	16.0	0	0.0
6	61	42	68.9	11	18.0	8	13.1
7	63	37	58.7	26	41.3	0	0.0
8	35	16	45.7	15	42.9	4	11.4
9–11	68	34	50.0	20	29.4	14	20.6
Total	338	220	65.1	92	27.2	26	7.7

Table 14 and Table 15 provide the counts for the content team acceptances, revisions, and rejections for ELA items field-tested during Phases A–C for single-EE and multi-EE testlets, respectively. In ELA, 251 items and 47 testlets were rejected following item review by the content team.

Table 14

ELA Team Responses to Flags From Phase A Through Phase C for Single-EE Testlets

	Flagged	Accept		Revise		Reject	
Grade	Items	*n*	%	*n*	%	*n*	%
3	94	75	79.8	9	9.6	10	10.6
4	39	25	64.1	4	10.3	10	25.6
5	40	24	60.0	7	17.5	9	22.5
6	77	39	50.6	7	9.1	31	40.3
7	85	43	50.6	9	10.6	33	38.8
8	71	47	66.2	4	5.6	20	28.2
9–10	94	21	22.3	6	6.4	67	71.3
11–12	131	51	38.9	12	9.2	68	51.9
Total	631	325	51.5	58	9.2	248	39.3

Table 15

ELA Team Responses to Flags From Phase A Through Phase C for Multi-EE Testlets

	Flagged	Accept		Revise		Reject	
Grade	Items	*n*	%	*n*	%	*n*	%
3	44	35	79.5	9	20.5	0	0.0
4	34	30	88.2	4	11.8	0	0.0
5	22	21	95.5	1	4.5	0	0.0
6	39	34	87.2	5	12.8	0	0.0
7	46	39	84.8	7	15.2	0	0.0
8	37	29	78.4	5	13.5	3	8.1
9	26	23	88.5	3	11.5	0	0.0
10	29	26	89.7	3	10.3	0	0.0
11	34	26	76.5	8	23.5	0	0.0
Total	311	263	84.6	45	14.5	3	1.0

Table 16 and Table 17 provide the counts for the content team acceptances, revisions, and rejections for mathematics items field-tested during Phases A–C for single-EE and multi-EE testlets, respectively. In mathematics, 158 items and 54 testlets were rejected.

Table 16

Mathematics Team Responses to Flags From Phase A Through Phase C for Single-EE Testlets

Grade	Flagged Items	Accept		Revise		Reject	
		n	%	*n*	%	*n*	%
3	122	52	42.6	48	39.3	22	18.0
4	125	53	42.4	61	48.8	11	8.8
5	116	60	51.7	41	35.3	15	12.9
6	122	76	62.3	40	32.8	6	4.9
7	142	63	44.4	73	51.4	6	4.2
8	152	80	52.6	61	40.1	11	7.2
9–11	466	208	44.6	219	47.0	39	8.4
Total	1,245	592	47.6	543	43.6	110	8.8

Table 17

Mathematics Team Responses to Flags From Phase A Through Phase C for Multi-EE Testlets

Grade	Flagged Items	Accept		Revise		Reject	
		n	%	*n*	%	*n*	%
3	71	34	47.9	34	47.9	3	4.2
4	55	31	56.4	22	40.0	2	3.6
5	76	24	31.6	49	64.5	3	3.9
6	91	32	35.2	43	47.3	16	17.6
7	119	43	36.1	65	54.6	11	9.2
8	84	30	35.7	46	54.8	8	9.5
9	85	36	42.4	49	57.6	0	0.0
10	33	14	42.4	16	48.5	3	9.1
11	74	32	43.2	40	54.1	2	2.7
Total	688	276	40.1	364	52.9	48	7.0

Initialization

The goal of DLM initialization is to provide an optimal match for students during their first DLM testing experience; that is, administered items should match students' knowledge, skill, and ability levels as closely as possible. The initialization algorithm was first examined following the 2013 pilot administration of ELA and mathematics assessments (see Clark, Kingston, Templin, & Pardos, 2014). Following this examination, the initialization algorithm was implemented for all field tests. Because the pilot sample size was low, the proportion of students classified to each complexity band was examined during each field test window to better understand the distribution of students within the population. This section of the report summarizes those findings.

Initialization During Field Testing

Initialization during the six field test events was based on educator responses to the First Contact survey. The specific items informing initialization included responses to items about expressive communication and questions about ELA and mathematics content areas. The First Contact survey questions used for initialization are summarized in Table 20.

Table 20

First Contact Survey Items Used for Initialization

Content Area	Item
ELA	Student’s approximate instructional reading level in print or braille
ELA	Student's ability to recognize single symbols presented visually or tactually[a]
Mathematics	Student's ability to sort objects by common properties (e.g., color, size, shape) [a]
Mathematics	Student's ability to add or subtract by joining or separating groups of objects [a]
Mathematics	Student's ability to form groups of objects for multiplication and division [a]
Mathematics	Student's ability to multiply or divide using numerals [a]
Communication	Student's use of speech, sign, or symbols to meet expressive communication needs
Communication	Student's highest level of expressive communication

[a]Response options included the percent of time the student demonstrates the behavior.

During the initialization process, students were assigned to one of four complexity bands for each content area based on responses to the First Contact items. Complexity bands spanned from Foundational to Band 3. Based on the student's assignment to a complexity band, a testlet at a specific linkage level was administered to the student. Table 21 shows the correspondence of complexity bands to linkage level testlets.

Table 21

Correspondence of Complexity Band to Linkage Level Testlets for Field Test 3 and Phase A

Complexity Band	Linkage Level
Foundational	Initial Precursor
Band 1	Distal Precursor
Band 2	Proximal Precursor
Band 3	Target or Successor

Student Classification to Complexity Bands

Two approaches to initialization were evaluated following the pilot administration of the DLM assessment (Clark, Kingston, Templin, & Pardos, 2014). The first approach to initialization was to calculate a complexity band for each content area based on the responses to that content area's First Contact items (see Table 20). The second approach was to also calculate an expressive communication band and use the lower of the complexity bands (content or communication) as the basis for assigning a testlet. Communication variables were included in the second approach because testlets at higher linkage levels require a certain level of expressive communication. Following the pilot event, the decision was made to include expressive communication in the algorithm under the assumption that it is better for a student's first testlet to be too easy rather than too hard and to avoid expressive communication limitations acting as a barrier to students demonstrating knowledge, skills, and understanding.

Based on the process of selecting the lower of the content or expressive communication complexity band, a small number of students were placed in a lower complexity band than if only the content area band were considered. Table 25 shows the number and percent of students reclassified to a lower complexity band after the expressive communication items were included in the initialization algorithm for the pilot and Field Tests 1–3. Because estimates from these four events were relatively stable, analyses were not repeated for Phases A–C. The percentage of students reclassified was comparable across testing events, each event impacting 10% or fewer students.

Table 25

Percent of Students Reclassified After Expressive Communication was Included in Initialization

	Event	*n*	%
ELA	Pilot	50	7
	Field Test 1	907	10
	Field Test 2	938	9
	Field Test 3	849	9
Math	Pilot	47	7
	Field Test 1	732	8
	Field Test 2	767	7
	Field Test 3	713	8

For each of the 2014 and 2015 field test windows, student classification to each complexity band was compared to baseline percentages from the spring 2013 First Contact survey administration and the percentages observed during the fall 2013 pilot administration.

The baseline data collected during spring 2013 included First Contact responses from the 13 states participating in the DLM Consortium at the time. Results from this initial completion of the survey produced baseline estimates of the percentage of students in each complexity band. Since the initial administration of the First Contact survey, the states included in the consortium have changed. As a result, the proportion of students in each complexity band has also changed. Table 23 and Table 24 summarize the percent of students by complexity band for each testing window for ELA and mathematics, respectively. As compared to the baseline values, the field test proportions indicate that more students were classified to the Foundational and Band 1 complexity bands, and fewer students were classified to Band 2 and Band 3. Across field test events, roughly 20% of the DLM population was classified to the Foundational band, 30% to Band 1, 35% to Band 2, and 15% to Band 3.

Table 23

Percent of Students Classified to ELA Complexity Bands, by Event

Complexity Band	Baseline (*N*=44,550)	Pilot (*N*=1,409)	FT1 (*N*=9,615)	FT2 (*N*=10,445)	FT3 (*N*=9,731)	Phase A (*N*=10,181)	Phase B (*N*=14,617)	Phase C (*N*=17,997)
Foundational	12	23	18	20	19	20	18	18
Band 1	26	33	29	30	29	30	31	30
Band 2	38	31	36	35	35	35	36	36
Band 3	23	13	17	16	16	15	15	16

Table 24

Percent of Students Classified to Mathematics Complexity Bands, by Event

Complexity Band	Baseline (*N*=44,549)	Pilot (*N*=1,409)	FT1 (*N*=9,615)	FT2 (*N*=10,445)	FT3 (*N*=9,731)	Phase A (*N*=10,181)	Phase B (*N*=14,617)	Phase C (*N*=17,997)
Foundational	13	24	20	21	21	21	20	20
Band 1	28	32	31	32	32	33	34	33
Band 2	41	36	38	37	37	36	36	37
Band 3	18	10	11	10	10	10	10	10

Accessibility

Overview

The DLM staff intentionally considers accessibility as part of the design of the assessment system. In addition to the use of universal design principles during test development, the design of the user interface to serve students with the most significant cognitive disabilities, and the use of First Contact survey results to guide testlet delivery, the DLM assessments include accessibility supports so educators can customize each student's testing experience. Educators choose each student's accessibility supports and mark them on the Personal Needs & Preferences (PNP) Profile. Supports that were available during field testing are summarized in Figure 13.

Supports Provided Via the PNP Profile	Supports Requiring Additional Tools or Materials	Supports Provided Outside the System
• Magnification • Invert Color Choice • Color Contrast • Overlay Color • Spoken Audio (Synthetic Text to Speech)	• Uncontracted braille • Single-switch system (PNP enabled) • Two-switch system • Administration via iPad • Adaptive equipment used by student • Individualized Manipulatives	• Human Read Aloud • Sign interpretation of text • Language translation of text • Test administrator entering of responses for student • Partner-Assisted Scanning

Figure 13. PNP supports available during field testing. Only the supports provided via the PNP Profile were recorded during Field Test 3. Spoken audio was only available during Field Test 3.

This report describes the use of accessibility supports during field testing. It also summarizes the field testing process and outcomes for the two types of alternate forms that may be delivered based on the PNP Profile: braille and alternate forms for students who are blind or have visual impairments (BVI). Feedback regarding educator and student experience with accessibility supports is described in the summary of DLM survey results (see Clark, Brussow, & Karvonen, 2016).

Accessibility Supports Selected During Field Testing

The accessibility features in the PNP Profile are listed in four categories: display enhancements, language and braille, audio and environment support, and other supports. Display enhancements, including color contrast, color overlay, and magnification, are provided within the testing platform. Other supports available during field testing require additional tools or materials, such as single- and two-switch systems and alternate forms for visually impaired students. Some supports in the PNP must be provided by the test administrator, such as human read aloud or sign interpretation of text.

Table 18 summarizes the accessibility features that educators activated in the PNP for Field Test 3, which included a limited number of accessibility supports. Note that the PNP file is cumulative; therefore, a file from Field Test 3 includes all PNP entries from the entire 2013–2014 academic year. During Field Test 3, Text to Speech was the most selected support, being used by 72% of students taking the assessment.

Table 18

Accessibility Features Selected During Field Test 3 (N = 9,755)

PNP	*n*	%
Braille	40	< 1
Color Contrast	1,364	14
Color Overlay	869	9
Invert Color Choice	452	5
Magnification	2,071	21
Spoken Audio (Synthetic Read Aloud)	7,024	72

During the 2014–2015 year, 19,662 students who participated in one or more of the field tests during Phase A, Phase B, and Phase C had a completed PNP Profile. Table 19 summarizes the number and percent of students participating in Phases A–C who had PNP features selected. System read aloud of testlets was not available for Phases A–C.

Table 19

Accessibility Features Selected for Students During Phase A through Phase C (N = 19,662)

PNP	*n*	%
Braille	31	< 1
Two-Switch System	175	1
Translation	198	1
Sign Interpretation	373	2
Alternate Form: Visual Impairment	488	2
Invert Color Choice	727	4
Color Overlay	967	5
Color Contrast	1,063	5
Partner-Assisted Scanning	1,056	5
Single-Switch System	1,153	6
Magnification	1,455	7
Individualized manipulatives	5,807	30
Test Administrator Entering of Responses	7,262	37
Human Read Aloud	17,867	91

Field Tests of Alternate Forms

The DLM assessment has two types of alternate testlets designed for students with vision-related disabilities. One or both alternate forms may be available for a testlet, depending on the EE and linkage level being tested. Alternate forms for students who are blind or have a visual impairment (BVI) were used on a limited basis, when the content of a general testlet could not be made accessible to students who are blind or have visual impairments even with the accessibility supports available for DLM assessments. The second type of alternate form was braille. Because of the cognitive complexity required of students to read braille, only testlets at the Target and Successor levels were made available in braille for all grades. Testlets at the Proximal Precursor level were also available in braille for grades 6–8 and high school. All braille testlets were copies of testlets previously field-tested; no new testlets were written specifically for braille forms.

The Phase B field test included a limited number of alternate BVI testlets. A total of 21 single-EE and 12 multi-EE BVI testlets were available during the window. Additionally, 40 ELA and 152 mathematics single-EE BVI testlets and 22 ELA and 139 mathematics multi-EE BVI testlets were available during the Phase C window. After field testing, BVI items that met sample size requirements were included in the content-team item evaluation summarized above. Any items not meeting sample size requirements were field-tested again. Where BVI testlets were needed to provide complete blueprint coverage in the spring operational window item pools, content teams reviewed items that had not met sample size requirements and made content-based judgments regarding the items' suitability for operational use based on the evaluation of similar testlets that had already met sample-size requirements.

Braille testlets were made available during a separate braille field testing window within Phase C, spanning from January 19 through February 13, 2015. A total of 20 ELA and 16 mathematics single-EE braille testlets were made available for field testing, along with 8 ELA and 6 mathematics multi-EE testlets. Participation in the braille field test was purely voluntary. In IM states, student names had to be submitted prior to the braille field test in order to participate. For YE states, students were automatically assigned the braille field test if braille was indicated on the student’s PNP Profile and a braille testlet was available for field-testing.

A total of six IM students participated in the field test, all from a single state. Participating students were in grades 6 and 8 and high school. These students received up to four forms each in ELA and mathematics. No students from YE states participated in the braille field test. Because of the small number of students participating, no braille items met the sample size threshold necessary for statistical item review.

Conclusions

To prepare for operational testing, six field test windows were implemented during 2014 and 2015 to collect data on single-EE and multi-EE testlets. DLM flagging rules were applied to all field tested items. Flagging rates were generally low across all grades, content areas, and testing models, with rates ranging from 3.5% to 27.8%.

The vast majority of content that was field-tested during 2014 and 2015 was either accepted outright or with revisions. Less than 1% of mathematics tasks and only 1% of ELA tasks that were field-tested were rejected. This finding supports DLM's approach to evidence-centered design, whereby tasks and testlets are specifically written using EE concept maps with accessibility, bias and sensitivity, and content in mind (see Clark, Karvonen, & Swinburne Romine, 2014; Clark, Swinburne Romine, Bell, & Karvonen, 2015).

Additional embedded field testing will be conducted during the 2015–2016 academic year to ensure the depth of item and testlet pools for each blueprint model and to support the psychometric modeling used for scoring and reporting. Single-EE testlet field-testing will occur during both instructionally embedded testing and spring windows, while multi-EE testlet field-testing will occur during the spring window. Once field-tested, single-EE content is divided between the instructionally embedded testing window and the spring IM window. All multi-EE content is delivered during the spring window for states in the YE blueprint model.

Item flagging criteria used through 2015 were based on classical item statistics because calibration data for operational scoring based on diagnostic classification modeling was not yet available. As the psychometric model is updated, future reviews of items and testlets will incorporate results from the type of diagnostic classification model used for scoring. This will include three additional flagging constraints for content teams to consider. Flagging will include non-informative items, which are items that do not demonstrate a notable increase in the likelihood of a correct response for students who master the assessed node over students who do not master the node; this indicator is similar to item discrimination indices. Node reversals will also be flagged for item review. Node reversals occur when non-masters of a node have a high chance of being a master on subsequent nodes. This may be due to the way items are written for the assessed node. Finally, node overspecification will be flagged to indicate nodes that do not appear to be distinct from one another. Again, this may be a result of items being written to assess the two nodes in too similar of a way to distinguish the two types of skills being assessed. As with the current flagging approach, content teams will review all flags in the context of the content being assessed to determine if edits are needed.

During 2014 and 2015, some field testing was conducted for alternate forms, including braille and BVI testlets. The field testing of alternate forms was limited by sample size constraints. The low number of students available for field testing BVI forms made it difficult to obtain samples large enough to evaluate item quality. Content teams relied more heavily on content-based reviews of the testlets than flagging results because most items did not reach the sample size threshold of 20 students. As sample size increases for alternate form testlets over additional testing windows, items will be reviewed for any flags that result.

Across the field test events, analysis of First Contact–based complexity band assignments indicated a shift in the student population, as shown by differences in the percentage of students classified to each complexity band between the baseline spring 2013 administration of the First Contact survey and the six field test windows in 2014 and 2015. This shift is likely due to changes in the states participating in the consortium, including the total number of states, and to state differences in eligibility criteria for alternate assessments. As compared to the baseline results, the field test data indicated a higher percentage of students assessed at the foundational band. Similarly, the field tests had a smaller percentage of students in the higher complexity bands, resulting in lower sample sizes for Target and Successor testlets.

Because classification to complexity bands during field testing was comparable to that observed during the fall 2013 pilot administration, the initialization algorithm implemented during field testing was maintained during spring 2015 operational testing and the 2015–2016 academic year. However, additional research is underway to evaluate whether modifications should be made to optimize the algorithm. The current research specifically examines the algorithm for mathematics testlets because field test data indicates that mathematics items may be more challenging than ELA items and because students' opportunity to learn the mathematics content measured on DLM assessments is broader than what is captured in the four First Contact survey items used in the current initialization algorithm.

References

Clark, A., Brussow, J., & Karvonen, M. (2016). *Summary of results from the Dynamic Learning Maps® teacher surveys* (Technical Report No. 15-05). Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation.

Clark, A., Karvonen, M., & Swinburne Romine, R. (2014). *Results from external review during the 2013–2014 academic year* (Technical Report No. 14-02). Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation.

Clark, A., Kingston, N., Templin, J., & Pardos, Z. (2014). *Summary of results from the fall 2013 pilot administration of the Dynamic Learning Maps® Alternate Assessment System* (Technical Report No. 14-01). Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation.

Clark, A., Swinburne Romine, R., Bell, B., & Karvonen, M. (2015). *Results from external review during the 2014–2015 academic year* (Technical Report No. 15-01). Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation.