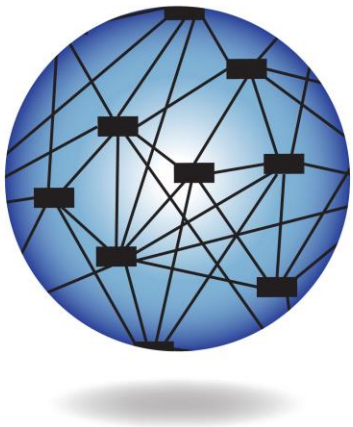

DYNAMIC™
LEARNING MAPS

Summary of Results from the Fall 2013 Pilot Administration of the Dynamic Learning Maps™ Alternate Assessment System

---

Technical Report #14-01

3/24/2014

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.

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 author(s), and no official endorsement by the U.S. Department should be inferred.

Table of Contents

Introduction

A pilot administration of the Dynamic Learning Maps™ Alternate Assessment System was conducted in the fall of 2013. The pilot assessment was available to teachers and students in states belonging to the Dynamic Learning Maps Consortium from October 21 to November 22, 2013. A total of 1,409 students completed assessments and 597 teachers responded to teacher surveys using the Dynamic Learning Maps (DLM®) system.

The mathematics and English language arts (ELA) content teams each selected a single Essential Element to be assessed in each of three grade bands: third-fourth grade, seventh-eighth grade, and high school. A fixed-form assessment was built for each grade band that assessed the Essential Element at three different linkage levels. All forms consisted of three testlets: a testlet at the initial precursor linkage level, a testlet at the distal or proximal precursor linkage level, and a testlet at the target linkage level. All students started at the least difficult level and were given the option to exit at any time.

The primary purpose of the pilot assessment was to evaluate the method for assigning students to an initial assessment within the system. The DLM test development team was also interested in teachers' perceptions of the assessment system. Specific research questions included the following:

- Will complexity bands support the KITE system to present an item that is relatively well matched to students' knowledge, skills, and abilities, as evidenced by teacher responses to the First Contact Survey?
- How do teachers perceive the current system features?
    - Do features meet student needs?
    - Do features function as expected?
    - What suggestions do teachers have for improvement?
    - How do students interact with the assessment system (e.g., level of engagement, level of independence)?
- What can we learn from teachers regarding administration and training recommendations?
    - What is the approximate number of testlets that teachers will likely administer during a single session?
    - Which areas should we concentrate on for improving the test delivery engine?
- What are the initial findings of exploratory modeling work when fitting cognitive diagnostic models and Bayesian networks to data from the pilot?

The report that follows includes a summary of findings from the pilot administration. The first section of the report provides an overview of the analyses conducted to evaluate the initialization process. The second section includes a summary

of exploratory findings from the initial modeling work conducted with data obtained from the pilot. The last section provides an overview of the feedback received from the teacher survey.

Initialization

The primary purpose of the pilot assessment was to gain information about initial student entry into the DLM assessment system. The goal is to provide an optimal match for students during their initial DLM testing experience; that is, items presented to students should provide the best possible match to their knowledge, skill, and ability levels. After students' initial testing experience, the system dynamically routes the student through the learning map based on their responses and provides testlets at the appropriate level of complexity.

Baseline data was obtained through the pilot to evaluate student initialization. A fixed-form assessment was administered in each of the three grade bands to ascertain how students with varying knowledge, skill, and ability levels responded to testlets spanning a range of complexity levels. Each form included three testlets that assessed a single Essential Element at three different linkage levels. The first testlet was administered at the initial precursor linkage level, the second testlet at the distal or proximal precursor level, and the final testlet at the target linkage level. Students were able to exit the assessment at any point if the content became too challenging. By administering to all students in a grade band the same set of testlets for a single Essential Element, the DLM test development team was able to gauge how a range of students responded to the varied levels of complexity and will use that data to inform initial assignment to a complexity band.

Responses to the First Contact Survey were used to create the initial assignment of testlet complexity. Two approaches were evaluated based on First Contact responses for students taking the pilot assessment:

1) Assign each student to an initial complexity band based solely on First Contact responses that pertain to academic performance in ELA or mathematics; or
2) Assign each student to an initial complexity band based on a combination of First Contact responses that pertain to ELA or mathematics *and* the student's expressive communication ability.

Content and special education experts selected the First Contact Survey items to be used for initialization. The survey items for ELA entry asked about each student's reading level and how often the student correctly recognizes single symbols. The survey items for mathematics entry asked how often a student correctly performs the following tasks: sorting; addition and subtraction; forming groups of objects to multiply or divide; and multiplication and division. The survey items for expressive communication asked whether a student expressively communicates using speech, sign language, or augmentative or alternative communication (AAC), and whether the student regularly combines 0, 1, 2, or 3 or more spoken words, signs, or symbols. The DLM test development team chose to include expressive communication variables in the second initialization approach because nodes assessed at higher linkage levels often require more concrete or abstract symbolic communication. Based on teacher responses to these First

Contact Survey items, students were assigned to one of four complexity bands, from foundational to complexity band 3, using a decision tree for each content area.

Comparison of Approaches

Data collected from the pilot assessment was used to evaluate the two proposed initialization methods. The percentages of students classified in each complexity band for ELA and mathematics are presented in Table 1. Similar values are evident in ELA and mathematics. These values indicate that the combined approach of using content area and expressive communication variables provides a slightly more conservative classification to complexity bands; thus a small percentage of students are placed at a lower complexity band after taking into account their expressive communication ability. The percentage of students impacted ranges from 5% to 9% based on grade and content area.

Table 1

Percentages of Students Classified into Complexity Bands

Complexity band	ELA		Mathematics	
	Content only	Combined	Content only	Combined
Foundational	20%	23%	20%	24%
Complexity band 1	31%	33%	32%	32%
Complexity band 2	33%	31%	36%	36%
Complexity band 3	16%	13%	12%	10%

Based on these findings, the DLM test development team decided to move forward with the combined algorithm of content and expressive communication items for initialization for the first field testing event and continue investigating complexity bands with additional data. Although the decision would result in a small portion of students being placed at an initially lower complexity level, the DLM test development team believes it is preferential to have students enter the assessment with items that are too easy than with items that are too difficult. The conservative approach would potentially provide more students with a positive initial testing experience. As previously stated, beyond the initial testing experience, students' response patterns will be used to modify students' complexity band classifications as needed.

Examination Within and Across Complexity Bands

To determine whether the complexity bands provided meaningful distinction between students at varying levels of knowledge, skill, and ability, analyses were conducted to determine the extent that students categorized to the four complexity bands differed from one another. One expected finding is, if the complexity bands provide a meaningful distinction between students, then the percentage of students responding correctly to items should increase as the complexity band increases, and the percentage of students within a complexity band who respond correctly to items should decrease as linkage level increases.

In Tables 2 and 3, the columns labeled 1, 2, and 3 represent items within the administered testlets. Items in testlet 1 were at the initial precursor linkage level, testlet 2 at the distal precursor level, and testlet 3 at the target level, with difficulty increasing over testlets. The rows represent students grouped by complexity bands, increasing from foundational (F) to complexity band 3 (CB 3). The tables provide the percentage of correct responses for each item administered in the seventh-eighth grade band assessment for ELA and mathematics, including items that were not attempted.

Table 2

Seventh-Eighth Grade ELA Percentage Correct by Item

Complexity band	Testlet 1			Testlet 2			Testlet 3			
	1	2	3	1	2	3	1	2	3	4
F (*N*=90)	39%	36%	43%	24%	28%	27%	27%	24%	26%	22%
CB 1 (*N*=92)	75%	46%	62%	32%	39%	42%	40%	34%	28%	41%
CB 2 (*N*=114)	96%	82%	79%	77%	59%	72%	50%	53%	75%	67%
CB 3 (*N*=54)	100%	94%	94%	93%	67%	93%	67%	78%	81%	83%

Table 3

Seventh-Eighth Grade Mathematics Percentage Correct by Item

Complexity band	Testlet 1				Testlet 2			Testlet 3		
	1	2	3	4	1	2	3	1	2	3
F (*N*=90)	39%	39%	33%	31%	27%	16%	26%	22%	25%	22%
CB 1 (*N*=92)	47%	48%	54%	51%	40%	30%	27%	22%	31%	19%
CB 2 (*N*=114)	68%	76%	78%	74%	67%	45%	55%	36%	41%	44%
CB 3 (*N*=54)	76%	89%	87%	78%	84%	64%	87%	80%	53%	76%

As expected, the percentage correct at the item level is lowest for students at the foundational level, and increases as the complexity band increases. Similarly, because the testlets are ordered from lowest linkage level to highest, percentage correct generally decreases from testlet 1 to testlet 3. Similar results were found across grade bands and content areas. These findings are one source of evidence indicating that the complexity bands are useful for creating a meaningful distinction among students in order to provide them with the best match of item complexity during the initial testing experience.

Because students were able to exit the assessment at any time, the DLM test development team was interested in determining how many students within each complexity band attempted all three testlets. Table 4 provides the percentages of students who attempted all three testlets, by grade band and content area. These findings indicate that students at the foundational level attempted all three testlets less frequently than students at higher complexity bands. This is an expected finding, as students at the foundational level would typically only be administered testlets at the initial precursor level, and only the first testlet in the pilot was at this level. Students at complexity band 3 would typically be assigned items at the target level or beyond and thus would be expected to be able to respond to all content presented in the pilot. Future analyses will examine completion rates within each testlet.

Table 4

Percentage of Students Who Attempted All Testlets by Grade and Content Area

Complexity band	ELA			Mathematics		
	3rd-4th	7th-8th	HS	3rd-4th	7th-8th	HS
F	53%	77%	81%	63%	69%	68%
CB 1	85%	79%	79%	86%	76%	72%
CB 2	84%	92%	90%	89%	88%	98%
CB 3	100%	94%	100%	75%	98%	95%

To further evaluate whether complexity bands successfully distinguished between groups of students, exploratory analyses were conducted to evaluate a portion of the teacher survey responses by complexity band. Responses for a small subset of survey items are presented here, with the full survey results included in the third section of this report.

Teachers were able to indicate on the teacher survey the reason for exiting a testlet prior to completion. These findings were examined across complexity bands to determine where similarities or differences were evident. Because a separate complexity band was calculated for each content area, results are presented by content area even though the survey questions were not content specific. Findings for each ELA complexity band are presented in Table 5. Note that percentages add up within a column rather than across a row. These values were consistent with those observed for the mathematics complexity band and, thus, only the ELA table is presented. Values indicate that the percentage of students who did not exit a testlet prior to completion increased across complexity bands. Of those students at the foundational level who did exit a testlet, the most frequent reason was the student did not know the content, while for students in complexity band 3, the most common reason was frustration or disengagement.

Table 5

Reasons for Exiting Testlets by Complexity Band

Reason for exiting	F		CB 1		CB 2		CB 3	
	n	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Did not exit	162	69%	162	70%	169	80%	71	89%
Extreme frustration or disengagement	15	6%	21	9%	4	2%	4	5%
Student's behavior or health interfered	7	3%	10	4%	1	1%	0	0%
Accidental exit	6	3%	7	3%	16	7%	1	1%
Student did not know anything about the content	36	15%	22	9%	9	4%	0	0%
Accessibility features were not working	1	1%	4	2%	1	1%	1	1%
Other reason	7	3%	6	3%	11	5%	3	4%

Student interaction with the testing system was also examined by complexity band to determine whether level of independence varied by band. Table 6 presents the findings for the mathematics complexity bands. Similar values were observed for the ELA complexity bands. Students classified in lower complexity bands had less independence when interacting with the system, while students classified in higher complexity bands had greater levels of independence. Few students at any complexity band interacted with the assessment system without any prompting, redirection, or support from a teacher.

Table 6

Student Interaction with the System by Mathematics Complexity Bands

Type of interaction	F		CB1		CB2		CB3	
	n	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Did *not* require supports *and* entered responses independently	1	2%	3	6%	8	17%	1	8%
Required supports *and* entered responses independently	2	5%	6	13%	27	58%	9	76%
Did *not* require supports *and* did *not* enter responses independently	2	5%	8	17%	10	21%	1	8%
Required supports *and* did *not* enter responses independently	39	88%	31	64%	2	4%	1	8%

Regression Analyses

To further evaluate the extent to which the proposed initialization algorithm was supported by the pilot data and to explore alternate modeling approaches, a series of regression analyses were conducted. A hierarchical ordinary least squares regression model was used to predict the total score for each content area assessment using the previously specified First Contact Survey variables. Many of the items had to be dummy-coded because they are categorical variables. This created a large number of predictors, so a reduced set of variables was used to remove redundancy in the number and overlap of variables for each skill. Variables included addition/subtraction and sorting for mathematics; two reading levels (up to primer level and beyond primer level) and symbol recognition for ELA; and a single expressive communication variable reflecting the student's highest level of expressive communication using spoken word, sign, or AAC. The hierarchical nature of the model was such that the mathematics and ELA First Contact predictors were added to the model first, followed by the expressive communication variable, to determine the extent to which additional variance was explained by its inclusion.

The hierarchical ordinary least squares regression models were statistically significant for both ELA and mathematics across all three grade-band assessments (see Table 7). The amount of variance explained by the mathematics First Contact predictors was between 14% and 38%, with an additional 3% to 5% of variance explained by including the expressive communication variable. For ELA, the amount of variance explained by the First Contact predictors was between 17% and 53%, with an additional 6% to 9% of variance explained by the inclusion of the expressive communication

variable. For all grade bands and content areas, the addition of expressive communication resulted in a significant change to model-data fit.

Table 7

Ordinary Least Squares Regression Results by Grade and Content Area

Grade and content area	*F*	*df*	*p*	R^2
3rd-4th grade mathematics	16.9	2, 258	< .001	.14
7th-8th grade mathematics	59.4	2, 247	< .001	.35
High school mathematics	21.3	2, 107	< .001	.38
3rd-4th grade ELA	14.3	2, 258	< .001	.17
7th-8th grade ELA	96.8	2, 247	< .001	.52
High school ELA	14.0	2, 107	< .001	.26

Next, a hierarchical ordinal logistic regression model was used to predict the probability of success for students at each linkage level testlet. Success at the testlet level was determined by obtaining a threshold of 67% correct. The same First Contact variables were used as predictors. Again, the mathematics variables were significant predictors of mathematics linkage level, χ^2 (7) = 165.24, p < .001, with a Nagelkerke pseudo R^2 value of .26. The expressive communication variable raised the value by .02. Similar findings were evident for ELA, χ^2 (4) =117.21, p < .001, with a Nagelkerke value of .18. The inclusion of the expressive communication variable increased the value by .04. For both content areas, the addition of the expressive communication variable resulted in a significant change to Nagelkerke pseudo R^2 values. Similar findings were obtained using binary logistic regression models to predict success at each linkage level testlet independently.

Predicted and observed values were compared and the root mean squared error (RMSE) was calculated to quantitatively capture how accurate each model was in predicting actual student values for the three linkage level categories. These values are presented in Tables 8 through 11. The RMSE values indicate a sample standard deviation of around 1.0 for all models. The addition of expressive communication variables to the models resulted in a slightly smaller RMSE value for both content areas, and more conservative classification to linkage levels.

Table 8

Ordinal Logistic Regression RMSE Values for Mathematics Only

Observed	Predicted Ordinal Logistic Regression			
	1	2	3	Total
1	155	0	54	209
2	46	0	73	119
3	57	0	187	244
Total	258	0	314	572
RMSE				0.99

Table 9

Ordinal Logistic Regression RMSE Values for Mathematics Combined

Observed	Predicted Ordinal Logistic Regression			
	1	2	3	Total
1	152	0	57	209
2	41	0	78	119
3	49	0	195	244
Total	242	0	330	572
RMSE				0.97

Table 10

Ordinal Logistic Regression RMSE Values for ELA Only

Observed	Predicted Ordinal Logistic Regression			
	1	2	3	Total
1	51	0	117	168
2	12	0	50	62
3	25	0	316	341
Total	88	0	483	571
RMSE				1.05

Table 11

Ordinal Logistic Regression RMSE Values for ELA Combined

Observed	Predicted Ordinal Logistic Regression			
	1	2	3	Total
1	94	0	74	168
2	19	0	43	62
3	65	0	276	341
Total	178	0	393	571
RMSE				1.04

RMSE values were also calculated for the decision tree approach based on First Contact Survey responses. The foundational level corresponded with linkage level 1, or the initial precursor level. Complexity bands 1 and 2 corresponded with linkage level 2, or the distal or proximal precursor levels. Complexity band 3 corresponded with linkage level 3, or the target level. Observed and predicted values are presented in Tables 12 through 15.

Table 12

RMSE Values for ELA Only Decision Tree

Observed	Predicted Decision Tree			
	1	2	3	Total
1	66	123	6	195
2	17	48	5	70
3	30	239	94	363
Total	113	410	105	628
RMSE				0.92

Table 13

RMSE Values for ELA Combined Decision Tree

Observed	Predicted Decision Tree			
	1	2	3	Total
1	71	121	3	195
2	19	46	5	70
3	43	241	79	363
Total	133	408	87	628
RMSE				0.95

Table 14

RMSE Values for Mathematics Only Decision Tree

Observed	Predicted Decision Tree			
	1	2	3	Total
1	98	143	6	247
2	15	104	7	126
3	21	175	64	260
Total	134	422	77	633
RMSE				0.84

Table 15

RMSE Values for Mathematics Combined Decision Tree

Observed	Predicted Decision Tree			
	1	2	3	Total
1	107	137	3	247
2	20	100	6	126
3	29	174	57	260
Total	156	411	66	633
RMSE				0.86

Summary

Taken together, the regression findings suggest that the First Contact Survey academic variables selected for learning map initialization were successful predictors of performance on the pilot assessment. The analyses also provided additional support for the inclusion of expressive communication in the initialization algorithm. Analyses will continue to be conducted with data obtained from the field tests to further evaluate the appropriateness of the proposed initialization algorithm.

Modeling

Pilot Test Modeling Overview

The DLM test development team is exploring the use of two modeling approaches, both of which will be used for long-term modeling work: cognitive diagnostic modeling and Bayesian networks. Although the sample size of the pilot was small, the DLM test development team wanted to conduct preliminary modeling using data from the pilot. Data obtained from the field tests and beyond will contain larger sample sizes and will continue to inform this work.

Cognitive Diagnostic Modeling

The main goals of the modeling work conducted with pilot test data were to evaluate the structure of tests and to identify any psychometric concerns for future large-scale implementation of these models in the DLM project. Modeling analyses were conducted for each grade band and content area separately. All items were analyzed. For each grade band and content area, three "super nodes" were modeled corresponding to the learning map. Each super node consisted of the items assessed for one of the three linkage level testlets used in the pilot assessment. The modeling framework selected was marginal maximum likelihood using an expectation maximum (EM) algorithm because of its frequent use in large scale testing.

Pilot test comparison psychometric models.

Four modeling approaches were compared using data obtained from the pilot. For the log-linear cognitive diagnosis model (LCDM) and hierarchical diagnostic classification model (HDCM), the three super nodes were classified for each student using a binary system of mastered or non-mastered. With HDCM, the super nodes reflected the hierarchical structure of the learning map, where mastery of one node precedes mastery of another; for LCDM any structure was permissible. A two-parameter logistic item response theory (IRT) model was also included, with a single continuous trait measured by all items. Finally, a confirmatory multidimensional IRT model was also used, where each super node represented multiple continuous traits. These modeling approaches were then repeated with continuous testlet factors to account for additional dependency in the pilot data, which was potentially caused by the use of testlet-item structure.

Pilot test model comparison results.

For data obtained from the mathematics pilot assessments, the four testlet-based models were nearly uniformly preferred based on model fit indices. The structure of the learning map was not always confirmed by the analysis; rather, a more general structure

appeared to be present. The use of the binary and continuous traits seemed equally prevalent based on the data obtained from the pilot. Similar findings were present for ELA.

Pilot test modeling findings summary.

In summary, the data obtained from the pilot assessment suggests that models nearly always require testlet effects to be included. In addition, the structure of the learning map did not always appear when a flexible model was used, which may be due to the non-overlap of items and the small pilot sample size. Analyses conducted with data obtained from the field tests will show more detail about the impact of map structure on modeling efforts.

Psychometric concerns for large scale implementation.

Based on analyses conducted with pilot data, the need to develop estimation methods that are flexible is evident. In addition, the inclusion of additional variables in the models, including testlet-level or student-level data, will be important to consider when conducting future DLM modeling work.

Bayesian Modeling

Does the map topology add diagnostic power?

Significant time and effort has gone into creating the nodes and the prerequisite relationships in the mathematics and ELA learning maps. These maps have important instructional value because they lay out principally derived pathways that teachers can use to arrive at the desired learning objectives for their students. But how much diagnostic information can be leveraged from the structure of the map?

We used the pilot data to investigate the diagnostic value of the prerequisite connections by comparing the predictive accuracy of a map that uses linkage level connection information versus a unidimensional model that connects all questions to a single latent trait. We used Bayesian inference to fit the model and make predictions with a cross-validation holdout strategy, a statistically strong validation technique used in machine learning. The result was that the linkage level map provided only negligible improvement in average prediction accuracy: 72.7% versus 72.3% with the unidimensional map. Both models were substantively better than using simple percentage correct to predict (63.2% accuracy). Among the six testlets evaluated, only the ELA 4.3 Essential Element showed statistically significant prediction improvement with the DLM linkage level maps (75.2% versus 73.1%). Essential Element 4.3 asks that students "use details from the text to describe characters in the story." This evaluation tested prediction on a random set of questions given response evidence from a complimentary random set.

A common decision in item selection will be to decide whether to give students more items at an easier level or a harder level based on their responses at the current level. We evaluated which model would be better at predicting responses outside of the current level and found that the unidimensional model predicted responses at the precursor level more accurately on average than the linkage level map (73.2% versus 71.6%); two Essential Elements had statistically significant results (ELA 7.3 and mathematics 3.4). The unidimensional model was less accurate than the linkage level map at predicting successor levels (71.3% versus 71.5%). However, this result had no statistical significance. A combination of models depending on task will be considered next.

Modeling next steps.

The pilot study only categorized items within a particular linkage level. Full node-level information will be modeled with field test data and compared against the linkage level and unidimensional modeling. We will also investigate, with more data, if a particular modeling level performs more accurately at predicting precursor and successor levels, the implications of this on testlet selection, and the effect of taking into account student First Contact Survey information on prediction accuracy. Lastly, we will investigate the strength of particular prerequisite relationship assumptions and integrate this model assessment into visualizations.

Teacher Survey

As part of the pilot testing event, teachers were asked to complete a survey about each participating student's experience with the assessment. This survey was designed to provide the DLM test development team with feedback on the initial use of the testing platform and user interface. All participating teachers were presented with seven survey items on a common form, followed by one of five randomly administered forms containing between 2 and 12 additional items. The survey contained a mix of multiple-choice and open-ended response items. Teachers were not required to respond to all items and could exit the survey at any time.

A total of 1,209 teacher responses to the survey were recorded, indicating a response rate of around 86%. The breakdown of responses by grade band is presented in Table 16. The DLM test development team is very pleased with the response rate of teachers participating in the survey at each grade band and plans to continue to include survey items as part of each field test event.

Table 16

Teacher Responses to Survey by Grade Band

Grade band	Students assessed	Teacher responses	%
3rd-4th grade	477	400	84%
7th-8th grade	546	464	85%
High school	393	324	82%

Multiple-Choice Items

Teachers were asked to provide a baseline rating of the assessment system, knowing that improvements would continue to be made for the field tests. Teachers were asked to rank the system using A, B, C, D, or F, with A being the highest rating. Teachers most often provided a midline system ranking of C. The number and percentage of each rating are presented in Table 17.

Table 17

Teacher Baseline Rating of Assessment System

Rating	*n*	%
A	51	4%
B	337	28%
C	407	34%
D	238	20%
F	164	14%

During the pilot administration, students were able to exit the assessment at any time. Teachers were asked to provide rationale for exiting a testlet prior to completion. Teachers indicated that a total of 436 students, or 36%, exited a testlet prior to its completion. The reasons for exiting a testlet are listed in Table 18. Teachers indicated that, of those students who exited a testlet early, approximately two-thirds returned to the test after exiting, while one-third did not return to the test.

Table 18

Reasons for Exiting the Assessment System Prior to Completion

Reason for exiting	*n*	%
Student was frustrated or disengaged	119	27%
Student did not know anything about the content	119	27%
Other reason	87	20%
Accidental exit	54	12%
Student's behavior or health interfered	38	9%
Accessibility features were not working	19	5%

An additional survey item asked teachers to indicate how many testlets they would be likely to administer to each student during a future testing session. A total of 38% indicated they would administer one testlet of three to five items in a single testing session, while 31% indicated they would administer two or three testlets, respectively.

The DLM test development team was interested in evaluating how independently students interacted with the testing system during the initial pilot experience. The team expects that as students become more familiar with the system over time, their level of independence will increase. Two survey items were included to gauge student

independence when interacting with the system. One of these items asked teachers to rate the student's level of independence when responding to multiple-choice items. The teachers chose ratings ranging from 1 (no independence) to 5 (complete independence). A total of 33% of teachers indicated their student had no independence; 15% indicated their student had complete independence; and about 17% indicated their student fell in between (ranks 2 to 4). A second survey item asked teachers to indicate which of four options best described their student's interaction with the system. These findings are presented in Table 19. Approximately 45% of students required prompting, support, or redirection from their teacher during the assessment *and* could not enter their own responses on the computer. In contrast, approximately 10% of students required no prompting, support, or redirection from their teacher and entered all responses independently.

Table 19

Student Interaction with the DLM Assessment System

Student interaction	*n*	%
Did *not* require supports *and* entered responses independently	36	10%
Required supports *and* entered responses independently	108	32%
Did *not* require supports *and* did *not* enter responses independently	43	13%
Required supports *and* did *not* enter responses independently	153	45%

Feedback on student engagement during the pilot was also obtained through the teacher survey. One item asked teachers to rate each student's level of engagement with multiple-choice items administered on the computer from 1 to 5, with 1 being completely unengaged and 5 being completely engaged. Findings indicated an even split of approximately 20% of responses at each possible rating. A similar item asked teachers to rate the student's engagement with test items administered by the teacher directly to the student. Using the same scale, responses were as follows: 1) 14%, 2) 19%, 3) 18%, 4) 22%, and 5) 27%, indicating slightly higher levels of engagement for tasks administered to the student by the teacher as compared to directly on the computer. This finding may be related to the discussion above about student independence when interacting with the computer and might also be expected to change over time, as the student has more interaction with the system.

Input from teachers was also sought to determine whether system functionality met the students' needs. Teachers indicated that about 45% of students made use of the highlighting feature, which allows students to highlight important text. It met the needs of about 56% of the students who used it. Most students (67%) did not make use of the magnification feature during the pilot. It met the needs of 52% of the students who used

it. Teachers indicated that the ability to leave the testing session inactive for up to 30 minutes met 72% of students' needs, and the ability to exit and return to the testlet later met 78% of students' needs.

Several survey items were also included about the format of items and testlets administered during the pilot. Two item types were included in the pilot test: multiple-choice and drag-and-drop items. Teachers indicated that the multiple-choice item type met 68% of students' needs, while the drag-and-drop item type, used only in high school ELA testlets, met 65% of students' needs. Teachers indicated the amount of text presented on a single screen met 69% students' needs, and the engagement activities at the beginning of each testlet met the needs of 56% of students.

Open-Ended Items

In addition to responding to multiple-choice items about the pilot, teachers were also asked to provide open-ended feedback pertaining to the pilot testing experience and teacher preparation for upcoming field test events.

Preparation.

With regard to teacher preparation for the field test event in January, 2014, the teachers provided a great deal of feedback. Teachers suggested having easier access to usernames and passwords. They also suggested preparing two types of administration manuals: a one-page "quick sheet" and a detailed step-by-step manual complete with screenshots to aid in administering the exam. Teachers wanted to be able to preview these materials prior to administering the field test with their students. In response to these suggestions, the pilot administration materials were reorganized and expanded upon to ensure teachers had the information they need in an accessible format for Field Test 1.

Teachers also requested more training and practice activities for the field test events. Additional training modules were prepared for Field Test 1 and were made available to teachers to view before the field test window opened.

Item content.

One common request from teachers was to have prior access to content that will be covered during the field test events. The DLM test development team will continue to make available to teachers the nodes in the learning map and the Essential Elements that will be assessed during each field test event. In addition, the DLM test development team will provide content-specific vocabulary that will appear on the field test items so teachers can use it during instruction.

Another frequently received comment from teachers pertained to the desire for a greater number of images in the items. For ELA testlets, the DLM test development team had previously determined that a single picture would be presented with each screen,

which typically contains a sentence or two of text. The DLM test development team also had designed texts to de-emphasize images as tools to support comprehension. As such, the images included may not completely represent the content presented on the screen. These comments do suggest a need for more teacher education about the intentional design of ELA texts. For mathematics, the pilot testlets contained fewer items with images than is typical in the pool of mathematics items overall. Many mathematics items that will be included in future testing events include images in the stem and/or answer options.

In addition to comments about the content of the items, many teachers also commented on the level of difficulty of the items included in the pilot. Many teachers stated that items were too challenging or too easy for their students. These comments were expected due to the structure of the fixed-form pilot assessment. The DLM test development team chose to administer fixed forms spanning a range of linkage level testlets, from initial precursor to target level, in order to obtain information about the ideal point of entry for students with varying levels of knowledge, skill, and ability. Because of this, students were presented with a wider range of testlet complexity than they ordinarily would receive during a DLM testing session. While upcoming field tests will continue to evaluate initial linkage level placement, data obtained from the pilot will help the DLM test development team administer content that is more closely aligned with each student's knowledge, skill, and ability level.

Testing platform.

Teachers were also asked to provide feedback on the functionality of the assessment system itself. One frequent comment from teachers was a desire for more information about the accessibility features available for students in the personal needs profile. A revised document will be available to teachers during the field test events with clear explanations of these features. In addition, logins for simulated students will continue to be available to teachers so they can preview accessibility features while engaging in practice test activities.

Summary

The DLM test development team received an exceptional rate of responses to the teacher survey administered as part of the pilot. Teacher feedback spanned a variety of areas, including item and testlet construction, presentation, and system functionality. The DLM test development team will use feedback from the teacher survey to improve the content, system functionality, and professional development for upcoming field test events.