DYNAMIC®
LEARNING MAPS

2017–2018 Technical Manual Update

Science

December 2018

Dynamic Learning Maps Consortium. (2018, December). *2017–2018 Technical Manual Update—Science*. Lawrence, KS: University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS).

Acknowledgements

The publication of this technical manual update builds upon the documentation presented in the *2015–2016 Technical Manual—Science* and annual technical manual updates. This document represents further contributions to a body of work in the service of supporting a meaningful assessment system designed to serve students with the most significant cognitive disabilities. Hundreds of people have contributed to this undertaking. We acknowledge them all for their contributions.

Many contributors made the writing of this technical manual update possible. Dynamic Learning Maps® (DLM®) staff who made significant writing contributions to this technical manual update are listed below with gratitude.

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

The authors wish to acknowledge Zakry Akagi-Bustin, Teri Millar, Chelsea Nehler, Noelle Pablo, and Michelle Shipman for their contributions to this update. For a list of project staff who supported the development of this manual through key contributions to design, development, or implementation of the Dynamic Learning Maps Alternate Assessment System, please see the *2015–2016 Technical Manual—Science*, and the subsequent annual technical manual updates.

We are also grateful for the contributions of the members of the DLM Technical Advisory Committee who graciously provided their expertise and feedback. Members of the Technical Advisory Committee during the 2017–2018 operational year include:

Russell Almond, Ph.D., *Florida State University*
Greg Camilli, Ph.D., *Rutgers University*
Karla Egan, Ph.D., *Independent Consultant*
Robert Henson, Ph.D., *University of North Carolina-Greensboro*
James Pellegrino, Ph.D., *University of Illinois-Chicago*
Edward Roeber, Ph.D., *Assessment Solutions Group/Michigan Assessment Consortium*
David Williamson, Ph.D., *Educational Testing Service*
Phoebe Winter, Ph.D., *Independent Consultant*

Contents

List of Tables

List of Figures

1. Introduction

During the 2017–2018 academic year, the Dynamic Learning Maps® (DLM®) Alternate Assessment System offered assessments of student achievement in mathematics, English Language Arts (ELA), and science for students with the most significant cognitive disabilities in grades 3-8 and high school. Due to differences in the development timeline for science, separate technical manuals were prepared for ELA and mathematics (see Dynamic Learning Maps Consortium [DLM Consortium], 2018b; DLM Consortium, 2018c).

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

The DLM Alternate Assessment System is based on the core belief that all students should have access to challenging, grade-level content. Online DLM assessments give students with the most significant cognitive disabilities opportunities to demonstrate what they know in ways that traditional, paper-and-pencil, multiple-choice assessments cannot. A year-end assessment is administered in the spring, and results from that assessment are reported for state accountability purposes and programs.

A complete technical manual was created for the first year of operational administration in science, 2015–2016. The current technical manual provides updates for the 2017–2018 administration; therefore, only sections with updated information are included in this manual. For a complete description of the DLM science assessment system, refer to the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a).

1.1. Background

In 2017–2018, DLM science assessments were administered to students in 14 states and one Bureau of Indian Education school: Alaska, Delaware, Illinois, Iowa, Kansas, Maryland, Missouri, New Hampshire, New Jersey, New York, Oklahoma, Rhode Island, West Virginia, Wisconsin, and Miccosukee Indian School.

In 2017–2018, the Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS) at the University of Kansas (KU) continued to partner with the Center for Literacy and Disability Studies at the University of North Carolina at Chapel Hill and the Center for Research Methods and Data Analysis at KU. The project was also supported by a Technical Advisory Committee (TAC).

1.2. Technical Manual Overview

This manual provides evidence collected during the 2017–2018 administration to evaluate the DLM Consortium's assertion of technical quality and the validity of assessment claims.

Chapter 1 provides a brief overview of the assessment and administration for the 2017–2018 academic year and a summary of contents of the remaining chapters. While subsequent chapters

describe the individual components of the assessment system separately, several key topics are addressed throughout this manual, including accessibility and validity.

Chapter 2 provides an overview of the purpose of the Essential Elements (EEs) for science, including the intended coverage with the *Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (National Research Council, 2012) and the Next Generation Science Standards (NGSS Lead States [NGSS], 2013). For a full description of the process by which the Essential Elements were developed, see the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a).

Chapter 3 outlines evidence related to test content collected during the 2017–2018 administration, including a description of test development activities and the operational and field test content available.

Chapter 4 provides an update on test administration during the 2017–2018 year. The chapter provides updated information about adaptive routing in the system, as well as teacher survey results regarding educator experience and system accessibility.

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

Chapter 6 was not updated for 2017–2018; no changes were made to the cut points used in scoring DLM assessments. See the *2015–2016 Technical Manual Update—Science* (DLM Consortium, 2017a) for a description of the methods, preparations, procedures, and results of the standard-setting meeting and the follow-up evaluation of the impact data.

Chapter 7 reports the 2017–2018 operational results, including student participation data. The chapter details the percentage of students at each performance level; subgroup performance by gender, race, ethnicity, and English-learner status; and the percentage of students who showed mastery at each linkage level. Finally, the chapter provides descriptions of changes to score reports and data files during the 2017–2018 administration.

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

Chapter 9 describes additional validation evidence collected during the 2017–2018 administration not covered in previous chapters. The chapter provides study results for the five critical sources of evidence: test content, internal structure, response process, relation to other variables, and consequences of testing.

Chapter 10 was not updated for 2017–2018. See Chapter 10 in the *2015–2016 Technical Manual Update—Science* (DLM Consortium, 2017a) for a description of the training and instructional activities that were offered across the DLM Science Consortium.

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

2. Essential Element Development

The Essential Elements (EEs) for science, which include three levels of cognitive complexity, are the conceptual and content basis for the Dynamic Learning Maps® (DLM®) Alternate Assessment System for science, with the overarching purpose of supporting students with the most significant cognitive disabilities (SCD) in their learning of science content standards. For a complete description of the process used to develop the EEs for science, based on the organizing structure suggested by the *Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (National Research Council, 2012, "*Framework*" hereafter) and the Next Generation Science Standards (NGSS, 2013), see Chapter 2 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a).

2.1. Purpose of EEs for Science

The EEs for science are specific statements of knowledge and skills linked to the grade-band expectations identified in the *Framework* and NGSS, and they are the content standards on which the alternate assessments are built. The general purpose of the DLM EEs is to build a bridge connecting the content in the *Framework* and NGSS with academic expectations for students with SCD. This section describes the intended breadth of coverage of the DLM EEs for science as it relates to the *Framework* and NGSS. For a complete summary of the process used to develop the EEs, see Chapter 2 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a).

As described in the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a), the *Framework* and NGSS served as the organizing structure for developing the DLM EEs for science. However, as the science state partners did not want to develop EEs for every sub-idea in the *Framework*, a crosswalk of states' existing alternate science standards was used to identify the intended foci for students with SCD and the DLM science assessment. This information was then used to map states' alternate standards to the *Framework* and NGSS. The DLM Science Consortium identified the most frequently assessed topics across states in the three content domains of physical science, life science, and Earth and space science. The analysis of states' alternate content standards resulted in a list of common cross-grade Disciplinary Core Ideas (DCIs) and sub-ideas seen in the *Framework* in states' science standards. From there, states requested that at least one EE be developed under each of the 11 DCIs. Their rationale included a desire for breadth of coverage across the DCIs defined by the *Framework* (i.e., not the breadth of coverage that represented the entire *Framework*), and included content that persisted across grade bands, as well as content that was most important for students with SCD to be prepared for college, career, and community life. As such, the intention was not to develop EEs for every sub-idea in the *Framework*, but rather for a selected subset of sub-ideas across all of the DCIs that would be an appropriate basis for developing alternate content standards for students with SCD.

3. Item and Test Development

Chapter 3 of the Dynamic Learning Maps® (DLM®) Alternate Assessment System *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a) describes item and test development procedures. This chapter provides an overview of updates to item and test development for the 2017–2018 academic year. The first portion of the chapter provides an overview of 2017–2018 item writers' characteristics. The next portion of the chapter describes the pool of operational and field test testlets administered during spring 2018.

For a complete description of item and test development for DLM assessments, including information on the use of evidence-centered design and Universal Design for Learning in the creation of concept maps to guide test development; external review of content; and information on the pool of items available for the pilot, field tests, and 2015–2016 administration, see the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a).

3.1. Items and Testlets

This section describes the items and testlets that are administered as part of the DLM assessment system. For a complete summary of item and testlet development procedures, see Chapter 3 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a).

3.1.1. Items

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

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

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

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

- The first answer option is the key.

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

Refer to Chapter 3 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a) for a complete description of the design of computer-delivered and teacher-administered testlets.

3.1.2. Item Writing

For the 2017–2018 year, only a limited number of items were written to replenish the pool. The item writing process for 2017–2018 began with an on-site event in January 2018. Following this initial event, item writing continued remotely via a secure online platform. A total of 159 testlets were written for science.

3.1.3. Item Writers

An item writer survey was used to collect demographic information about the teachers and other professionals who were hired to write DLM testlets. In total, 17 item writers wrote testlets for the 2017–2018 year. The median and range of years of teaching experience in four areas the item writers had is shown in Table 3.1. The median years of experience was at least 12 years for item writers of science testlets in pre-K–12 and special education, as well as the science subject area.

Table 3.1. Item Writers' Years of Teaching Experience

Area	Median	Range
Pre-K–12	17	3-35
Science	16	3-35
Special Education	12	3-28

Item writers were also asked to indicate the grade or grades they had experience teaching. There were seven item writers with experience at the elementary level (grades 3–5), eleven with experience in middle school (grades 6–8), and seven with experience in high school (grades 9-12).

The distribution and types of degrees held by item writers are shown in Table 3.2. All item writers held at least a Bachelor's degree, with the most common field of study being education, followed by a science-specific field. A majority (n = 15; 88%) also held a Master's degree, and the most common field of study was special education.

Table 3.2. Item Writers' Level and Type of Degree

Degree	*n*	%
Bachelor's Degree	**17**	**100.0**
Education	9	52.9
Content Specific	4	23.5
Special Education	2	11.8
Other	2	11.8
Master's Degree	**15**	**88.2**
Education	2	11.8
Content Specific	2	11.8
Special Education	7	41.2
Other	4	23.5
Other Advanced Degree	**2**	**11.8**

Most item writers had experience working with students with disabilities, as summarized in Table 3.3. Teachers collectively had the most experience working with students with a mild cognitive disability, multiple disabilities, emotional disability, or specific learning disability.

Table 3.3. Item Writers' Experience with Disability Categories

Diability Category	*n*	%
Blind/Low Vision	6	35
Deaf/Hard of Hearing	2	12
Emotional Disability	9	53
Mild Cognitive Disability	10	59
Multiple Disabilities	10	59
Orthopedic Impairment	6	35
Other Health Impairment	8	47
Severe Cognitive Disability	7	41
Specific Learning Disability	9	53
Speech Impairment	6	35
Traumatic Brain Injury	6	35
None of the above	4	24

Of the items writers, 59% had experience administering an Alternate Assessment of Alternate Achievement Standards (AA-AAS) prior to their work on the DLM project, and 35% reported working with students eligible for AA-AAS at the time of the survey.

3.2. External Reviews

Due to the implementation of a new external review timeline, there were limited external review activities during the 2017–2018 year. Because of this, external review activities for recently developed testlets were scheduled for an on-site external review event during summer of 2018 and will be

documented in the *2018–2019 Technical Manual Update—Science*.

3.3. Operational Assessment Items for Spring 2018

A total of 297,859 operational test sessions were administered during the spring testing window. One test session is one testlet taken by one student. Only test sessions that were complete at the close of each testing window counted toward the total sessions.

Testlets were made available for operational testing in spring 2018 based on the 2016–2017 operational pool and the testlets field-tested during 2016–2017 that were promoted to the operational pool following their review. Table 3.4 summarizes the total number of operational testlets for spring 2018 for science. There were 153 operational testlets available across grade bands and courses. This total included 35 Essential Element (EE)/linkage level combinations for which both a general version and a version for students who are blind or visually impaired or read braille were available.

Table 3.4. Distribution of Spring 2018 Operational Testlets, by Grade Band or Course (N = 153)

Grade Band or Course	*n*
Elementary	44
Middle School	45
High School	43
Biology	30

Note: Three EEs are shared across the high school and biology assessment.

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

Figure 3.1 shows the p-values for each operational item in science. To prevent items with small sample sizes from potentially skewing the results, the sample size cutoff for inclusion in the p-value plots was 20. The p-values for most science items were between .5 and .7.

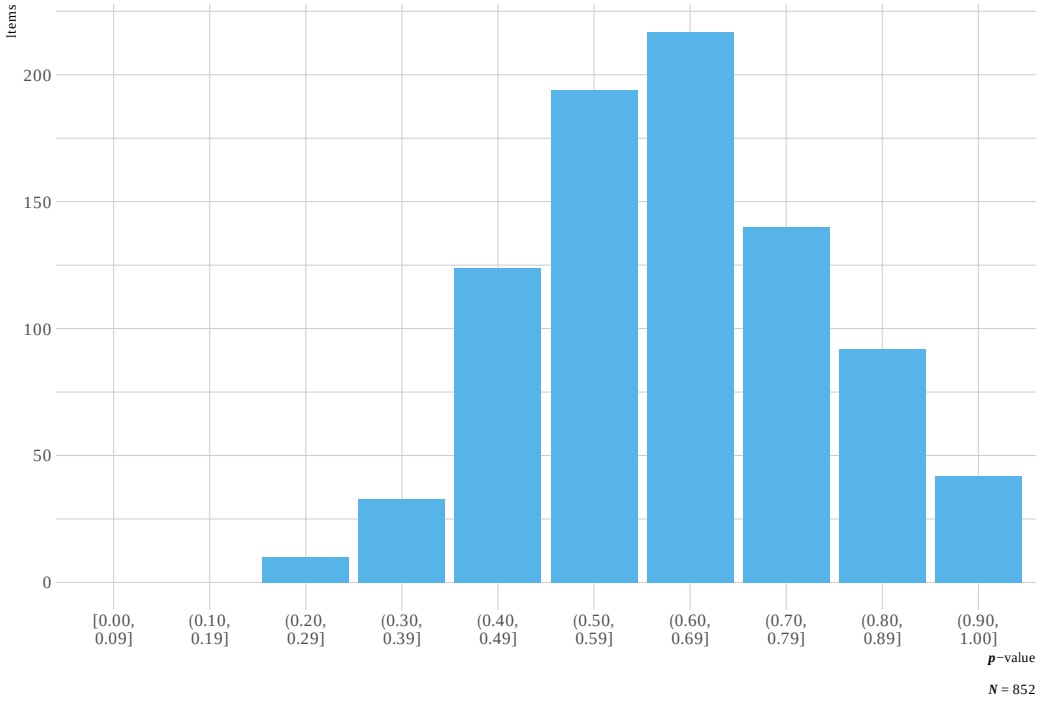


Figure 3.1. *p*-values for science 2018 operational items. *Note*. Items with a sample size of less than 20 were omitted.

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

Figure 3.2 summarizes the standardized difference values for operational items for science. Almost all items fell within two standard deviations of the mean of all items measuring the respective EE and linkage level. As additional data are collected and decisions are made regarding item pool replenishment, test development teams will consider item standardized difference values when determining which items and testlets are recommended for retirement.

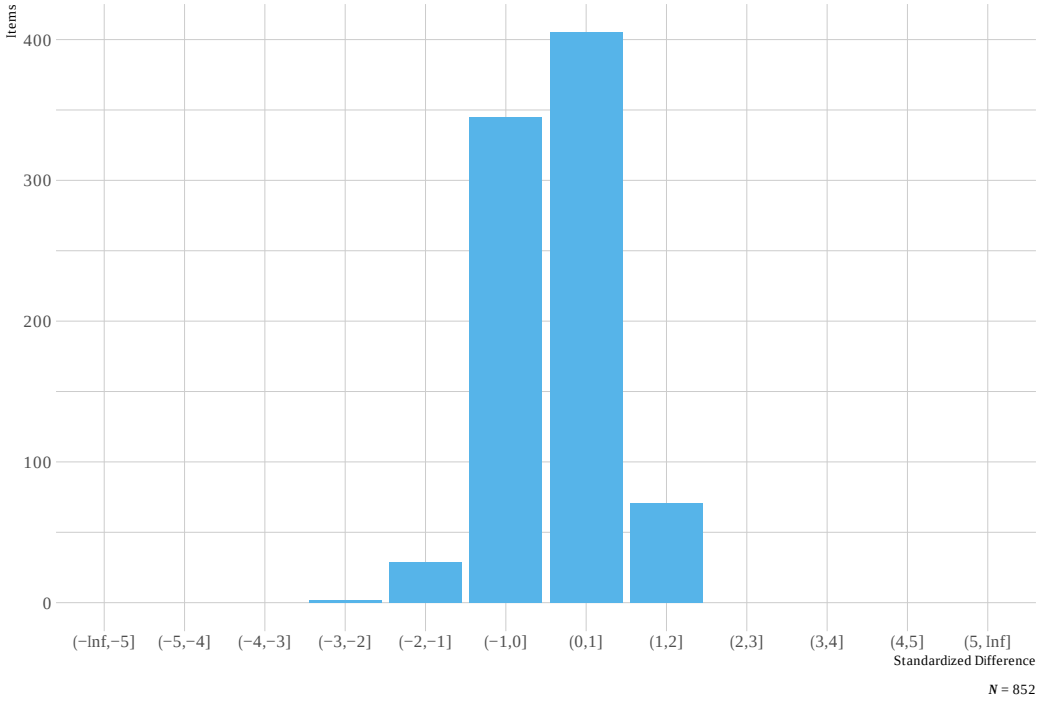


Figure 3.2. Standardized difference *z*-scores for science 2018 operational items. *Note*. Items with a sample size of less than 20 were omitted.

3.4. Field Testing

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

A summary of prior field test events can be found in the *Summary of the Dynamic Learning Maps Science Alternate Assessment Development Process* (Nash & Bechard, 2016).

3.4.1. Description of Field Tests

Field test testlets were administered during the spring window. During the spring administration, all students received a field test testlet upon completion of all operational testlets.

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

test content came from the existing operational pool.

For the spring field test, one EE was selected for field testing from each grade band (elementary, middle school, and high school). Students participating in the end-of-instruction high school biology assessment received the same EE for field testing as the standard high school assessment. This resulted in a total of three EEs being selected for the spring field test. There were three testlets available for each grade band, corresponding with the three linkage levels of the selected EE for each grade band.

Participation in spring field testing was not required in any state, but teachers were encouraged to administer all available testlets to their students. In total, 26,503 (78%) students took at least one field test form. High participation rates allowed for a significant increase in the amount of cross-linkage-level data, furthering modeling research into the structure of the linkage levels with EEs (see Chapter 5 of this manual for future directions). The purpose of the spring field test was to collect additional cross-linkage-level data, and thus the design utilized the pool of currently available operational testlets; therefore, test development team review of items included in the field test was not necessary.

4. Test Administration

Chapter 4 of the Dynamic Learning Maps® (DLM®) Alternate Assessment System *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a) describes general test administration and monitoring procedures. This chapter describes updated procedures and data collected in 2017–2018, including a summary of adaptive routing, total testing time, Personal Needs and Preferences (PNP) profile selections, and teacher survey responses regarding user experience and accessibility.

Overall, administration features remained consistent with the prior year's implementation, including spring administration of testlets, adaptive delivery, and the availability of accessibility supports.

For a complete description of test administration for DLM assessments, including information on administration time, available resources and materials, and information on monitoring assessment administration, see the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a).

4.1. Overview of Key Administration Features

This section describes the testing windows for DLM test administration for 2017–2018. For a complete description of key administration features, including information on assessment delivery, the KITE® system, and linkage level selection, see Chapter 4 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a). Additional information about administration can also be found in the *Test Administration Manual 2017–2018* (DLM Consortium, 2017d) and the *Educator Portal User Guide* (DLM Consortium, 2017c).

4.1.1. Test Windows

New in 2017–2018, instructionally embedded science assessments were available for teachers to optionally administer between September 20 and December 20, 2017, and between January 2 and February 28, 2018. During the consortium-wide spring testing window, which occurred between March 12 and June 8, 2018, students were assessed on each Essential Element (EE) on the blueprint. Each state sets its own testing window within the larger consortium spring window.

4.2. Administration Evidence

This section describes evidence collected during the spring 2018 operational administration of the DLM Science alternate assessment. The categories of evidence include data relating to the adaptive delivery of testlets in the spring window, user experience, and accessibility.

4.2.1. Adaptive Delivery

During the spring 2018 test administration, the science assessment was adaptive between testlets, following the same routing rules applied in prior years. That is, the linkage level associated with the next testlet a student received was based on the student's performance on the most recently administered testlet, with the specific goal of maximizing the match of student knowledge and skill to the appropriate linkage level content.

- The system adapted up one linkage level if the student responded correctly to at least 80% of

the items measuring the previously tested EE. If the previous testlet was at the highest linkage level (i.e., Target), the student remained at that level.
- The system adapted down one linkage level if the student responded correctly to less than 35% of the items measuring the previously tested EE. If the previous testlet was at the lowest linkage level (i.e., Initial), the student remained at that level.
- Testlets remained at the same linkage level if the student responded correctly to between 35% and 80% of the items on the previously tested EE.

The linkage level of the first testlet assigned to a student was based on First Contact survey responses. The correspondence between the First Contact complexity bands and first assigned linkage levels are shown in Table 4.1.

Table 4.1. Correspondence of Complexity Bands and Linkage Level

First Contact complexity band	Linkage level
Foundational	Initial
1	Initial
2	Precursor
3	Target

For a complete description of adaptive delivery procedures, see Chapter 4 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a).

Following the spring 2018 administration, analyses were conducted to determine the mean percentage of testlets that adapted up a linkage level, stayed at the same linkage level, or adapted down a linkage level from the first to second testlet administered for students within a grade band or course and complexity band. The aggregated results can be seen in Table 4.2.

Overall, results were similar to those found in the previous years. For the majority of students across all grade bands who were assigned to the Foundational Complexity Band by the First Contact survey, testlets did not adapt to a higher linkage level after the first assigned testlet (ranging from 56% to 100%). Consistent patterns were not as apparent for students who were assigned Complexity Band 1, Complexity Band 2, or Complexity Band 3. Distributions across the three categories were more variable across grade bands. Further investigation is needed to evaluate reasons for these different patterns.

The 2017–2018 results build on earlier findings from previous years of operational assessment administration and suggest that the First Contact survey complexity band assignment is an effective tool for assigning most students content at appropriate linkage levels. Most students assigned to the Foundational Complexity Band and Complexity Band 3 did not adapt, with between 0% and 44% of students adapted to the available adjacent linkage level, suggesting that the available content served the majority of students' needs. Results also indicate that students assigned to Band 2 were more variable with respect to the direction in which they move between the first and second testlets. Several factors may help explain these results, including more variability in student characteristics within this group and content-based differences across grade bands. Further exploration is needed in this area. Finally, results show that students assigned to Band 1 tended to adapt up a linkage level more frequently, which is an expected finding given that the Foundational and Band 1 students are

both assigned content at the Initial linkage level. Additional analyses are planned to evaluate the adaptation pathways for students assigned to Band 1 in order to determine if changes to the assignment process are needed. For a description of previous findings, see Chapter 4 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a), and the *2016–2017 Technical Manual Update—Science* (DLM Consortium, 2018a).

Table 4.2. Adaptation of Linkage Levels Between First and Second Science Testlets (N = 33,933)

Grade	Foundational		Band 1		Band 2			Band 3	
	Adapted Up (%)	**Did Not Adapt (%)**	**Adapted Up (%)**	**Did Not Adapt (%)**	**Adapted Up (%)**	**Did Not Adapt (%)**	**Adapted Down (%)**	**Did Not Adapt (%)**	**Adapted Down (%)**
3–5	44.1	55.9	79.0	21.0	30.6	43.4	26.0	67.1	32.9
6–8	35.0	65.0	66.6	33.4	39.2	41.4	19.4	67.1	32.9
9–12	32.4	67.6	57.9	42.1	44.5	38.5	17.0	80.8	19.2
Biology	0.0	100.0	26.9	73.1	14.5	40.0	45.5	64.4	35.6

Note: Foundational and Band 1 correspond to testlets at the lowest linkage level, so testlets could not adapt down a linkage level. Band 3 corresponds to testlets at the highest linkage level in science, so testlets could not adapt up a linkage level.

4.2.2. Administration Incidents

As in all previous operational years, testlet assignment during the spring 2018 assessment window was monitored to ensure students were correctly assigned to testlets. Administration incidents that have the potential to affect scoring are reported to states in a supplemental Incident File. Improving on the previous operational years, no incidents were observed during the spring 2018 administration. Assignment to testlets will continue to be monitored in subsequent years to track any potential incidents and report them to state partners.

4.3. Implementation Evidence

This section describes evidence collected during the spring 2018 operational implementation of the DLM Science alternate assessment. The categories of evidence include data relating to user experience and accessibility.

4.3.1. User Experience with the DLM System

User experience with the 2017–2018 assessments was evaluated through the spring 2018 survey, which was disseminated to teachers who had administered a DLM assessment during the spring window. In 2018, the survey was distributed to teachers in KITE Client, where students completed assessments. Each student was assigned a survey for their teacher to complete. The survey included three sections. The first and third sections were fixed across all students, while the second section was spiraled across students, with teachers responding to a block of questions pertaining to accessibility, Educator Portal and KITE Client feedback, the relationship of assessment content to instruction by subject, and teacher experience with the system.

A total of 11,542 teachers from states participating in DLM science assessments responded to the survey (with a response rate of 78.7%) for 24,431 students.

Participating teachers responded to surveys for between one and 29 students. Teachers most frequently reported having 0 to 5 years of experience in science and with students with significant cognitive disabilities. The median response to the number of years of experience in both of these areas was 6 to 10 years. Approximately 56% indicated they had experience administering the DLM science assessment in all three operational years.

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

4.3.1.1. Educator Experience

Survey respondents were asked to reflect on their own experience with the assessments as well as their comfort level and knowledge administering them. Most of the questions required teachers to respond on a four-point scale: *strongly disagree*, *disagree*, *agree*, or *strongly agree*. Responses are summarized in Table 4.3.

Nearly all teachers (96.4%) agreed or strongly agreed that they were confident administering DLM testlets. Most respondents (90%) agreed or strongly agreed that the required test administrator

training prepared them for their responsibilities as test administrators. Most teachers also responded that manuals and the Educator Resources page helped them understand how to use the system (90.4%); that they knew how to use accessibility supports, allowable supports, and options for flexibility (93.6%); and that the Testlet Information Pages helped them deliver the testlets (90%).

Table 4.3. Teacher Responses Regarding Test Administration

Statement	SD		D		A		SA		A+SA	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
Confidence in ability to deliver DLM testlets	63	1.3	109	2.3	2,127	43.9	2,543	52.5	4,670	96.4
Test administrator training prepared respondent for responsibilities of test administrator	130	2.7	348	7.2	2,495	51.7	1,849	38.3	4,344	90.0
Manuals and DLM Educator Resources Page materials helped respondent understand how to use assessment system	116	2.4	346	7.2	2,677	55.6	1,679	34.8	4,356	90.4
Respondent knew how to use accessibility features, allowable supports, and options for flexibility	81	1.7	229	4.7	2,649	54.9	1,866	38.7	4,515	93.6
Testlet Information Pages helped respondent to deliver the testlets	127	2.6	357	7.4	2,596	53.7	1,753	36.3	4,349	90.0

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

4.3.1.1.1. KITE System

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

The software used for the administration of DLM testlets is KITE Client. Teachers were asked to consider their experiences with KITE Client and respond to each question on a five-point scale: *very hard, somewhat hard, neither hard nor easy, somewhat easy,* or *very easy*. Table 4.4 summarizes teacher responses to these questions.

Respondents found it to be either *somewhat easy* or *very easy* to log in to the system (76.1%), to navigate within a testlet (80.9%), to record a response (83.7%), to submit a completed testlet (85%), and to administer testlets on various devices (74.2%). Open-ended survey response feedback indicated testlets were easy to administer and that technology had improved compared to previous years.

Table 4.4. Ease of Using KITE Client

	VH		SH		N		SE		VE		SE+VE	
Statement	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Enter the site	63	1.2	299	5.8	875	16.9	1,578	30.5	2,360	45.6	3,938	76.1
Navigate within a testlet	41	0.8	174	3.4	770	14.9	1,478	28.6	2,700	52.3	4,178	80.9
Record a response	39	0.8	115	2.2	684	13.3	1,336	25.9	2,975	57.8	4,311	83.7
Submit a completed testlet	30	0.6	76	1.5	664	12.9	1,280	24.9	3,091	60.1	4,371	85.0
Administer testlets on various devices	71	1.4	210	4.1	1,048	20.4	1,479	28.8	2,331	45.4	3,810	74.2

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

Educator Portal is an area of the KITE system used to store and manage student data and enter PNP and First Contact information. Teachers were asked to assess the ease of navigating and using Educator Portal for its intended purposes. The data are summarized in Table 4.5 using the same scale used to rate experiences with KITE Client. Overall, respondents' feedback was mixed to favorable: a majority of teachers found it to be either *somewhat easy* or *very easy* to navigate the site (66.9%), enter PNP and First Contact information (72.1%), manage student data (66.5%), manage their accounts (68.6%), or manage tests (68.2%).

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

Table 4.5. Ease of Using Educator Portal

	VH		SH		N		SE		VE		SE+VE	
Statement	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Navigate the site	109	2.4	533	11.8	855	18.9	1,504	33.3	1,517	33.6	3,021	66.9
Enter Access Profile and First Contact information	51	1.1	312	6.9	894	19.9	1,607	35.7	1,639	36.4	3,246	72.1
Manage student data	98	2.2	441	9.8	974	21.6	1,636	36.3	1,363	30.2	2,999	66.5
Manage my account	82	1.8	358	7.9	973	21.6	1,644	36.4	1,454	32.2	3,098	68.6
Manage tests	122	2.7	425	9.4	887	19.7	1,548	34.3	1,529	33.9	3,077	68.2

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

Finally, respondents were asked to rate their overall experience with KITE Client and Educator Portal on a four-point scale: *poor, fair, good,* and *excellent*. Results are summarized in Table 4.6. The majority of respondents reported a positive experience with KITE Client. A total of 83.3% of respondents rated their KITE Client experience as *good* or *excellent*, while 75% rated their overall experience with Educator Portal as *good* or *excellent*.

Table 4.6. Overall Experience With KITE Client and Educator Portal

	Poor		Fair		Good		Excellent	
Statement	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
KITE Client	133	2.6	736	14.2	2,570	49.7	1,737	33.6
Educator Portal	240	4.6	1,056	20.3	2,649	51.0	1,246	24.0

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

4.3.2. Accessibility

Accessibility supports provided in 2017–2018 were the same as those available in previous years. DLM accessibility guidance, in accordance with DLM Consortium (2017b), distinguishes among accessibility supports that are provided in KITE Client via the Access Profile[1], require additional tools or materials, and are provided by the test administrator outside the system.

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

Table 4.7 shows selection rates for the three categories of accessibility supports. The most commonly selected supports were human read aloud, test administrator enters responses for student, and individualized manipulatives. For a complete description of the available accessibility supports, see Chapter 4 in the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a).

Table 4.7. Accessibility Supports Selected for Students (N = 30,948)

Support	*n*	%
Supports provided in KITE Client via Access Profile		
Spoken audio	5,123	16.6
Magnification	3,134	10.1
Color contrast	2,492	8.1
Overlay color	1,648	5.3
Invert color choice	1,128	3.6
Supports requiring additional tools/materials		
Individualized manipulatives	13,057	42.2
Calculator	9,870	31.9
Single-switch system	833	2.7
Alternate form - visual impairment	794	2.6
Two-switch system	413	1.3
Uncontracted braille	13	0.0
Supports provided outside the system		
Human read aloud	27,327	88.3
Test administrator enters responses for student	15,807	51.1
Partner assisted scanning	2,521	8.1
Language translation of text	579	1.9
Sign interpretation of text	478	1.5

Table 4.8 describes teacher responses to survey items about the accessibility supports used during administration. Teachers were asked to respond to two items using a four-point Likert-type scale (*strongly disagree*, *disagree*, *agree*, or *strongly agree*) or indicate if the item did not apply to the student. The majority of teachers agreed that students were able to effectively use accessibility supports (81.6%), and that accessibility supports were similar to ones students used for instruction (82.4%). These data support the conclusions that the accessibility supports of the DLM alternate assessment were effectively used by students, emulated accessibility supports used during instruction, and met student needs for test administration. Additional data will be collected during the spring 2019 survey to determine whether results improve over time.

Table 4.8. Teacher Report of Student Accessibility Experience

	SD		D		A		SA		A+SA		N/A	
Statement	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Student was able to effectively use accessibility features.	163	3.0	185	3.4	2401	44.5	2000	37.1	4401	81.6	649	12.0
Accessibility features were similar to ones student uses for instruction.	146	2.7	207	3.8	2364	44.0	2064	38.4	4428	82.4	597	11.1

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

4.4. Conclusion

During the 2017–2018 academic year, the DLM system was available during two testing windows: an optional instructionally embedded window and the required spring window. Implementation evidence was collected in the form of teacher survey responses regarding user experience, accessibility, and Access Profile selections. Results from the teacher survey indicated that teachers felt confident administering testlets in the system, that KITE Client was easy to use, and that Educator Portal posed some challenges but had improved since the prior year.

5. Modeling

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

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

5.1. Overview of the Psychometric Model

Learning map models, which are networks of sequenced learning targets, are at the core of the DLM assessments in science. Because of the underlying map structure and the goal of providing more fine-grained information beyond a single raw or scale score value when reporting student results, the assessment system provides a profile of skill mastery to summarize student performance. This profile is created using latent class analysis, a form of diagnostic classification modeling, to provide information about student mastery of multiple skills measured by the assessment. Results are reported for each alternate content standard, called an Essential Element (EE), at the three levels of complexity for which science assessments are available: Initial, Precursor, and Target.

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

A description of the DLM scoring model for the 2017–2018 administration follows. Using latent class analysis, a probability of mastery was calculated on a scale from 0 to 1 for each linkage level within each EE. Each linkage level within each EE was considered the latent variable to be measured. Students were then classified into one of two classes for each linkage level of each EE: master or non-master. As described in Chapter 6 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a), a posterior probability of at least .8 was required for mastery classification. Consistent with the assumption of item fungibility, a single set of probabilities of masters and non-masters providing a correct response was estimated for all items within a linkage level. Finally, a structural parameter, which is the proportion of masters for the linkage level (i.e., the analogous map parameter), was also estimated. In total, three parameters per linkage level are specified in the DLM scoring model: a fungible probability for non-masters, a fungible probability for masters, and the proportion of masters.

Following calibration, students' results for each linkage level were combined to determine the highest linkage level mastered for each EE. Although the connections between linkage levels were

not modeled empirically, they were used in the scoring procedures. In particular, if the latent class analysis determined a student had mastered a given linkage level within an EE, then the student was assumed to have mastered all lower levels within that EE.

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

5.2. Calibrated Parameters

As stated in the previous section, the comparable *item parameters* for diagnostic assessments are the conditional probabilities of masters and non-masters providing a correct response to the item. Because of the assumption of fungibility, parameters are calculated for each of the 102 linkage levels in science (3 linkage levels $\times$ 34 EEs). Parameters include a conditional probability of non-masters providing a correct response and a conditional probability of masters providing a correct response. Across all linkage levels, the conditional probability that masters will provide a correct response is generally expected to be high, while it is expected to be low for non-masters. A summary of the operational parameters used to score the 2017–2018 assessment is provided in the following sections.

5.2.1. Probability of Masters Providing Correct Response

When items measuring each linkage level function as expected, students who have mastered the linkage level have a high probability of providing a correct response to items measuring the linkage level. Using the 2018 operational calibration, Figure 5.1 depicts the conditional probability of masters providing a correct response to items measuring each of the 102 linkage levels. Because the point of maximum uncertainty is .5, masters should have a greater than 50% chance of providing a correct response. The results in Figure 5.1 demonstrate that all linkage levels (n = 102, 100.0%) performed as expected.

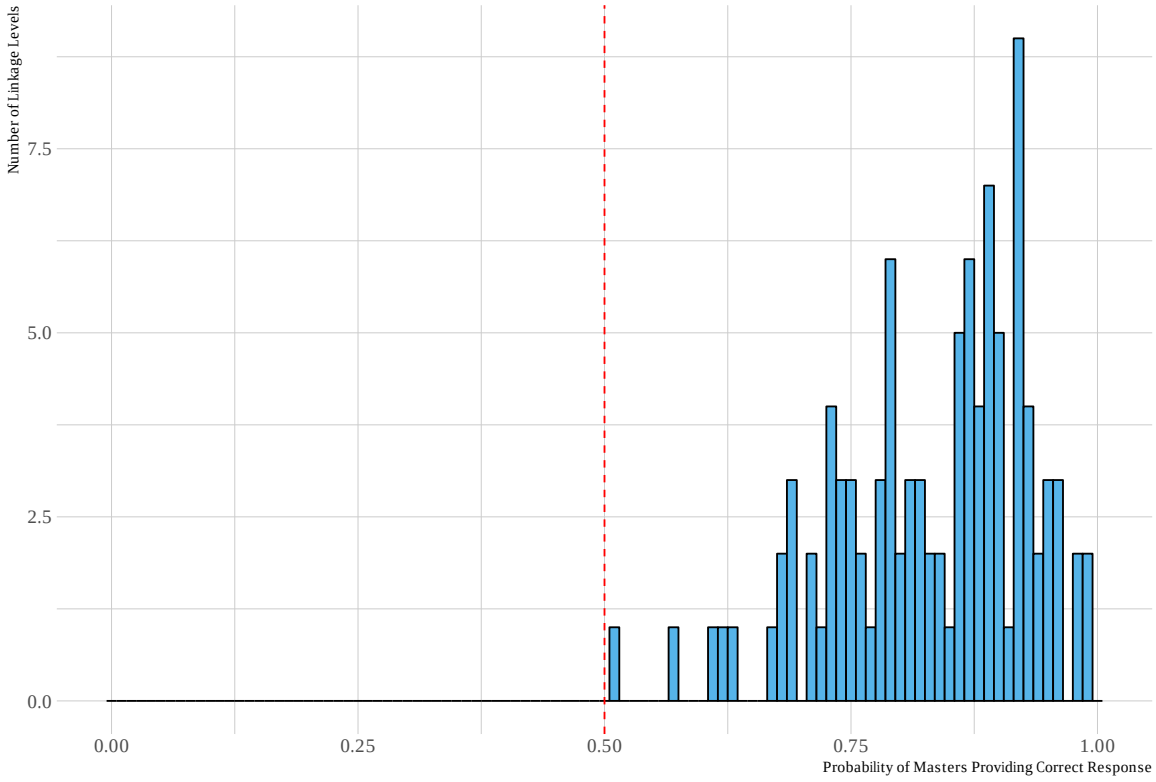


Figure 5.1. Probability of masters providing a correct response to items measuring each linkage level. *Note*: Histogram bins are shown in increments of .01. Reference line indicates .5.

5.2.2. Probability of Non-Masters Providing Correct Response

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

Figure 5.2 summarizes the probability of non-masters providing correct responses to items measuring each of the 102 linkage levels. There is greater variation in the probability of non-masters providing a correct response to items measuring each linkage level than was observed for masters, as shown in Figure 5.2. While most linkage levels (n = 80, 78.4%) performed as expected, non-masters sometimes had a greater than chance (> .5) likelihood of providing a correct response to items measuring the linkage level. This may indicate the items (and linkage level as a whole, since the item parameters are shared) were easily guessable or did not discriminate well between the two groups of students.

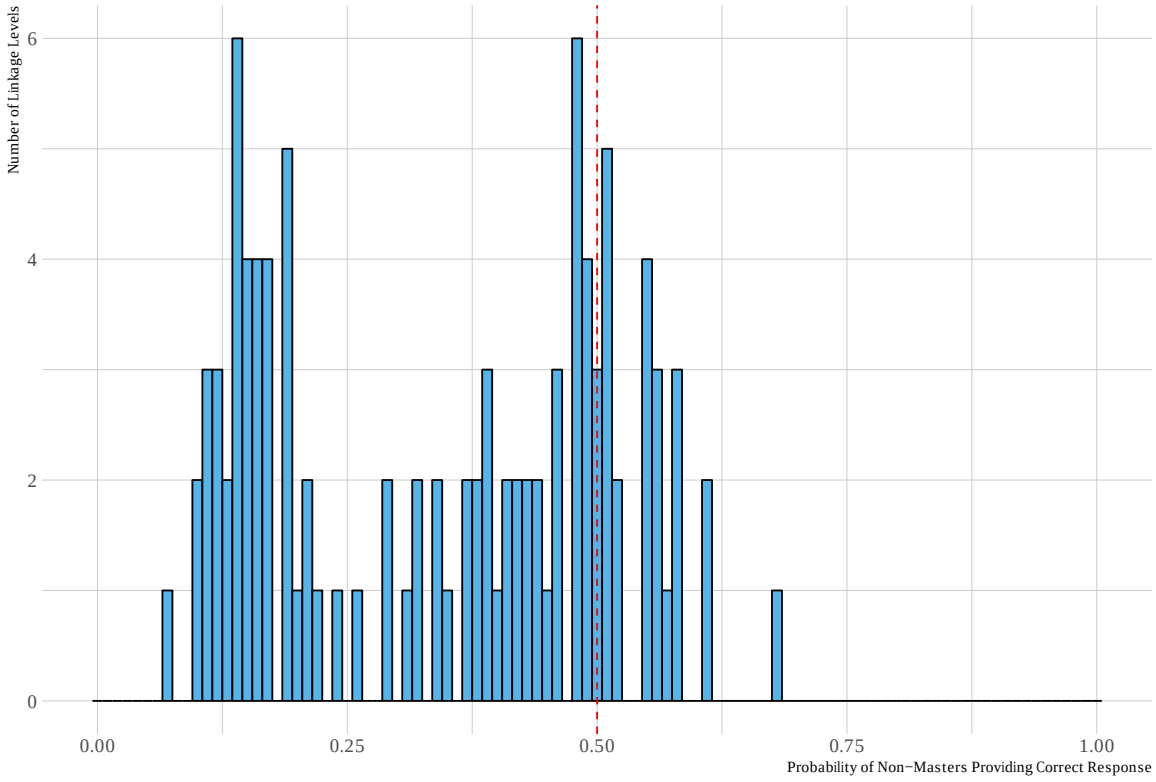


Figure 5.2. Probability of non-masters providing a correct response to items measuring each linkage level. *Note*: Histogram bins are in increments of .01. Reference line indicates .5.

5.3. Mastery Assignment

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

The two-down scoring rule is designed to avoid excessively penalizing students who do not show mastery of their tested linkage levels. This rule is used to assign mastery to untested linkage levels. Take, for example, a student who tested only on the Target linkage level of an EE. If the student demonstrated mastery of the Target linkage level, as defined by the .8 posterior probability of mastery cutoff or the 80% correct rule, then all linkage levels below and including the Target level would be categorized as mastered. If the student did not demonstrate mastery on the tested Target linkage level, then mastery would be assigned at two linkage levels below the tested linkage level (i.e., the Initial level). Theoretical evidence for the use of two-down rule is presented in Chapter 2 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a).

To evaluate the degree to which each mastery assignment rule contributed to students' linkage level mastery status during the 2017–2018 administration of DLM assessments, the percentage of mastery

statuses obtained by each scoring rule was calculated, as shown in Figure 5.3. Posterior probability was given first priority. That is, if multiple scoring rules agreed on the highest linkage level mastered within an EE (e.g., the posterior probability and 80% correct both indicate the Target linkage level as the highest mastered), the mastery status was counted as obtained via the posterior probability. If mastery was not demonstrated by meeting the posterior probability threshold, the 80% scoring rule was imposed, followed by the two-down rule. Approximately 75% to 82% of mastered linkage levels were derived from the posterior probability obtained from the modeling procedure. The next approximately 4% to 7% of linkage levels were assigned mastery status by the percentage correct rule. The remaining approximately 13% to 19% of mastered linkage levels were determined by the minimum mastery, or two-down rule.

Because correct responses to all items measuring the linkage level are often necessary to achieve a posterior probability above the .8 threshold, the percentage correct rule overlapped considerably (but was second in priority) with the posterior probabilities. The percentage correct rule did, however, provide mastery status in those instances where correctly responding to all or most items still resulted in a posterior probability below the mastery threshold. The agreement between these two methods was quantified by examining the rate of agreement between the highest linkage level mastered for each EE for each student. For the 2017–2018 operational year, the rate of agreement between the two methods was 83%. However, in instances where the two methods disagreed, the posterior probability method indicated a higher level of mastery (and was therefore was implemented for scoring) in 67% of cases. Thus, in some instances the posterior probabilities allowed students to demonstrate mastery when the percentage correct was lower than 80% (e.g., a student completed a four-item testlet and answered three of four items correctly).

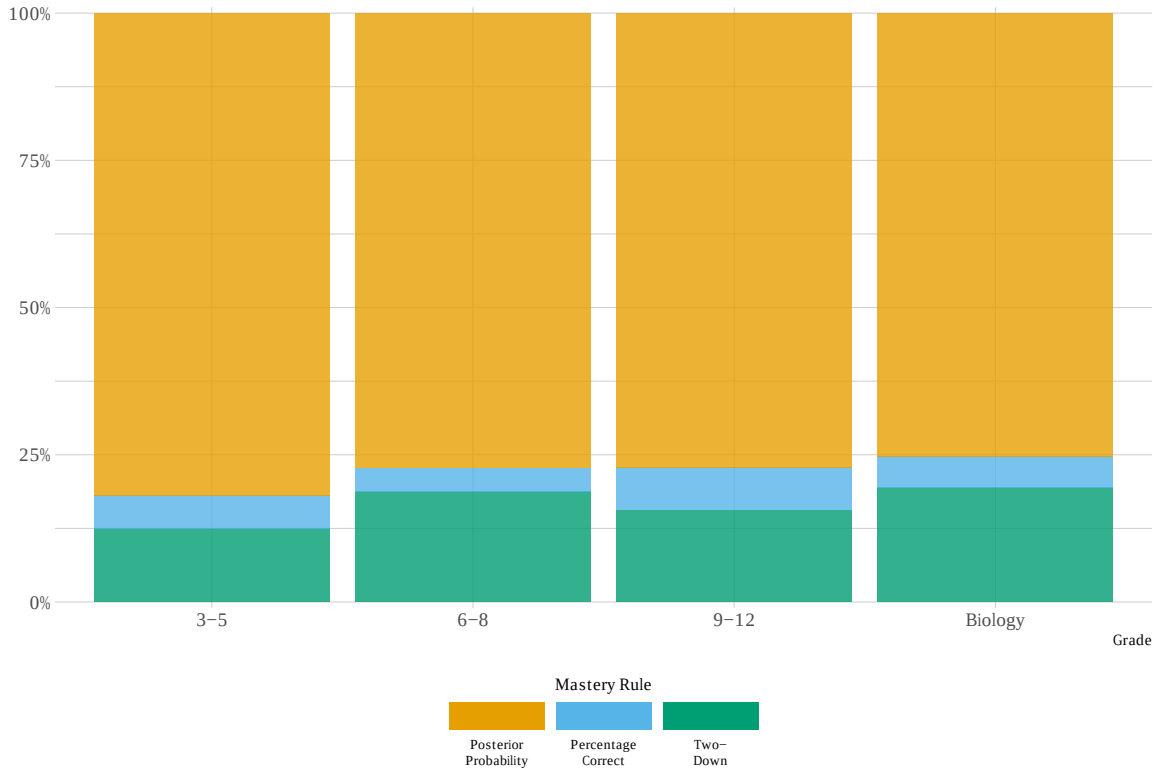


Figure 5.3. Linkage level mastery assignment by mastery rule for each grade band and course.

5.4. Model Fit

Model fit has important implications for the validity of inferences that can be made from assessment results. If the model used to calibrate and score the assessment does not fit the data well, results from the assessment may not accurately reflect what students know and can do. Relative and absolute model fit were compared following the 2017 administration. Model fit research was also prioritized during the 2017–2018 operational year, and frequent feedback was provided by the DLM technical advisory committee (TAC) modeling subcommittee, a subgroup of TAC members focused on reviewing modeling-specific research. During the 2017–2018 year, the modeling subcommittee reviewed research related to Bayesian methods for assessing modeling fit using posterior predictive model checks (Gelman & Hill, 2006; Gelman, Meng, & Stern, 1996) and a newly defined model with partial equivalency of model parameters.

For a complete description of the methods and process used to evaluate model fit, see Chapter 5 of the *2016–2017 Technical Manual Update—Science* (DLM Consortium, 2018a).

5.5. Conclusion

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

6. Standard Setting

The standard setting process for the Dynamic Learning Maps® (DLM®) Alternate Assessment System in science derived cut points for assigning students to four performance levels. For a description of the process, including the development of policy performance level descriptors, the 3-day standard setting meeting, follow-up evaluation of impact data and cut points, and specification of grade-specific performance level descriptors, see Chapter 6 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a).

7. Assessment Results

Chapter 7 of the Dynamic Learning Maps® (DLM®) Alternate Assessment System *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a) describes assessment results for the 2015–2016 academic year, including student participation and performance summaries, and an overview of data files and score reports delivered to state partners. This chapter presents 2017–2018 student participation data; the percentage of students achieving at each performance level; and subgroup performance by gender, race, ethnicity, and English learner (EL) status. This chapter also reports the distribution of students by the highest linkage level mastered during spring 2018. Finally, this chapter describes updates made to score reports and data files during spring 2018. For a complete description of score reports and interpretive guides, see Chapter 7 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a).

7.1. Student Participation

During spring 2018, science assessments were administered to 33,935 students in 14 states and one Bureau of Indian Education (BIE) school. Counts of students tested in each state and BIE are displayed in Table 7.1. The assessments were administered by 14,262 educators in 8,246 schools and 3,104 school districts.

Table 7.1. Student Participation by State (N = 33,935)

State	Students (*n*)
Alaska	239
Delaware	474
Illinois	4,750
Iowa	991
Kansas	1,251
Maryland	2,216
Miccosukee Indian School	9
Missouri	3,180
New Hampshire	360
New Jersey	4,491
New York	9,180
Oklahoma	2,329
Rhode Island	404
West Virginia	734
Wisconsin	3,327

Table 7.2 summarizes the number of students tested in each grade and course. More than 10,000 students participated in each of the elementary (grades 3-5) and the middle school (grades 6-8) grade

bands.[2] In high school (grades 9-12) over 12,300 students participated. The differences in grade-level participation within each band can be traced to differing state-level policies about the grade in which students are assessed.

Table 7.2. Student Participation by Grade or Course (N = 33,935)

Grade	Students (n)
3	145
4	3,711
5	6,642
6	228
7	241
8	10,662
9	3,984
10	1,005
11	6,872
12	276
Biology	169

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

[2]In an effort to increase science instruction beyond the tested grades, several states promoted participation in the science assessment at all grade levels (i.e., did not restrict participation to the grade levels required for accountability purposes). Grade levels 3 and 7 are not tested for accountability purposes in the current DLM science states.

Table 7.3. Demographic Characteristics of Participants (N = 33,935)

Subgroup	*n*	%
Gender		
Male	22,618	66.65
Female	11,315	33.34
Missing	2	0.01
Race		
White	20,418	60.17
African American	7,916	23.33
Two or more races	2,744	8.09
Asian	1,626	4.79
American Indian	954	2.81
Native Hawaiian or Pacific Islander	167	0.49
Alaska Native	89	0.26
Missing	21	0.06
Hispanic ethnicity		
No	27,455	80.90
Yes	6,459	19.03
Missing	21	0.06
English learner (EL) participation		
Not EL eligible or monitored	31,850	93.86
EL eligible or monitored	2,084	6.14
Missing	1	<0.01

In addition to the spring administration, instructionally embedded science assessments are also made available for teachers to administer to students during the year. Results from the instructionally embedded science assessments do not contribute to final summative scoring but can be used to guide instructional decision-making. Table 7.4 summarizes the number of students participating in instructionally embedded testing by state. A total of 3,707 students took at least one instructionally embedded testlet during the 2017–2018 academic year.

Table 7.4. Students Completing Instructionally Embedded Science Testlets by State (N = 3,707)

State	*n*
Delaware	2
Illinois	3
Iowa	741
Kansas	343
Missouri	2,584
New Hampshire	2
New York	7
Oklahoma	24
West Virginia	1

Table 7.5 summarizes the number of instructionally embedded test sessions taken in science. Across all states, students took 28,835 total testlets during the instructionally embedded window.

Table 7.5. Number of Instructionally Embedded Science Test Sessions, by Grade or Course (N = 28,835)

Grade	*n*
3	703
4	629
5	7040
6	1055
7	1005
8	7019
9	1195
10	1157
11	6987
12	2045
Biology	0

7.2. Student Performance

Student performance on DLM assessments is interpreted using cut points, determined during standard setting, which separate student scores into four performance levels. For a full description of the standard-setting process, see Chapter 6 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a). A student receives a performance level based on the total number of linkage levels mastered across the assessed Essential Elements (EEs).

For the spring 2018 administration, student performance was reported using the same four performance levels approved by the DLM Consortium for prior years:

- The student demonstrates Emerging understanding of and ability to apply content knowledge

and skills represented by the EEs.
- The student's understanding of and ability to apply targeted content knowledge and skills represented by the EEs is Approaching the Target.
- The student's understanding of and ability to apply content knowledge and skills represented by the EEs is At Target.
- The student demonstrates Advanced understanding of and ability to apply targeted content knowledge and skills represented by the EEs.

7.2.1. Overall Performance

Table 7.6 reports the percentage of students achieving at each performance level from the spring 2018 administration for science.

The spring 2018 results were fairly consistent with performance in prior years, with the majority of students achieving at either the Emerging or Approaching the Target performance levels. At the elementary level, the percentage of student who achieved at the At Target or Advanced levels ranged from approximately 6% to 23%; in middle school the range was 14% to 23%; and in high school and end-of-instruction biology, the percentages ranged from 10% to 27%.

Table 7.6. Percentage of Students by Grade and Performance Level

Grade	Emerging (%)	Approaching (%)	Target (%)	Advanced (%)	Target+ Advanced (%)
3 (n = 145)	82.1	12.4	4.8	0.7	5.5
4 (n = 3,711)	57.0	19.6	15.4	8.0	23.4
5 (n = 6,642)	64.0	22.1	12.9	0.9	13.8
6 (n = 228)	66.7	18.9	11.4	3.1	14.5
7 (n = 241)	65.6	20.7	13.3	0.4	13.7
8 (n = 10,662)	52.0	24.7	20.5	2.8	23.3
9 (n = 3,984)	47.3	25.7	20.2	6.8	27.0
10 (n = 1,005)	59.0	27.9	10.9	2.2	13.1
11 (n = 6,872)	53.8	27.8	14.5	3.9	18.4
12 (n = 276)	77.9	12.0	8.3	1.8	10.1
Biology (n = 169)	60.9	18.3	17.2	3.6	20.7

7.2.2. Subgroup Performance

Data collection for DLM assessments includes demographic data on gender, race, ethnicity, and EL status. Table 7.7 summarizes the disaggregated frequency distributions for science, collapsed across all assessed grade levels. Although states each have their own rules for minimum student counts needed to support public reporting of results, small counts are not suppressed here because results are aggregated across states, and individual students cannot be identified. Rows labeled Missing indicate the student's demographic data were not entered into the system.

Table 7.7. Students at Each Performance Level, by Demographic Subgroup (N = 33,935)

Subgroup	Emerging		Approaching		Target		Advanced	
	n	%	*n*	%	*n*	%	*n*	%
Gender								
Male	12,370	54.7	5,476	24.2	3,877	17.1	895	4.0
Female	6,466	57.1	2,741	24.2	1,771	15.7	337	3.0
Missing	2	100.0	0	0.0	0	0.0	0	0.0
Race								
White	11,410	55.9	4,938	24.2	3,425	16.8	645	3.2
African American	4,237	53.5	1,984	25.1	1,324	16.7	371	4.7
Two or more races	1,561	56.9	645	23.5	442	16.1	96	3.5
Asian	1,040	64.0	336	20.7	200	12.3	50	3.1
American Indian	444	46.5	247	25.9	207	21.7	56	5.9
Native Hawaiian or Pacific Islander	72	43.1	42	25.1	40	24.0	13	7.8
Alaska Native	62	69.7	23	25.8	4	4.5	0	0.0
Missing	12	57.1	2	9.5	6	28.6	1	4.8
Hispanic ethnicity								
No	15,328	55.8	6,695	24.4	4,485	16.3	947	3.4
Yes	3,497	54.1	1,519	23.5	1,161	18.0	282	4.4
Missing	13	61.9	3	14.3	2	9.5	3	14.3
English learner (EL) participation								
Not EL eligible or monitored	17,861	56.1	7,712	24.2	5,186	16.3	1,091	3.4
EL eligible or monitored	976	46.8	505	24.2	462	22.2	141	6.8
Missing	1	100.0	0	0.0	0	0.0	0	0.0

7.2.3. Linkage Level Mastery

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

Table 7.8 reports the percentage of students who mastered each linkage level as the highest linkage level across all EEs for each grade. For example, across all third-grade EEs, the Initial level was the highest level that students mastered 44% of the time. The percentage of students who mastered as high as the Target linkage level ranged from approximately 20% in grade 12 to 47% in grade 9.

Table 7.8. Students' Highest Linkage Level Mastered Across Science EEs, by Grade

Grade	Linkage Level			
	No evidence (%)	Initial (%)	Precursor (%)	Target (%)
3 (n = 145)	16.6	44.1	14.5	24.8
4 (n = 3,711)	5.7	33.7	16.1	44.4
5 (n = 6,642)	5.3	35.3	17.4	42.0
6 (n = 228)	18.9	22.4	30.7	28.1
7 (n = 241)	12.4	25.7	28.2	33.6
8 (n = 10,662)	4.8	16.0	33.2	46.0
9 (n = 3,984)	5.7	23.5	23.7	47.1
10 (n = 1,005)	8.5	27.7	29.6	34.3
11 (n = 6,872)	6.2	27.6	26.3	39.9
12 (n = 276)	26.8	41.7	11.6	19.9
Biology (n = 169)	1.2	41.4	22.5	34.9

7.3. Data Files

Data files were made available to DLM state partners following the spring 2018 administration. Similar to prior years, the General Research File (GRF) contained student results, including each student's highest linkage level mastered for each EE and final performance level for the subject for all students who completed any testlets. In addition to the GRF, the DLM Consortium delivered several supplemental files. Consistent with prior years, the Special Circumstances File provided information about which students and EEs were affected by extenuating circumstances (e.g., chronic absences), as defined by each state. State partners also received a supplemental file to identify exited students. The exited students file was updated in spring 2018 to include all students who exited at any point during the academic year, rather than only including students who had exited and did not later re-entered the system. Additional demographic fields were also added to this file in order to assist in the matching of students across the multiple return files. In the event of observed incidents during assessment delivery, state partners are provided with an Incident File describing students impacted. Because no incidents were observed during the spring 2018 administration, these files were not delivered.

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

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

7.4. Score Reports

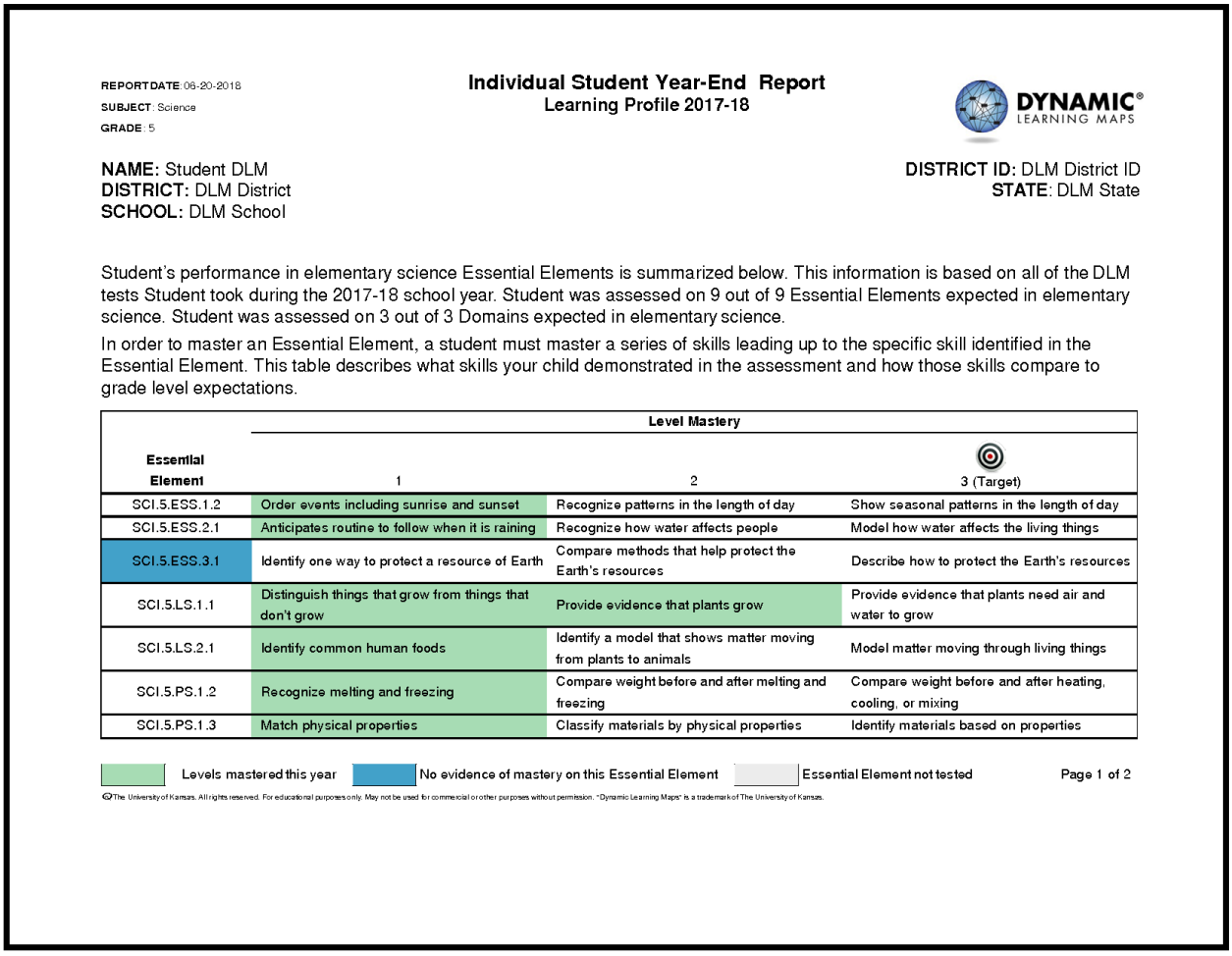
The DLM Consortium provides assessment results to all member states to report to parents/guardians, educators, and state and local education agencies. Individual Student Score Reports summarized student performance on the assessment by subject. Several aggregated reports were provided to state and local education agencies, including reports for the classroom, school, district, and state. No changes were made to the structure of aggregated reports during spring 2018; however, classroom and school reports were generated by the system in Educator Portal following final GRF upload (as the district and state reports were beginning in 2016–2017), rather than being generated outside the system by the score report program. Changes to the Individual Student Score Reports are summarized below. For a complete description of score reports, including aggregated reports, see Chapter 7 of the *2014–2015 Technical Manual—Integrated Model* (DLM Consortium, 2016).

7.4.1. Individual Student Score Reports

During the 2017–2018 year, minor changes were made to the Individual Student Score Reports. On the Learning Profile[3] portion of the report, text description of the shading in the Learning Profile was removed from the narrative to support printing in color or gray scale.

A sample Individual Student Score Report reflecting the 2018 changes is provided in Figure 7.1.

[3]Consistent with prior years, only states that follow the integrated assessment model for DLM English language arts and mathematics receive the Learning Profile in all three subject areas. Year-end states requested this information be omitted for science to be consistent with their ELA and mathematics reports.



REPORT DATE: 06-20-2018
SUBJECT: Science
GRADE: 5

Individual Student Year-End Report
Learning Profile 2017-18

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

DISTRICT ID: DLM District ID
STATE: DLM State

Student's performance in elementary science Essential Elements is summarized below. This information is based on all of the DLM tests Student took during the 2017-18 school year. Student was assessed on 9 out of 9 Essential Elements expected in elementary science. Student was assessed on 3 out of 3 Domains expected in elementary science.

In order to master an Essential Element, a student must master a series of skills leading up to the specific skill identified in the Essential Element. This table describes what skills your child demonstrated in the assessment and how those skills compare to grade level expectations.

	Level Mastery		
Essential Element	1	2	3 (Target)
SCI.5.ESS.1.2	Order events including sunrise and sunset	Recognize patterns in the length of day	Show seasonal patterns in the length of day
SCI.5.ESS.2.1	Anticipates routine to follow when it is raining	Recognize how water affects people	Model how water affects the living things
SCI.5.ESS.3.1	Identify one way to protect a resource of Earth	Compare methods that help protect the Earth's resources	Describe how to protect the Earth's resources
SCI.5.LS.1.1	Distinguish things that grow from things that don't grow	Provide evidence that plants grow	Provide evidence that plants need air and water to grow
SCI.5.LS.2.1	Identify common human foods	Identify a model that shows matter moving from plants to animals	Model matter moving through living things
SCI.5.PS.1.2	Recognize melting and freezing	Compare weight before and after melting and freezing	Compare weight before and after heating, cooling, or mixing
SCI.5.PS.1.3	Match physical properties	Classify materials by physical properties	Identify materials based on properties

Levels mastered this year | No evidence of mastery on this Essential Element | Essential Element not tested

Page 1 of 2

Figure 7.1. Example page of the Learning Profile for spring 2018.

7.5. Quality Control Procedures for Data Files and Score Reports

No changes were made to the manual or automated quality control procedures for spring 2018. For a complete description of quality control procedures, see Chapter 7 in the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a).

7.6. Conclusion

Following the spring 2018 administration, four data files were delivered to state partners. Overall, between 6% and 27% of students achieved at the At Target or Advanced levels across grades, which is consistent with prior years. No incidents were observed during the spring 2018 administration, so an incident file was not needed. Minor changes were made to score reports, including the removal of color-specific narrative text to support printing in grayscale.

8. Reliability

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

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

8.1. Background Information on Reliability Methods

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

Consistent with the levels at which DLM results are reported, this chapter provides results for six types of reliability evidence. For more information on DLM reporting, see Chapter 7 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a). The types of reliability evidence for DLM assessments include (a) classification to overall performance level (performance level reliability); (b) the total number of linkage levels mastered for the subject (subject reliability); (c) the number of linkage levels mastered within each domain (domain reliability); (d) the number of linkage levels mastered within each Essential Element (EE; EE reliability); (e) the classification accuracy of each linkage level within each EE (linkage level reliability); and (f) classification accuracy summarized for the three linkage levels (conditional evidence by linkage level). As described in the next section, reliability evidence comes from simulation studies in which model-specific test data are generated for students with known levels of attribute mastery.

8.2. Methods of Obtaining Reliability Evidence

Standard 2.1: "The range of replications over which reliability/precision is being evaluated should be clearly stated, along with a rationale for the choice of this definition, given the testing situation" (AERA et al., 2014, p. 42).

The simulation used to estimate reliability for DLM versions of scores and classifications considers the unique design and administration of DLM assessments. The use of simulation is necessitated by two factors: the assessment blueprint and the results that classification-based administrations give. Because of the limited number of items students complete to cover the blueprint, students take only

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

8.2.1. Reliability Sampling Procedure

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

1. Draw with replacement the student record of one student from the operational testing data. Use the student's originally scored pattern of linkage level mastery and non-mastery as the true values for the simulated student data.
2. Simulate a new set of item responses to the set of items administered to the student in the operational testlet. Item responses are simulated from calibrated model parameters[4] for the items of the testlet, conditional on the profile of linkage level mastery or non-mastery for the student.
3. Score the simulated item responses using the operational DLM scoring procedure, estimating linkage level mastery or non-mastery for the simulated student. See Chapter 5 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a) for more information.[5]
4. Compare the estimated linkage level mastery or non-mastery to the known values from Step 2 for all linkage levels at which the student was administered items.
5. Repeat Steps 1 through 4 for 2,000,000 simulated students.

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

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

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

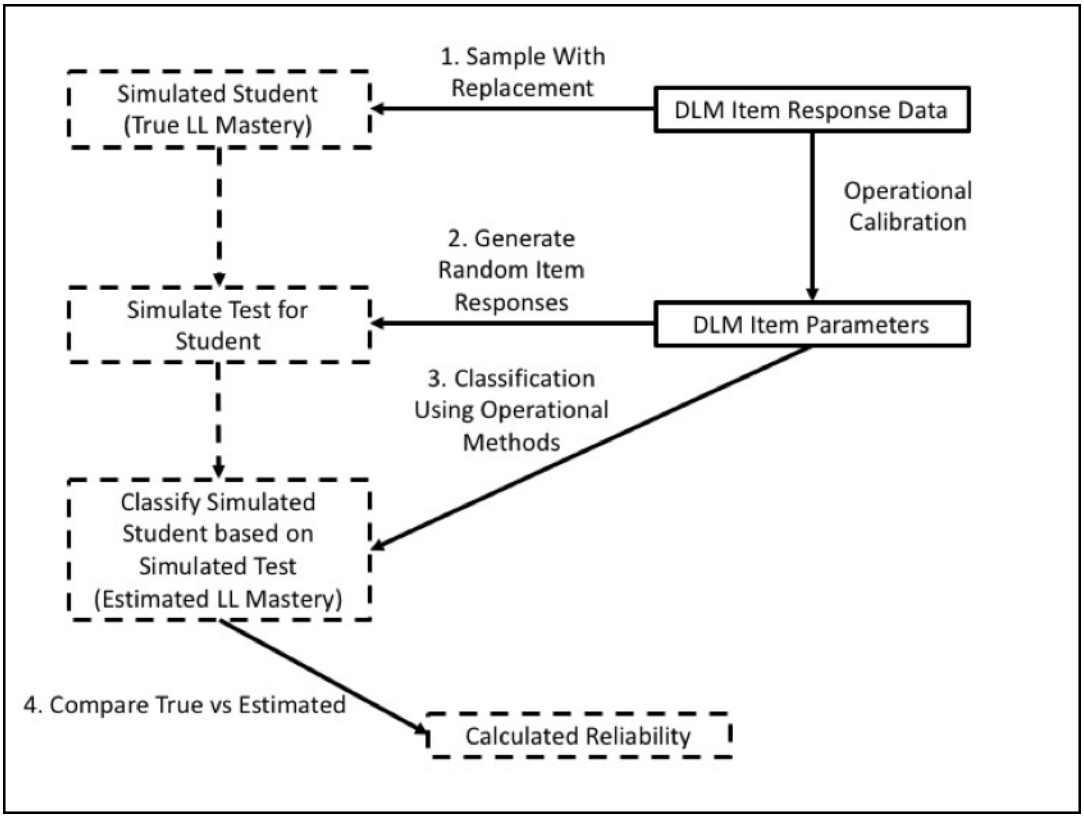


Figure 8.1. Simulation process for creating reliability evidence. *Note*: LL = linkage level.

8.3. Reliability Evidence

Standard 2.2: "The evidence provided for the reliability/precision of the scores should be consistent with the domain of replications associated with the testing procedures, and with the intended interpretations for use of the test scores" (AERA et al., 2014, p. 42).

Standard 2.5: "Reliability estimation procedures should be consistent with the structure of the test" (AERA et al., 2014, p. 43).

Standard 2.12: "If a test is proposed for use in several grades or over a range of ages, and if separate norms are provided for each grade or each age range, reliability/precision data should be provided for each age or grade-level subgroup, not just for all grades or ages combined" (AERA et al., 2014, p. 45).

Standard 2.16: "When a test or combination of measures is used to make classification decisions, estimates should be provided of the percentage of test takers who would be classified in the same way on two [or more] replications of the procedure" (AERA et al., 2014, p. 46).

Standard 2.19: "Each method of quantifying the reliability/precision of scores should be described clearly and expressed in terms of statistics appropriate to the method" (AERA et al., 2014, p. 47).

This chapter provides reliability evidence for six levels of data: (a) performance level reliability, (b)

subject reliability, (c) domain reliability, (d) EE reliability, (e) linkage level reliability, and (f) conditional reliability by linkage level. With 34 EEs, each comprising three linkage levels, the procedure includes 102 analyses to summarize reliability results. Because of the number of analyses, this chapter includes a summary of the reported evidence. An online appendix[6] provides a full report of reliability evidence for all 102 linkage levels and 34 EEs. The full set of evidence is furnished in accordance with Standard 2.12.

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

8.3.1. Performance Level Reliability Evidence

The DLM Consortium reports results using four performance levels. The scoring procedure sums the linkage levels mastered across all tested EEs, and cut points are applied to distinguish between performance categories.

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

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

Table 8.1 presents this information across all grades and subjects. Polychoric correlations between true and estimated performance level range from .93 to .98. Correct classification rates range from .78 to .90 and Cohen's kappa values are between .83 and .91. These results indicate that the DLM scoring procedure of assigning and reporting performance levels based on total linkage levels mastered results in reliable classification of students into performance level categories.

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

Table 8.1. Summary of Performance Level Reliability Evidence

Grade	Polychoric correlation	Correct classification rate	Cohen's kappa
3	.968	.892	.857
4	.964	.806	.885
5	.962	.861	.866
6	.944	.823	.846
7	.945	.838	.846
8	.931	.783	.830
9	.966	.814	.883
10	.962	.854	.869
11	.964	.836	.876
12	.981	.898	.909
Biology	.969	.854	.899

8.3.2. Subject Reliability Evidence

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

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

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

Table 8.2 shows the three summary values for each grade and subject. Classification rate information is provided in accordance with Standard 2.16. The two summary statistics included in Table 8.2 also meet Standard 2.19. The correlation between true and estimated number of linkage levels mastered ranges from .92 to .96. Students' average correct classification rates range from .97 to .99 and average Cohen's kappa values range from .94 to .98. These values indicate the DLM scoring procedure of reporting the number of linkage levels mastered provides reliable results of student performance.

Table 8.2. Summary of Subject Reliability Evidence

Grade	Linkage levels mastered correlation	Average student correct classification	Average student Cohen's kappa
3	.940	.987	.973
4	.949	.977	.950
5	.943	.977	.951
6	.928	.981	.961
7	.934	.979	.960
8	.918	.971	.941
9	.956	.979	.959
10	.945	.983	.966
11	.950	.982	.964
12	.963	.990	.980
Biology	.951	.980	.958

8.3.3. Domain Reliability Evidence

Within the subject of science, students are assessed on EEs in three domains. Because Individual Student Score Reports summarize the number and percentage of linkage levels students mastered for each science domain (see Chapter 7 of this manual for more information), reliability evidence is also provided for each domain.

Domain reliability provides consistency evidence for the number of linkage levels mastered across all EEs in each science domain for each grade. Because domain reporting summarizes the total number of linkage levels a student mastered, the statistics reported for domain reliability are the same as those reported for subject reliability.

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

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

Table 8.3 shows the three summary values for each domain by grade. Values range from .69 to 1.00, indicating that, overall, the DLM method of reporting the total and percentage of linkage levels mastered by domain results in values that can be reliably reproduced.

Table 8.3. Summary of Science Domain Reliability Evidence

Grade	Domain	Linkage levels mastered correlation	Average student correct classification	Average student Cohen's kappa
3	ESS	.769	.996	.994
3	LS	.757	.998	.997
3	PS	.923	.994	.992
4	ESS	.795	.993	.990
4	LS	.700	.997	.996
4	PS	.931	.993	.990
5	ESS	.782	.993	.990
5	LS	.694	.997	.996
5	PS	.929	.993	.990
6	ESS	.744	.994	.992
6	LS	.843	.994	.992
6	PS	.829	.994	.991
7	ESS	.798	.994	.992
7	LS	.844	.994	.992
7	PS	.830	.994	.991
8	ESS	.757	.993	.990
8	LS	.834	.993	.990
8	PS	.822	.993	.991
9	ESS	.853	.994	.991
9	LS	.816	.994	.991
9	PS	.915	.996	.995
10	ESS	.846	.995	.993
10	LS	.797	.995	.993
10	PS	.883	.996	.995
11	ESS	.842	.994	.992
11	LS	.809	.994	.992
11	PS	.904	.996	.995
12	ESS	.861	.996	.995
12	LS	.848	.996	.995
12	PS	.929	.997	.996
Biology	LS1.A	.821	.995	.992
Biology	LS1.B	1.000	.999	.999
Biology	LS2.A	.733	.996	.995
Biology	LS3.B	1.000	.999	.999
Biology	LS4.C	.892	.996	.995

Note: ESS = Earth and space science; LS = life science; PS = physical science.

8.3.4. EE Reliability Evidence

Moving from higher-level aggregation to EEs, the reliability evidence shifts slightly. That is, because EEs are collections of linkage levels with an implied order, EE-level results are reported as the highest

linkage level mastered per EE. Considering subject scores as total scores from an entire test, evidence at the EE level is finer grained than reporting at a subject strand level, which is commonly reported by other testing programs. EEs are specific standards within the subject itself.

Three statistics are used to summarize reliability evidence for EEs:

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

Because there are 34 EEs, the summaries are reported herein according to the number and proportion of EEs that fall within a given range of an index value. Results are given in both tabular and graphical forms. Table 8.4 and Figure 8.2 provide the proportions and the number of EEs, respectively, falling within prespecifed ranges of values for the three reliability summary statistics (i.e., correct classification rate, kappa, correlation). In general, the reliability summaries for number of linkage levels mastered within EEs show strong evidence of reliability.

Table 8.4. Reliability Summaries Across All EEs: Proportion of EEs Falling Within a Specified Index Range

Reliability Index	Index range								
	< .60	**0.60-0.64**	**0.65-0.69**	**0.70-0.74**	**0.75-0.79**	**0.80-0.84**	**0.85-0.89**	**0.90-0.94**	**0.95-1.00**
Polychoric correlation	<.001	<.001	.059	.059	.088	.235	.294	.206	.059
Correct classification rate	<.001	<.001	<.001	<.001	<.001	.206	.618	.176	<.001
Cohen's kappa	.059	.088	.059	.206	.206	.176	.147	.059	<.001

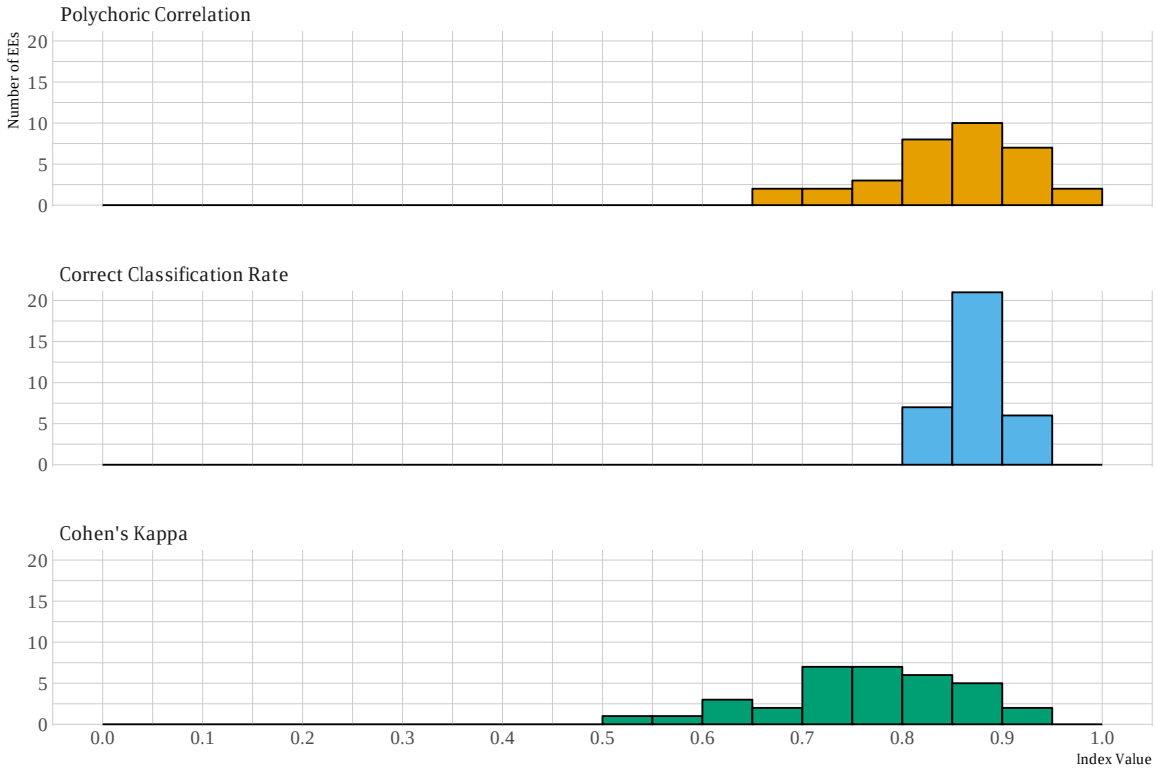


Figure 8.2. Number of linkage levels mastered within EE reliability summaries.

8.3.5. Linkage Level Reliability Evidence

Evidence at the linkage level comes from comparing the true and estimated mastery status for each of the 102 linkage levels in the operational DLM assessment.[7] This level of reliability reporting is even finer grained than the EE level. While it does not have a comparable classical test theory or item response theory analog, its inclusion is important because it is the level at which mastery classifications are made for DLM assessments. All reported summary statistics are based on the resulting contingency tables: the comparison of true and estimated mastery statuses across all simulated examinees. As with any contingency table, a number of summary statistics are possible.

For each statistic, figures are given comparing the results of all 102 linkage levels. Three summary statistics are presented:

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

[7]The linkage level reliability evidence presented here focuses on consistency of measurement given student responses to items. For more information on how students were assigned linkage levels during assessment, see Chapter 4 in the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a).

As there are 102 total linkage levels across all 34 EEs, the summaries reported herein are based on the proportion and number of linkage levels that fall within a given range of an index value. Results are given in both tabular and graphical forms. Table 8.5 and Figure 8.3 provide proportions and number of linkage levels, respectively, that fall within prespecified ranges of values for the three reliability summary statistics (i.e., correct classification rate, kappa, correlation). The kappa value and tetrachoric correlation for one linkage level could not be computed because all students were labeled as masters of the linkage level.

The correlations and correct classification rates show reliability evidence for the classification of mastery at the linkage level. Across all linkage levels, two had tetrachoric correlation values below .6, zero had a correct classification rate below .6, and 13 had a kappa value below 0.6.

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

Reliability Index	Index range								
	< .60	**0.60-0.64**	**0.65-0.69**	**0.70-0.74**	**0.75-0.79**	**0.80-0.84**	**0.85-0.89**	**0.90-0.94**	**0.95-1.00**
Tetrachoric correlation	.020	.010	.010	.020	.020	.020	.108	.186	.608
Correct classification rate	<.001	<.001	<.001	<.001	<.001	.020	.167	.559	.255
Cohen's kappa	.127	.059	.098	.098	.137	.216	.167	.078	.020

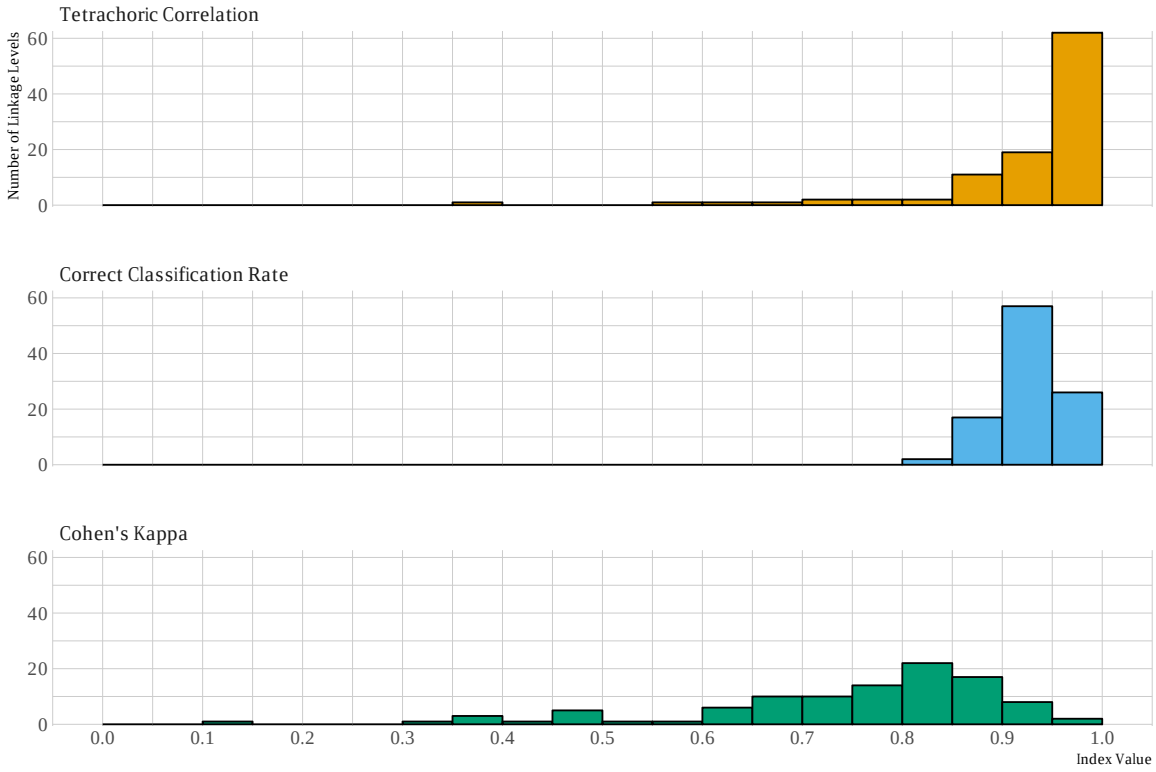


Figure 8.3. Summaries of linkage level reliability.

8.3.6. Conditional Reliability Evidence by Linkage Level

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

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

Figure 8.4 provides the number of linkage levels that fall within prespecified ranges of values for the three reliability summary statistics (i.e., tetrachoric correlation, correct classification rate, kappa). The correlations and correct classification rates generally indicate that all three linkage levels provide reliable classifications of student mastery; results are fairly consistent across all linkage levels for each of the three statistics reported.

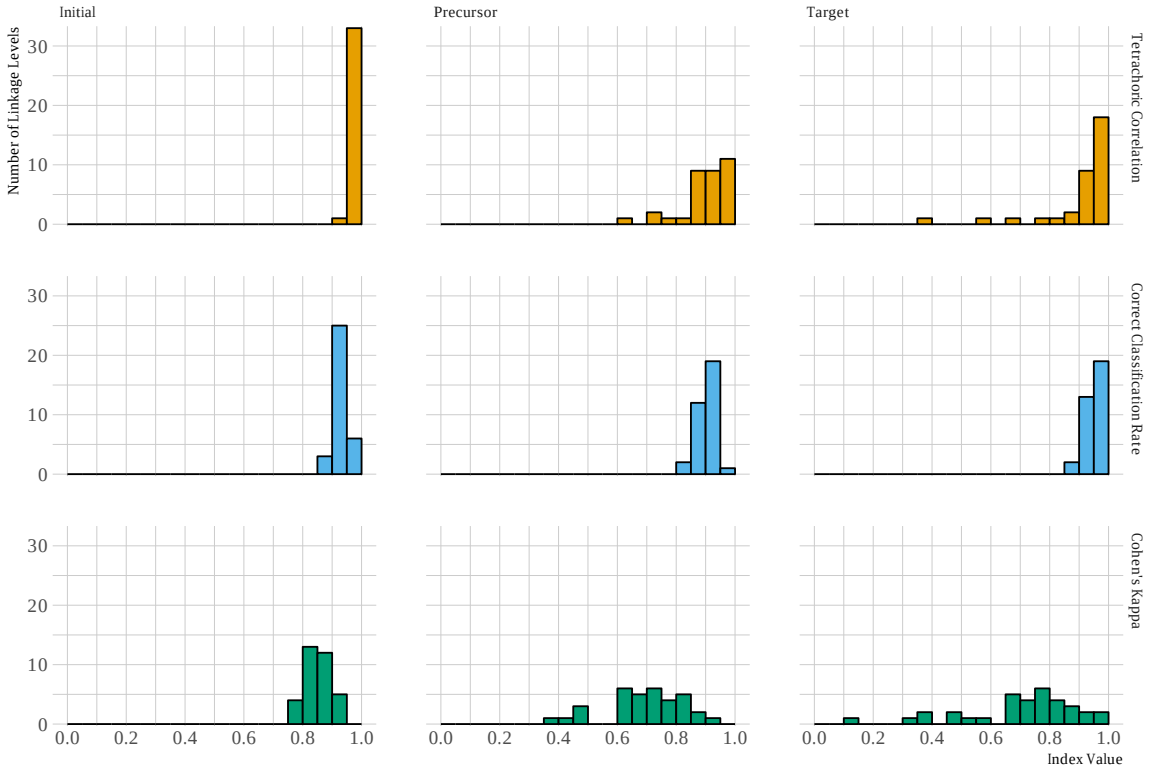


Figure 8.4. Conditional reliability evidence summarized by linkage level.

8.4. Conclusion

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

9. Validity Studies

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

9.1. Evidence Based on Test Content

Evidence based on test content relates to the evidence "obtained from an analysis of the relationship between the content of the test and the construct it is intended to measure" (AERA et al., 2014, p. 14). This section presents results from data collected during 2017–2018 regarding student opportunity to learn the assessed content. For additional evidence based on test content, including the alignment of test content to content standards via the DLM maps (which underlie the assessment system), see Chapter 9 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a).

9.1.1. Opportunity to Learn

After completing administration of the spring 2018 operational assessments, teachers were invited to complete a survey about the assessment (see Chapter 4 of this manual for more information on recruitment and response rates). The survey included three blocks of items. The first and third blocks were fixed forms assigned to all teachers. For the second block, teachers received one randomly assigned section.

The first block of the survey served several purposes.[8] One item provided information about the relationship between students' learning opportunities before testing and the test content (i.e., testlets) they encountered on the assessment. The survey asked teachers to indicate the extent to which they judged test content to align with their instruction across all testlets; Table 9.1 reports the results. Approximately 50% of responses (n = 11,257) reported that most or all science testlets matched instruction. More specific measures of instructional alignment are planned to better understand the extent that content measured by DLM assessments matches students' academic instruction.

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

None		Some (< half)		Most (> half)		All		N/A	
n	%	*n*	%	*n*	%	*n*	%	*n*	%
2,233	9.9	6,904	30.6	7,653	33.9	3,604	16.0	2,170	9.6

The survey also asked teachers to indicate the approximate number of hours they spent instructing students on each of the DLM science core ideas and in the science and engineering practices.

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

Teachers responded using a five-point scale: *none*, *1-10 hours*, *11-20 hours*, *21-30 hours*, or *more than 30 hours*. Table 9.2 and Table 9.3 indicate the amount of instructional time spent on DLM science core ideas and science and engineering practices, respectively. For all science core ideas and science and engineering practices, the most commonly selected response was *1-10 hours*.

Table 9.2. Instructional Time Spent on Science Core Ideas

	Number of hours									
	None		1-10		11-20		21-30		>30	
Core Idea	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Physical Science										
Matter and its interactions	374	19.8	838	44.5	368	19.5	185	9.8	120	6.4
Motion and stability: Forces and interactions	459	24.4	791	42.0	363	19.3	165	8.8	105	5.6
Energy	412	22.1	816	43.7	354	19.0	174	9.3	112	6.0
Life Science										
From molecules to organisms: Structures and processes	516	27.5	745	39.7	330	17.6	165	8.8	119	6.3
Ecosystems: Interactions, energy, and dynamics	347	18.5	769	41.1	383	20.5	234	12.5	139	7.4
Heredity: Inheritance and variation of traits	706	38.1	649	35.0	265	14.3	150	8.1	84	4.5
Biological evolution: Unity and diversity	651	35.0	676	36.3	294	15.8	151	8.1	88	4.7
Earth and Space Science										
Earth's place in the universe	435	23.3	761	40.7	361	19.3	182	9.7	130	7.0
Earth's systems	451	24.1	736	39.3	370	19.8	193	10.3	121	6.5
Earth and human activity	378	20.1	773	41.0	381	20.2	216	11.5	136	7.2

Table 9.3. Instructional Time Spent on Science and Engineering Practices

	Number of hours									
	None		1-10		11-20		21-30		>30	
Science and engineering practice	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Developing and using models	473	25.1	868	46.0	292	15.5	152	8.1	102	5.4
Planning and carrying out investigations	365	19.4	836	44.5	368	19.6	189	10.1	119	6.3
Analyzing and interpreting data	327	17.4	809	43.0	398	21.2	196	10.4	150	8.0
Using mathematics and computational thinking	335	17.8	741	39.4	357	19.0	207	11.0	239	12.7
Constructing explanations and designing solutions	538	28.7	747	39.9	330	17.6	161	8.6	97	5.2
Engaging in argument from evidence	648	34.6	722	38.5	284	15.1	135	7.2	86	4.6
Obtaining, evaluating, and communicating information	369	19.6	768	40.8	361	19.2	217	11.5	169	9.0

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

Table 9.4 summarizes the Spearman rank-order correlations between domain instructional time and linkage levels mastered in the domain. Correlations ranged from .13 to .16. Based on guidelines from Cohen (1988), the observed correlations were small.

Table 9.4. Correlation Between Instuction Time in Science Domain and Linkage Levels Mastered

Domain	Correlation with instructional time
Physical science	0.13
Life science	0.16
Earth and space science	0.15

9.2. Evidence Based on Response Processes

The study of test takers' response processes provides evidence about the fit between the test construct and the nature of how students actually experience test content (AERA et al., 2014). The validity studies presented in this section include teacher survey data collected in spring 2018 regarding students' ability to respond to testlets and test administration observation data collected during 2017–2018. For additional evidence based on response process, including studies on student and teacher behaviors during testlet administration and evidence of fidelity of administration, see Chapter 9 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a).

9.2.1. Evaluation of Test Administration

After administering spring operational assessments in 2018, teachers provided feedback via a teacher survey. Survey data that inform evaluations of assumptions regarding response processes include teacher perceptions of students' ability to respond as intended, free of barriers, and with necessary supports available.[9]

One of the fixed-form sections of the spring 2018 teacher survey included three items about students' ability to respond. Teachers were asked to use a four-point scale (*strongly disagree, disagree, agree,* or *strongly agree*). Results were combined in the summary presented in Table 9.5. The majority of teachers (more than 84%) agreed or strongly agreed that their students (a) responded to items to the best of their knowledge and ability; (b) were able to respond regardless of disability, behavior, or health concerns; and (c) had access to all supports necessary to participate. These results are similar to those observed in previous years and suggest that students are able to effectively interact with and respond to the assessment content.

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

Table 9.5. Teacher Perceptions of Student Experience With Testlets

Statement	SD		D		A		SA		A+SA	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
The student responded to items to the best of their knowledge and ability.	896	3.9	1,501	6.6	12,192	53.5	8,193	36.0	20,385	89.5
The student was able to respond regardless of disability, behavior, or health concerns.	1,601	7.0	2,026	8.9	12,347	54.1	6,865	30.1	19,212	84.2
The student had access to all supports necessary to participate.	635	2.8	738	3.2	11,840	51.8	9,624	42.1	21,464	93.9

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

9.2.2. Test Administration Observations

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

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

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

While all consortium states are encouraged to submit test administration observations, these observations are optional. Test administration observations were received from three states during

the 2017–2018 academic year. A total of 28 test administration observations were collected. Of those, 8 (28.6%) were of computer-delivered assessments and 20 (71.4%) were of teacher-administered testlets.

To investigate the assumptions that underlie the claims of the validity argument, several parts of the test administration observation protocol were designed to provide information corresponding to the assumptions. One assumption addressed is that educators allow students to engage with the system as independently as they are able. For computer-delivered testlets, related evidence is summarized in Table 9.6; behaviors were identified as supporting, neutral, or nonsupporting. For example, clarifying directions (25% of observations) removes student confusion about the task demands as a source of construct-irrelevant variance and supports the student's meaningful, construct-related engagement with the item. In contrast, using physical prompts (e.g., hand-over-hand guidance) indicates that the teacher directly influenced the student's answer choice. Overall, 60% of observed behaviors were classified as supporting, with no observed behaviors reflecting nonsupporting actions.

Table 9.6. Test Administrator Actions During Computer-Delivered Testlets (n = 8)

Action	*n*	%
Supporting		
Navigated one or more screens for the student	5	62.5
Read one or more screens aloud to the student	5	62.5
Repeated question(s) before student responded	3	37.5
Clarified directions or expectations for the student	2	25.0
Neutral		
Used pointing or gestures to direct student attention or engagement	4	50.0
Used verbal prompts to direct the student's attention or engagement (e.g., "look at this")	4	50.0
Asked the student to clarify or confirm one or more responses	1	12.5
Used materials or manipulatives during the administration process	1	12.5
Allowed student to take a break during the testlet	0	0.0
Repeated question(s) after student responded (i.e., gave a second trial at the same item)	0	0.0
Nonsupporting		
Physically guided the student's hand to an answer choice	0	0.0
Reduced the number of answer choices available to the student	0	0.0

Note: Respondents could select multiple responses to this question.

For DLM assessments, interaction with the system includes interaction with the assessment content as well as physical access to the testing device and platform. The fact that educators navigated one or more screens in 62% of the observations does not necessarily indicate the student was prevented from engaging with the assessment content as independently as possible. Depending on the student, test administrator navigation may either support or minimize students' independent, physical interaction

with the assessment system. While not the same as interfering with students' interaction with the content of assessment, navigating for students who are able to do so independently conflicts with the assumption that students are able to interact with the system as intended. The observation protocol did not capture why the test administrator chose to navigate, and the reason was not always obvious.

A related assumption is that students are able to interact with the system as intended. Evidence for this assumption was gathered by observing students taking computer-delivered testlets, as shown in Table 9.7. Independent response selection was observed in 75% of the cases. Non-independent response selection may include allowable practices, such as test administrators entering responses for the student. The use of materials outside of KITE Client was seen in 12% of the observations. Verbal prompts for navigation and response selection are strategies within the realm of allowable flexibility during test administration. These strategies, which are commonly used during direct instruction for students with the most significant cognitive disabilities, are used to maximize student engagement with the system and promote the type of student-item interaction needed for a construct-relevant response. However, they also indicate that students were not able to sustain independent interaction with the system throughout the entire testlet.

Table 9.7. Student Actions During Computer-Delivered Testlets (n = 8)

Action	*n*	%
Selected answers independently	6	75.0
Navigated screens independently	4	50.0
Navigated screens after verbal prompts	3	37.5
Selected answers after verbal prompts	3	37.5
Navigated screens after test administrator pointed or gestured	2	25.0
Skipped one or more items	1	12.5
Used materials outside of KITE student portal to indicate responses to testlet items	1	12.5
Independently revisited a question after answering it	0	0.0
Revisited one or more questions after verbal prompt(s)	0	0.0

Note: Respondents could select multiple responses to this question.

Another assumption in the validity argument is that students are able to respond to tasks irrespective of sensory, mobility, health, communication, or behavioral constraints. This assumption was evaluated by having observers note whether there was difficulty with accessibility supports (including lack of appropriate available supports) during observations of teacher-administered testlets. Of the 20 observations of teacher-administered testlets, observers noted difficulty in zero cases (0%). For computer-delivered testlets, evidence to evaluate the assumption was collected by noting students indicating responses to items using varied response modes such as eye gaze (0%) and using manipulatives or materials outside of KITE (12%). Additional evidence for this assumption was gathered by observing whether students were able to complete testlets. Of the 28 test administration observations collected, students completed the testlet in 28 cases (100%).

Another assumption underlying the validity argument is that test administrators enter student responses with fidelity. To record student responses with fidelity, test administrators needed to observe multiple modes of communication, such as verbal, gesture, and eye gaze. Table 9.8

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

Table 9.8. Primary Response Mode for Teacher-Administered Testlets (n = 20)

Response mode	*n*	%
Gestured to indicate response to test administrator who selected answers	10	50.0
Verbally indicated response to test administrator who selected answers	9	45.0
Used computer/device to respond independently	5	25.0
Eye-gaze system indication to test administrator who selected answers	0	0.0
Used switch system to respond independently	0	0.0
No response	0	0.0

Note: Respondents could select multiple responses to this question.

Computer-delivered testlets provided another opportunity to confirm fidelity of response entry when test administrators entered responses on behalf of students. This support is recorded on the Personal Needs and Preferences Profile and is recommended for a variety of situations (e.g., students who have limited motor skills and cannot interact directly with the testing device even though they can cognitively interact with the onscreen content). Observers recorded whether the response entered by the test administrator matched the student's response. In two of eight (25%) observations of computer-delivered testlets, the test administrator entered responses on the student's behalf. In all of those cases, observers indicated that the entered response matched the student's response.

9.3. Evidence Based on Internal Structure

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

9.3.1. Evaluation of Item-Level Bias

Differential item functioning (DIF) addresses the challenge created when some test items are "asked in such a way that certain groups of examinees who are knowledgeable about the intended concepts are prevented from showing what they know" (Camilli & Shepard, 1994, p. 1). DIF analyses can uncover internal inconsistency if particular items function differently in a systematic way for identifiable subgroups of students (AERA et al., 2014). While identification of DIF does not always indicate weakness in a test item, it can point to construct-irrelevant variance or unexpected multidimensionality, posing considerations for validity and fairness.

9.3.1.1. Method

DIF analyses for 2018 followed the same procedure used in previous years, including data from 2015–2016 through 2016–2017 to flag items for evidence of DIF. Items were selected for inclusion in the DIF analyses based on minimum sample-size requirements for the two gender subgroups: male and female. Within the DLM population, the number of female students responding to items is smaller than the number of male students by a ratio of approximately 1:2; therefore, a threshold for item inclusion was retained from previous years whereby the female group must have at least 100 students responding to the item. The threshold of 100 was selected to balance the need for a sufficient sample size in the focal group with the relatively low number of students responding to many DLM items.

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

Using the above criteria for inclusion, 406 (72%) items on science testlets were selected. In total, 107 were evaluated in the elementary school grade band, 126 items in the middle school grade band, 119 items in the high school grade band, and 54 items in the biology end-of-instruction assessment. Item sample sizes ranged from 275 to 6,019.

Of the 158 items that were not included in the DIF analysis, 156 (98.7%) had a focal group sample size of less than 100 and 2 (1.3%) had an item *p*-value greater than .95. Table 9.9 shows the number and percent of items that failed each inclusion criteria, broken down by the linkage level the items assess. The majority of non-included items come from the Precursor linkage level and are excluded due to insufficient sample size of the focal group.

Table 9.9. Items Not Included in DIF Analysis, by Subject and Linkage Level

	Sample Size		Item Proportion Correct		Subgroup Proportion Correct	
Subject and Linkage Level	***n***	**%**	***n***	**%**	***n***	**%**
Initial	12	7.7	0	0.0	0	0.0
Precursor	144	92.3	0	0.0	0	0.0
Target	0	0.0	2	100.0	0	0.0

For each item, logistic regression was used to predict the probability of a correct response, given group membership and performance in the subject. Specifically, the logistic regression equation for each item included a matching variable comprised of the student's total linkage levels mastered in the subject of the item and a group membership variable, with females coded 0 as the focal group and males coded 1 as the reference group. An interaction term was included to evaluate whether nonuniform DIF was present for each item (Swaminathan & Rogers, 1990); the presence of non-uniform DIF indicates that the item functions differently because of the interaction between total linkage levels mastered and gender. When non-uniform DIF is present, the gender group with the highest probability of a correct response to the item differs along the range of total linkage levels

mastered, thus one group is favored at the low end of the spectrum and the other group is favored at the high end.

Three logistic regression models were fitted for each item:

$$\text{M}_0\text{: logit}(\pi_i) = \beta_0 + \beta_1 \text{X} \tag{9.1}$$
$$\text{M}_1\text{: logit}(\pi_i) = \beta_0 + \beta_1 \text{X} + \beta_2 G \tag{9.2}$$
$$\text{M}_2\text{: logit}(\pi_i) = \beta_0 + \beta_1 \text{X} + \beta_2 G + \beta_3 \text{X} G; \tag{9.3}$$

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

Because of the number of items evaluated for DIF, Type I error rates were susceptible to inflation. The incorporation of an effect-size measure can be used to distinguish practical significance from statistical significance by providing a metric of the magnitude of the effect of adding gender and interaction terms to the regression model.

For each item, the change in the Nagelkerke pseudo R^2 measure of effect size was captured, from M_0 to M_1 or M_2, to account for the effect of the addition of the group and interaction terms to the equation. All effect-size values were reported using both the Zumbo and Thomas (1997) and Jodoin and Gierl (2001) indices for reflecting a negligible, moderate, or large effect. The Zumbo and Thomas thresholds for classifying DIF effect size are based on Cohen's (1992) guidelines for identifying a small, medium, or large effect. The thresholds for each level are .13 and .26; values less than .13 have a negligible effect, values between .13 and .26 have a moderate effect, and values of .26 or greater have a large effect.

The Jodoin and Gierl approach expanded on the Zumbo and Thomas effect-size classification by basing the effect-size thresholds for the simultaneous item bias test procedure (Li & Stout, 1996), which, like logistic regression, also allows for the detection of both uniform and nonuniform DIF and uses classification guidelines based on the widely accepted ETS Mantel-Haenszel classification guidelines. The Jodoin and Gierl threshold values for distinguishing negligible, moderate, and large DIF are more stringent than those of the Zumbo and Thomas approach, with lower threshold values of .035 and .07 to distinguish between negligible, moderate, and large effects. Similar to the ETS Mantel-Haenszel method, negligible effect is denoted with an A, moderate effect with a B, and large effect with a C.

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

9.3.1.2. Results

9.3.1.2.1. Uniform DIF Model

A total of 40 items were flagged for evidence of uniform DIF when comparing M_0 to M_1. Table 9.10 summarizes the total number of items flagged for evidence of uniform DIF by grade band for each model. The percentage of items flagged for uniform DIF ranged from 4% to 13%.

Table 9.10. Items Flagged for Evidence of Uniform Differential Item Functioning

Grade Band or Course	Items flagged (*n*)	Total items (*N*)	Items flagged (%)	Items with moderate or large effect size (*n*)
Elementary	10	107	9.3	0
Middle	12	126	9.5	0
High	16	119	13.4	0
Biology	2	54	3.7	0

Using the Zumbo and Thomas (1997) effect-size classification criteria, all items were found to have a negligible effect-size change after the gender term was added to the regression equation. Similarly, using the Jodoin and Gierl (2001) effect-size classification criteria, all items were found to have a negligible effect-size change after the gender term was added to the regression equation.

9.3.1.2.2. Combined Model

A total of 57 items were flagged for evidence of DIF when both the gender and interaction terms were included in the regression equation, as shown in equation (9.3). Table 9.11 summarizes the number of items flagged by grade band or course. The percentage of items flagged for each grade band or course ranged from 12% to 18%.

Table 9.11. Items Flagged for Evidence of Differential Item Functioning for the Combined Model

Grade Band or Course	Items flagged (*n*)	Total items (*N*)	Items flagged (%)	Items with moderate or large effect size (*n*)
Elementary	19	107	17.8	0
Middle	15	126	11.9	1
High	16	119	13.4	0
Biology	7	54	13.0	0

Using the Zumbo and Thomas (1997) effect-size classification criteria, all items had a negligible change in effect size after adding the gender and interaction terms to the regression equation.

Using the Jodoin and Gierl (2001) effect-size classification criteria, zero items had a moderate change in effect size, one had a large change in effect size, and the remaining 56 items were found to have a negligible change in effect size after adding the gender and interaction terms to the regression equation. Information about the flagged items with a non-negligible change in effect size is summarized in Table 9.12. The one flagged item favored the female group at higher levels of ability and males at lower levels of ability (as indicated by a positive $\beta_3 XG$). Appendix A includes a plot that displays the best-fitting regression line for each gender group, with jitter plots representing the total linkage levels mastered for individuals in each gender group for the one science item with a non-negligible effect-size change in the combined model.

Table 9.12. Items Flagged for Differential Item Functioning With Moderate or Large Effect Size for the Combined Model

Item ID	Grade Band	EE	χ^2	p-value	$\beta_2 G$	R^2	$\beta_3 XG$	Z&T*	J&G*
51584	Middle	MS.LS.1.5	9.00	<.01	-0.36	0.07	.89	C	C

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

9.3.1.3. Test Development Team Review of Flagged Items

The science test development team was provided with a data file that contained information about the item flagged with a large effect size. To avoid biasing the review of the item, the file did not indicate which group was favored.

During their review of the flagged item, the test development team was asked to consider facets of the item that may lead one gender group to provide correct responses at a higher rate than the other. Because DIF is closely related to issues of fairness, the bias and sensitivity external review criteria (see Clark, Beitling, Bell, & Karvonen, 2016) were provided for the test development team to consider as they reviewed the items. After reviewing a flagged item and considering its context in the testlet, including the engagement activity, the test development team was asked to provide one of three decision codes.

1. Accept: There is no evidence of bias favoring one group or the other. Leave item as is.
2. Minor revision: There is a clear indication that a fix will correct the item if the edit can be made within the allowable edit guidelines.
3. Reject: There is evidence the item favors one gender group over the other. There is no allowable edit to correct the issue. The item is slated for retirement.

After review, the item flagged with a large effect size was given a decision code of 1 by the test development team. No evidence could be found in the item indicating the content favored one gender group over the other.

As additional data are collected in subsequent operational years, the scope of DIF analyses will be expanded to include additional items, subgroups, and approaches to detecting DIF.

9.3.2. Internal Structure Within Linkage Levels

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

Chapter 5 of this manual includes a summary of the parameters used to score the assessment, which includes the probability of a master providing a correct response to items measuring the linkage level and the probability of a non-master providing a correct response to items measuring the linkage level. Because a fungible model is used for scoring, these parameters are the same for all items measuring the linkage level.

When linkage levels perform as expected, masters should have a high probability of providing a correct response, and non-masters should have a low probability of providing a correct response. As indicated in Chapter 5 of this manual, for 102 (100.0%) linkage levels, masters had a greater than .5 chance of providing a correct response to items. Similarly, for 80 (78.4%) linkage levels, non-masters had a less than .5 chance of providing a correct response to items.

Chapter 3 of this manual includes additional evidence of internal consistency in the form of standardized difference figures. Standardized difference values are calculated to indicate how far from the linkage level mean each item's *p*-value falls. Across all linkage levels, 592 (99.2%) of items fell within two standard deviations of the mean for the linkage level.

These sources, combined with procedural evidence for developing fungible testlets at the linkage level, provide evidence of the consistency of measurement at the linkage levels. For more information on the development of fungible testlets, see the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a). In instances where linkage levels and the items measuring them do not perform as expected, test development teams review flags to ensure the content measures the construct as expected.

9.4. Evidence Based on Consequences of Testing

Validity evidence must include the evaluation of the overall soundness of proposed interpretations of test scores for their intended uses (AERA et al., 2014, p. 19). To establish sound score interpretations, the assessment must measure important content that informs instructional choices and goal setting.

Consistent with previous years, one source of evidence was collected in spring 2018 via teacher survey responses regarding teacher perceptions of assessment content. An additional study was conducted to evaluate teachers' use of report contents for instructional planning and decision making.

9.4.1. Teacher Perception of Assessment Content

On the spring 2018 survey,[10] teachers were asked three questions about their perceptions of assessment content: whether the content measured important academic skills and knowledge,

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

whether the content reflected high expectations, and whether the testlet activities were similar to instructional activities in the classroom. Table 9.13 summarizes their responses. Teachers generally agreed or strongly agreed that content reflected high expectations for their students (84%), measured important academic skills (72%), and was similar to instructional activities used in the classroom (70%).

While the majority of teachers agreed with these statements, 16%-30% disagreed. DLM assessments represent a departure from the breadth of academic skills assessed by many states' previous alternate assessments. Given the short history of general curriculum access for this population and the tendency to prioritize the instruction of functional academic skills (Karvonen, Wakeman, Browder, Rogers, & Flowers, 2011), teachers' responses may reflect awareness that DLM assessments contain challenging content. However, teachers were divided on its importance in the educational programs of students with the most significant cognitive disabilities.

Table 9.13. Teacher Perceptions of Assessment Content

	Strongly Disagree		Disagree		Agree		Strongly Agree		Agree + Strongly Agree	
Statement	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Content measured important academic skills and knowledge for this student.	2,552	11.1	3,835	16.7	13,432	58.6	3,117	13.6	16,549	72.2
Content reflected high expectations for this student.	1,187	5.2	2,406	10.6	13,916	61.0	5,286	23.2	19,202	84.2
Activities in testlets were similar to instructional activities used in the classroom.	2,062	9.0	4,833	21.2	12,836	56.3	3,060	13.4	15,896	69.7

9.4.2. Use of Reports for Instruction

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

[11] Individual student score reports include a Performance Profile, which summarizes overall performance in the subject. States who either participate only in DLM science assessments or who also participate in the integrated model for English language arts and mathematics, also receive a Learning Profile, which summarizes specific skills mastered by EE.

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

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

Interested teachers were asked to complete a Qualtrics survey listing their background information and responding to the three eligibility questions. A total of 135 teachers responded to the survey. Of those, 40 responded *yes* to all three eligibility questions and were contacted to set up a time to participate. Of those contacted, 17 participated in the virtual meetings. Because of attrition challenges between scheduling and conducting phone calls, the number of participants per call ranged from one to five. This resulted in several focus groups being conducted as one-on-one interviews; they are collectively referred to as focus groups throughout this section. While the views described in this section are based on a limited sample of teachers, their feedback provides some evidence as to how reports are used in the subsequent academic year. Additional consequential evidence is planned (see Chapter 11 of this manual).

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

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

9.4.2.1. Receiving Reports

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

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

9.4.2.2. Using Reports to Inform Instruction

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

Teachers who received a Learning Profile described their processes for using fine-grained results to create instructional plans in the subsequent academic year. They described evaluating the skills mastered in the prior grade, as shown on the 2017 score report, and comparing those to skills available in the current 2017–2018 grade's content standards. Prioritization varied based on student needs. For some students, teachers described focusing less on skills that had already been mastered to provide greater breadth of instruction and assessment; for others, they described prioritizing the next level of skill acquisition within a similar standard to provide greater depth of instruction and assessment.

Teachers also described using the Learning Profile section of reports to inform IEP goals. As one teacher stated, "Their IEP goals are very similar to their linkage level [statement]. I can say, 'Hey, let's look at this linkage level and let's look at this target skill and this is what we're working on in your IEP.' It's real easy for me to tie all these things together so we don't have this weird zigzag of skills [across the Learning Profile]. [It's] more streamlined and better growth." She went on to say, "I really feel like this holds kids to a higher standard. I think it keeps teachers from writing cop-out goals."

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

9.4.2.3. Talking With Parents

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

For parents of students new to the DLM assessment system, teachers reported some confusion about the reports. "Parents seemed a little confused because they had never seen a report before. So I don't think they really knew exactly what they were looking at since it was something so new presented to

them." The teacher went on to share, "We just went over exactly what was on the report step by step. I pointed out some of the IEP objectives and how they were related to what was on the report."

Most teachers reported that while their district shared a copy of the report to give to parents, they were not provided with the DLM Parent Interpretive Guide to accompany the report, and teachers were not aware it existed.

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

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

9.5. Conclusion

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

The final chapter of this manual, Chapter 11, references evidence presented through the technical manual, including Chapter 9, and expands the discussion of the overall validity argument. Chapter 11 also provides areas for further inquiry and ongoing evaluation of the DLM Alternate Assessment System, building on the evidence presented in the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a) and the subsequent annual technical manual update (DLM Consortium, 2018a), in support of the assessment's validity argument.

10. Training and Instructional Activities

Chapter 10 of the Dynamic Learning Maps® (DLM®) Alternate Assessment System *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a) describes the training offered in 2015–2016 to state and local education agency staff, the required test administrator training, the optional science module for test administrators, and the optional science instructional activities. No changes were made to training or optional science resources in 2017–2018.

11. Conclusion and Discussion

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

The DLM system completed its third operational administration year in 2017–2018. This technical manual update provides updated evidence from the 2017–2018 year intended to evaluate the propositions and assumptions that undergird the assessment system as described at the onset of its design in the DLM theory of action. The contents of this manual address the information summarized in Table 11.1. Evidence summarized in this manual builds on the original evidence included in the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a) and in the subsequent year (DLM Consortium, 2018a). Together, the documents summarize the validity evidence collected to date.

Table 11.1. Review of Technical Manual Update Contents

Chapter	Contents
1	Provides an overview of information updated for the 2017–2018 year
2	Not updated for 2017–2018
3, 4	Provides procedural evidence collected during 2017–2018 of test content development and administration, including field-test information, and teacher-survey results
5	Describes the statistical model used to produce results based on student responses, along with a summary of item parameters
6	Not updated for 2017–2018
7, 8	Describes results and analyses from the third operational administration, evaluating how students performed on the assessment, the distributions of those results, including aggregated and disaggregated results, and analysis of the consistency of student responses
9	Provides additional studies from 2017–2018 focused on specific topics related to validity and to evaluate the score propositions and intended uses
10	Not updated for 2017–2018

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

11.1. Validity Evidence Summary

The accumulated evidence available by the end of the 2017–2018 year provides additional support for the validity argument. Four interpretation and use claims are summarized in Table 11.2. Each claim is addressed by evidence in one or more of the sources of validity evidence defined in the *Standards for*

Educational and Psychological Testing (AERA et al., 2014). While many sources of evidence contribute to multiple propositions, Table 11.2 lists the primary associations. For example, Proposition 4 is indirectly supported by content-related evidence described for Propositions 1 through 3. Table 11.3 shows the titles and sections for the chapters cited in Table 11.2.

Table 11.2. DLM Alternate Assessment System Claims and Sources of Updated Evidence for 2017–2018

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

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

Table 11.3. Evidence Sources Cited in Table 11.2

Evidence no.	Chapter	Section
3.1	3	Items and Testlets
3.2	3	External Reviews
3.3	3	Operational Assessment Items for 2017–2018
3.4	3	Field Testing
4.1	4	Writing Testlet Assignment
4.2	4	Instructionally Embedded Administration
4.3	4	User Experience With the DLM System
4.4	4	Accessibility
5.1	5	All
7.1	7	Student Performance
7.2	7	Score Reports
8.1	8	All
9.1	9	Evidence Based on Test Content
9.2	9	Evidence Based on Response Processes
9.3	9	Evidence Based on Internal Structure
9.4	9	Evidence Based on Consequences of Testing

11.2. Continuous Improvement

11.2.1. Operational Assessment

As noted previously in this manual, 2017–2018 was the third year the DLM Alternate Assessment System was operational. While the 2017–2018 assessments were carried out in a manner that supports the validity of inferences made from results for the intended purposes, the DLM Alternate Assessment Consortium is committed to continual improvement of assessments, teacher and student experiences, and technological delivery of the assessment system. Through formal research and evaluation as well as informal feedback, some improvements have already been implemented for 2018–2019. This section describes significant changes from the second to third year of operational administration, as well as examples of improvements to be made during the 2018–2019 year.

Overall, there were no significant changes to the learning map models, item-writing procedures, item flagging outcomes, the modeling procedure used to calibrate and score assessments, or the method for quantifying the reliability of results from previous years to 2017–2018.

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

The validity evidence collected in 2017–2018 expands upon the data compiled in the first two operational years for four of the critical sources of evidence as described in *Standards for Educational and Psychological Testing* (AERA et al., 2014): evidence based on test content, internal structure, response process, and consequences of testing. Specifically, analysis of opportunity to learn contributed to the evidence collected based on test content. Teacher-survey responses on test administration further contributed to the body of evidence collected based on response process, in addition to test-administration observations and evaluation of interrater agreement on the scoring of student writing products. Evaluation of item-level bias via differential item functioning analysis, along with item-pool statistics and model parameters, provided additional evidence collected based on internal structure. Teacher-survey responses also provided evidence based on consequences of testing, as well as a summary of findings from score-report focus groups collecting teacher feedback on their use of summative reports in the subsequent academic year. Studies planned for 2018–2019 to provide additional validity evidence are summarized in the following section.

11.2.2. Future Research

The continuous improvement process also leads to future directions for research to inform and improve the DLM Alternate Assessment System in 2018–2019 and beyond. The manual identifies some areas for further investigation.

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

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

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

Other ongoing operational research is also anticipated to grow as more data become available. For example, differential item functioning analyses will be expanded to include evaluating items across expressive communication subgroups, as identified by the First Contact survey.

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

12. References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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

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

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. London, England: Routledge.

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

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

Dynamic Learning Maps Consortium. (2017a). *2015–2016 Technical Manual—Science*. University of Kansas, Center for Educational Testing and Evaluation. Lawrence, KS.

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

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

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

Dynamic Learning Maps Consortium. (2018a). *2016–2017 Technical Manual Update—Science*. University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS). Lawrence, KS.

Dynamic Learning Maps Consortium. (2018b). *2017–2018 Technical Manual Update—Integrated Model*. University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS). Lawrence, KS.

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

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

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

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

Karvonen, M., Wakeman, S. Y., Browder, D. M., Rogers, M. A., & Flowers, C. (2011). Academic curriculum for students with significant cognitive disabilities: Special education teacher perspectives a decade after IDEA 1997. Retrieved from ERIC database.

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

Nash, B. & Bechard, S. (2016). *Summary of the Science Dynamic Learning Maps Alternate Assessment Development Process* (tech. rep. No. 16-02). University of Kansas, Center for Educational Testing and Evaluation. Lawrence, KS.

National Research Council. (2012). *A Framework for K-12 science education: Practice, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, by States*. Washington, DC: The National Academies Press.

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

Templin, J. & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*(2), 251–275. doi:10.1007/s00357-013-9129-4[12]

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

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

A. Differential Item Functioning Plots

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

A.1. Uniform Model

No items had a non-negligible effect-size change when comparing equation (9.2) to equation (9.1). In this model, the probability of a correct response was modeled as a function of ability and gender.

A.2. Combined Model

These plots show items that had a non-negligible effect-size change when comparing equation (9.3) to equation (9.1). In this model, the probability of a correct response was modeled as a function of ability, gender, and their interaction.

Item 51584

χ^2 = 9.00, p = 0.0027; Nagelkerke's R^2 = 0.89, Zumbo & Thomas: *large*, Jodoin & Gierl: *large*