DYNAMIC®
LEARNING MAPS

2018–2019 Technical Manual Update

Science

December 2019

Dynamic Learning Maps Consortium. (2019, December). *2018–2019 Technical Manual Update—Science*. Lawrence, KS: University of Kansas, 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.

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

The authors wish to acknowledge Teri Millar, 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 2018–2019 operational year include:

Russell Almond, Ph.D., *Florida State University*
Greg Camilli, Ph.D., *Rutgers University*
Karla Egan, Ph.D., *EdMetric*
Claudia Flowers, Ph.D, *University of North Carolina-Charlotte*
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 2018–2019 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], 2019a; DLM Consortium, 2019b).

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 2018–2019 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 2018–2019, DLM science assessments were administered to students in 16 states and one Bureau of Indian Education school: Alaska, Arkansas, Delaware, District of Columbia, Illinois, Iowa, Kansas, Maryland, Missouri, New Hampshire, New Jersey, New York, Oklahoma, Rhode Island, West Virginia, Wisconsin, and Miccosukee Indian School.

In 2018–2019, 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.

1.2. Technical Manual Overview

This manual provides evidence collected during the 2018–2019 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 2018–2019 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 2018–2019 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 2018–2019 year. The chapter provides updated information about adaptive routing in the system, Personal Needs and Preferences Profile selections, and teacher survey results regarding educator experience and system accessibility.

Chapter 5 provides a brief summary of the psychometric model used in scoring DLM assessments. This chapter includes a summary of 2018–2019 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 provides a summary of the changes made to the cut points used in scoring DLM assessments for grade 3 and grade 7 during the 2018–2019 administration. 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 2018–2019 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 2018–2019 administration.

Chapter 8 summarizes reliability evidence for the 2018–2019 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 2018–2019 administration not covered in previous chapters. The chapter provides study results for four of the five critical sources of evidence: test content, internal structure, response process, and consequences of testing.

Chapter 10 was not updated for 2018–2019. 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 2018–2019 academic year. The first portion of the chapter includes an analysis of answer option selection and provides an overview of 2018–2019 item writers' characteristics. The next portion of the chapter describes the pool of operational and field test testlets administered during spring 2019.

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

During 2018–2019 we analyzed answer-option selection for the operational pool. 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 an 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.

We evaluated the current the current operational item pool[1] to determine the number of items for which each answer option (A, B, or C) was the correct option, also called the key. As mentioned, the first answer option is always the key for all teacher-administered items (i.e., items measuring the initial linkage level); therefore, Table 3.1 shows the number and percentage of items for which each answer option is the key for computer-administered items (i.e., items measuring the Precursor and Target linkage levels). Across items, the key was fairly evenly distributed between the three answer options, with A having a slightly higher prevalence than B or C in the Precursor linkage level. Answer option A was nearly twice as prevalent as answer option C in the Target linkage level.

Table 3.1. Number and Percentage of Computer-Delivered Items by Answer Key

	Precursor		Target	
Answer Key	***n***	**%**	***n***	**%**
A	98	41.2	46	43.8
B	68	28.6	33	31.4
C	72	30.3	26	24.8

An additional analysis was conducted to determine if item difficulty differed by answer key. A weighted *p*-value was calculated for items with each answer option as the key, weighted by each item's sample size. Table 3.2 presents the weighted *p*-values for computer-delivered three-option multiple-choice items. Results suggest that for both linkage levels, items that have B as the answer key may be, on average, slightly more difficult than items where A or C is the key. Because of adaptive routing, students take items at different linkage levels across the Essential Elements (EEs). Because *p*-values are sample-dependent, values are not directly comparable to one another. In other words, fluctuations in *p*-values may also reflect differences in the samples of students who took the items.

[1]These analyses include items that were in the operational item pool and administered during the testing window.

Table 3.2. Weighted *p*-values by Answer Key for Computer-Delivered Items

	Precursor		Target	
Answer Key	***p*-value**	**SE**	***p*-value**	**SE**
A	0.592	0.001	0.712	0.001
B	0.556	0.001	0.674	0.001
C	0.607	0.001	0.778	0.002

3.1.2. Item Writing

For the 2018–2019 year, items were written to replenish the pool. The item writing process for 2018–2019 began with an on-site event in January 2019. Following this initial event, item writing continued remotely via a secure online platform. A total of 265 testlets were written for science.

3.1.2.1. Accessibility and Fairness Considerations for Item Writing

A hybrid item writing model was implemented in January 2019. The model consisted of an online advance training course, a three-day face-to-face onsite training event, and continuous targeted training and feedback throughout a 6-month remote item writing session. This section describes the training item writers received regarding accessibility considerations and the writing of items and testlets that are appropriately challenging while maintaining links to grade-level content and minimizing barriers to students with specific needs.

Item writers were trained to use Essential Element Concept Maps (EECMs), which are graphical organizers, structured around the core evidence centered design (ECD) principles of design patterns, development specifications, and task templates, guide the development of accessible items and testlets aligned to the linkage level. The EECMs provide specific guidance on accessible concepts and language use, define the skill development for each linkage level, and identify content, through the use of an accessibility flag, that may require an alternative approach to assessment for some students (e.g., braille). For more information about the content of EECMs and the ECD approach used by the DLM system, see Chapter 4 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a). The development of the EECM template was aimed at generating clear and easy-to-use task templates to aid item writers in writing items and testlets that include all of the essential features needed for a valid assessment for students with significant cognitive disabilities. This included the application of Universal Design for Learning principles to ensure a wide variety of supports, reflecting the diversity of the population of students with significant cognitive disabilities. For a complete description of the development process for EECMs for English language arts (ELA) and mathematics, which were the basis for the science EECMs, see Bechard et al. (2019).

Item writers were also trained to use the DLM Taxonomy of Cognitive Process Dimensions to make judgments on the complexity of the items and testlets they write to ensure student accessibility to the construct. Other training involved using accessible language in items to ensure writers use clear language free of unnecessary or distracting verbiage to minimize the need for inferences and prior knowledge, while also maintaining a link to grade-level content. The training also addressed using accessible vocabulary, which includes using high-frequency words, single-syllable words, and decodable words, while avoiding multiple-meaning words. Science testlets include a science story, which is intended to engage the student in the content of the testlet. Item writers received additional

training on improving access to the science content through a science story. The training included how to break multi-step problems down to address one step at a time in the display of the science story and items, which responds to the barrier of limited working memory for some students. Also, training included avoidance of emotional content, which can be a barrier to students' demonstration of knowledge, skills, and understanding.

Item writers also received training on using fair and people-positive language. The training included ensuring the use of language that does not require background knowledge outside the bounds of the targeted construct and that the language neither prevents nor promotes any regional or cultural group from demonstrating what they know about the targeted construct. It also included information about using people-first language for individuals with disabilities and to ensure populations are not depicted stereotypically. Item writers were trained on how to select and request accessible graphics. The training provided information about what makes a graphic accessible for students, such as, only including the information about the written content and being easy to describe with alternate text for students who are blind or have visual impairments. Item writers were trained to peer-review their partner's testlets, which included information about assessing each of the accessibility checks on a peer-review checklist and providing feedback if needed. The accessibility checks included ensuring the language, word choice, sentence structure, and graphics and images were appropriate for the EE and linkage level and maximized accessibility for all students.

3.1.2.2. 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, 25 item writers wrote testlets for the 2018–2019 year. The median and range of years of teaching experience in four areas the item writers had is shown in Table 3.3. The median years of experience was at least 13 years for item writers of science testlets in pre-K–12, special education, and science.

Table 3.3. Item Writers' Years of Teaching Experience

Area	Median	Range
Pre-K–12	17	6-30
Science	13.5	0-26
Special Education	14	0-30

The level and types of degrees held by item writers are shown in Table 3.4. All item writers held at least a Bachelor's degree, with the most common field of study being education (n = 8; 32%). A majority (n = 24; 96%) also held a Master's degree, and the most common field of study was special education (n = 11; 44%).

Table 3.4. Item Writers' Level and Type of Degree

Degree	*n*	%
Bachelor's Degree	**25**	**100.0**
Education	8	32.0
Content Specific	0	0.0
Special Education	6	24.0
Other	7	28.0
Missing	4	16.0
Master's Degree	**24**	**96.0**
Education	1	4.0
Content Specific	0	0.0
Special Education	11	44.0
Other	10	40.0
Missing	0	0.0
Other Advanced Degree	**10**	**40.0**

Item writers reported a range of experience working with students with different disabilities, as summarized in Table 3.5. Teachers collectively had the most experience working with students with multiple disabilities, significant cognitive disability, or speech impairment.

Table 3.5. Item Writers' Experience with Disability Categories

Disability Category	*n*	%
Blind/Low Vision	9	36.0
Deaf/Hard of Hearing	9	36.0
Emotional Disability	13	52.0
Mild Cognitive Disability	13	52.0
Multiple Disabilities	17	68.0
Orthopedic Impairment	7	28.0
Other Health Impairment	11	44.0
Significant Cognitive Disability	17	68.0
Specific Learning Disability	13	52.0
Speech Impairment	14	56.0
Traumatic Brain Injury	8	32.0
Not reported	6	24.0

3.2. External Reviews

As described in Chapter 3 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a), the purpose of external review is to evaluate items and testlets developed for the DLM alternate assessment before field test administration. The DLM external review criteria were developed in partnership with members of the DLM governance board and used at an external review pilot event in 2013. Educators and governance partners participated in and provided feedback on the pilot external review event. The clarity and appropriateness of the review criteria were evaluated after the

event, and minor modifications were made. While originally developed for English language arts and mathematics, the DLM Science Consortium adopted the same external review process, procedures and review criteria for evaluating science items and testlets. There are three panel types in the external review process: accessibility, bias and sensitivity, and content. This section describes the updated external review criteria for the three panels and the external reviewer training.

3.2.1. External Review Panelist Training

External review events are conducted onsite in Kansas City, MO, with educators representing the DLM partner states. Participating educators complete an online advance training course before attending the onsite event, during which additional training is provided. The training prioritizes participant understanding of each criterion within the item and testlets to support judgments of accept, revise, or reject for the items and testlets.

In addition to training on how to interpret and use the external review criteria, panelists received training on the overview of the review process, as well as training specific to the type of panel they were assigned. For example, the accessibility panel received training about the DLM system and how each system component is intended to maximize accessibility for students, including how the assessment delivery platform allows for a variety of customized accessibility supports (i.e., color contrast, whole-screen magnification, text-to-speech, etc.). Additionally, panelist training includes information on how to judge if the language used in the items is accessible, (i.e., uses simple sentence structure, avoids pronouns, avoids multiple-meaning words, etc.); graphics in the items are accessible, (i.e., only contains the necessary elements, are clearly marked or labeled [if necessary], does not add information beyond the text, etc.); and the content is accessible through the use of science stories. Finally, the training for evaluating the accessibility of testlets contains information on identifying if the testlet is instructionally relevant and barrier-free (i.e., used single-step problems, used simple language structure, etc.).

The external review panelists' bias and sensitivity training includes information about how to judge if items are fairly assessing the construct by not requiring background knowledge; representing the topic accurately; not biased towards a subgroup of the population; and/or not measuring group membership more than the content objective. The external review panelists are trained to identify potentially sensitive content in a testlet; demeaning or offensive material; and religious references.

External review panelists rate items and testlets individually on each review criteria. The panelist reviews all items and testlets that receive a revise or reject rating. A table facilitator leads a consensus-building conversation, and a final revise or reject rating is determined for each item and testlet. If a revise rating is made, the panelists include a specific suggested revision to the item(s) or testlet to address the identified issue.

3.2.2. Review Recruitment, Assignments, and Training

In April 2018, a volunteer survey was used to recruit external review panelists. Volunteers for the external review process completed the Qualtrics survey to capture demographic information as well as information about their education and experience. The candidates were screened by the implementation and test development teams to ensure they qualified. These data were then used to identify panel types (content, bias and sensitivity, and accessibility) for which the candidate would be eligible. A total of 19 individuals were placed on external review panels.

Each reviewer was assigned to one of the three panel types. There were 19 science reviewers: 6 on accessibility panels, 9 on content panels, and 4 on bias and sensitivity panels.

Panelists completed 6 rounds of reviews. Each round consisted of 1 collection of testlets that ranged from 6 testlets to 26 testlets, dependent on the panel type. Content panels had the smallest number of testlets per collection, and bias and sensitivity panels had the largest number of testlets per collection.

The professional roles reported by the 2017–2018 reviewers are shown in Table 3.6. Reviewers who reported "Other" roles included state education agency (SEA) staff, speech language therapists, principals, and process coordinators.

Table 3.6. Professional Roles of External Reviewers

	Science	
Role	***n***	**%**
Classroom Teacher	11	57.9
District Staff	4	21.1
Instructional Coach	0	0.0
Other	4	21.1

Reviewers had varying experience teaching students with the most significant cognitive disabilities. Science reviewers had a median of 15 years of experience, with a minimum of 4 and a maximum of 29 years of experience.

Population density of schools in which reviewers taught or held a position is reported in Table 3.7. Rural was defined as a population living outside settlements of 1,000 or fewer inhabitants, suburban was defined as an outlying residential area of a city of 2,000–49,000 or more inhabitants, and urban was defined as a city of 50,000 inhabitants or more.

Table 3.7. Population Density for Schools of External Reviewers

	Science	
Population Density	***n***	**%**
Rural	6	31.6
Suburban	4	21.1
Urban	8	42.1
Not Applicable	1	5.3

Prior to attending the on-site external review event, panelists completed an advance training course. The course included two modules that all panelists had to complete: DLM Overview and External Review Process. After each module, the panelists had to complete a quiz and receive a score of at least 80% to continue to the next module. After completing the first two modules and quizzes, each panelist was then directed to a module and quiz that was catered towards their subject and panel type. While the bias and sensitivity and accessibility modules were universal for all subjects, each content module was subject-specific. Panelists were required to complete advance training prior to

reviewing any testlets at the event.

Review of testlets was completed during the on-site two day training. The panelists reviewed each testlet on their own and then reviewed them together as a group. Each group came to a consensus for each item and testlet, and the facilitator recorded that recommendation for the test development teams to consider.

3.2.3. Results of Reviews

Most of the externally reviewed content was included in the 2019 fall and 2020 spring windows. For science, the percentage of items and testlets rated as ***accept*** ranged from 50% to 88% and 31% to 88%, respectively. The percentage of items and testlets rated as ***revise*** ranged from 12% to 40% and 6% to 69%, respectively. The rate at which items and testlets were recommended for rejection ranged from 0% to 1% and 0% to 1%, respectively, across grades, panels, and rounds of review. A summary of the test development team decisions and outcomes is provided here.

3.2.4. Test Development Team Decisions

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

DLM test development teams documented the decision category applied by external reviewers to each item and testlet. Following this process, test development teams made a final decision to accept, revise, or reject each of the items and testlets. The science content team retained 98% of items and testlets sent out for external review. Most revisions made to items and teslets were minor. The science team made 121 minor revisions to items and 11 minor revisions to testlets.

3.3. Operational Assessment Items for Spring 2019

A total of 333,694 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 2019 based on the 2017–2018 operational pool and the testlets field-tested during 2017–2018 that were promoted to the operational pool following their review. Table 3.8 summarizes the total number of operational testlets for spring 2019 for science. There were 151 operational testlets available across grade bands and courses. This total included 1 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 3.8. Distribution of Spring 2019 Operational Testlets, by Grade Band or Course (N = 151)

Grade Band or Course	*n*
Elementary	42
Middle School	45
High School	43
Biology	31

Note: Ten 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 .4 and .6.

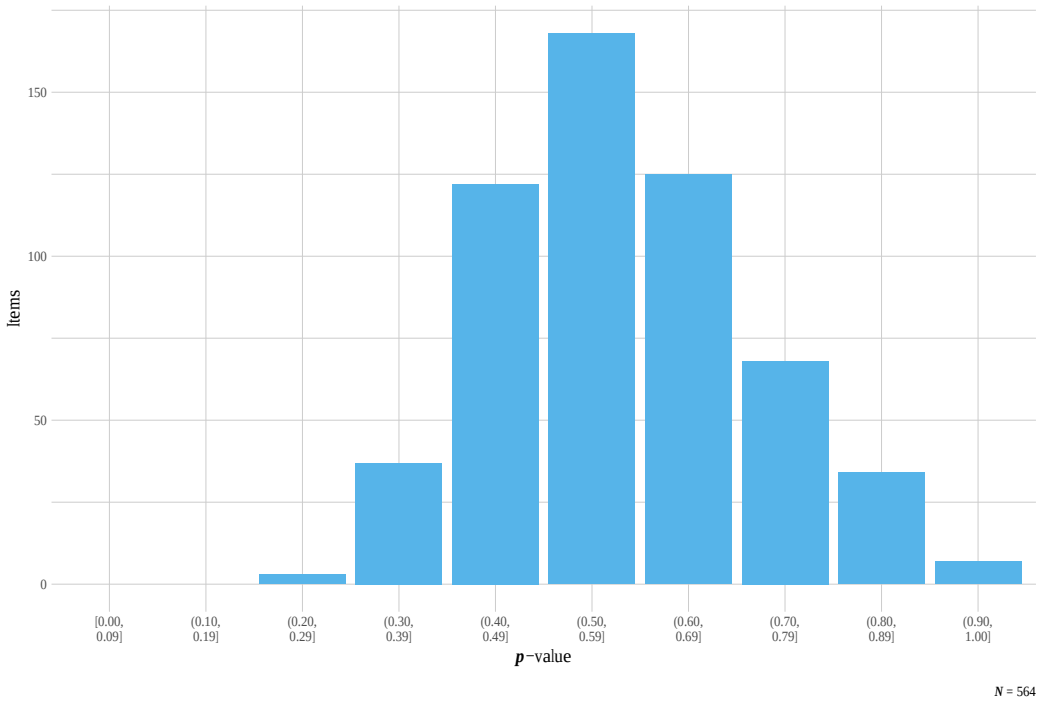


Figure 3.1. p-values for science 2019 operational items. *Note.* Items with a sample size of less than 20 were omitted.

Standardized difference values were also calculated for all operational items with a student sample size of at least 20 required 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. All items fell within two standard deviations of the mean of all items measuring the EE and linkage level. As additional data are collected and decisions are made regarding item pool replenishment, test development teams will consider item standardized difference values when determining which items and testlets are recommended for retirement.

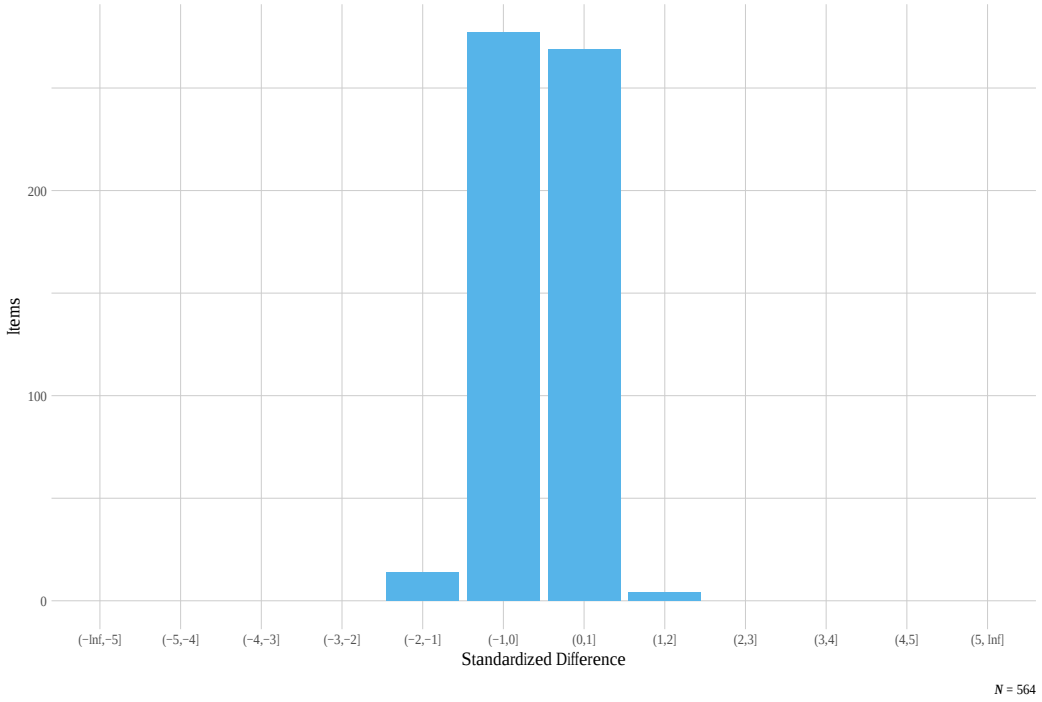


Figure 3.2. Standardized difference *z*-scores for science 2019 operational items. *Note*. Items with a sample size of less than 20 were omitted.

3.4. Field Testing

During the spring 2019 administration, DLM field tests were administered to collect student data on linkage levels adjacent to those taken during the operational assessment. By collecting this data, we are better able to empirically evaluate the relationships between linkage levels.

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. 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 spring 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.

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, 28,767 (76%) 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.

3.5. Conclusion

During the 2018–2019 academic year, the test development teams conducted events for both item writing and external review. Overall, 265 testlets were written for science. Additionally, during external review, 98% of science testlets were retained with no or minor changes. Of the content already in the operational pool, all items had p-values within two standard deviations of the average for the the EE and linkage level. Field testing in 2018–2019 focused on collecting data from students at linkage levels adjacent to those administered during the operational assessment to support future modeling work. Field testing in 2019-–2020 will be focused on collecting data for the content that was retained during the external review event described in this chapter.

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 2018–2019, 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 2018–2019. For a complete description of key administration features, including information on assessment delivery, Kite Student Portal, 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 2018–2019* (DLM Consortium, 2018d) and the *Educator Portal User Guide* (DLM Consortium, 2018c).

4.1.1. Test Windows

Instructionally embedded science assessments were available for teachers to optionally administer between September 19 and December 19, 2018, and between January 2 and February 27, 2019. During the consortium-wide spring testing window, which occurred between March 11 and June 7, 2019, 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 2019 operational administration of the DLM Science alternate assessment. The categories of evidence include data relating to administration time, the adaptive delivery of testlets in the spring window, user experience, and accessibility.

4.2.1. Administration Time

Estimated administration time varies by student and subject. During the spring testing window, estimated total testing time was between 45-135 minutes per student, with each testlet taking approximately 5-15 minutes. Actual testing time per testlet varies depending on each student's unique characteristics.

Kite Student Portal captured start and end dates and time stamps for every testlet. Actual testing time per testlet was calculated as the difference between start and end times. Table 4.1 shows the distribution of test times per testlet. Most testlets took approximately 2-3 minutes to complete.

Testlets time out after 90 minutes.

Table 4.1. Distribution of Response Times per Testlet in Minutes

Grade	Min	Median	Mean	Max	25Q	75Q	IQR
3-5	0.07	2.15	2.94	89.90	1.33	3.45	2.12
6-8	0.08	2.08	2.87	89.80	1.27	3.37	2.10
9-12	0.00	2.28	3.14	89.22	1.38	3.67	2.29

Note: 25Q = lower quartile; 75Q = upper quartile; IQR = interquartile range.

4.2.2. Adaptive Delivery

During the spring 2019 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 4.2.

Table 4.2. 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 2019 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 4.3.

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 57% to 73%). A similar pattern was seen for students assigned to Complexity Band 3, with the majority of students not adapting down to a lower linkage level after the first assigned testlet (ranging from 68% to 86%). Consistent patterns were not as apparent for students who were assigned Complexity Band 1 or Complexity Band 2. Distributions across the three categories were more variable across grade bands. Further investigation is needed to evaluate reasons for these different patterns.

The 2018–2019 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 14% and 43% 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. However, patterns of adaptation beyond the first adaptation opportunity (e.g., between the second and third testlets, third and forth testlets, etc.), indicate that majority of Band 1 students adapt back down to the Initial level during the assessment, rather than remaining at the Precursor level. Thus, changes to the assignment process are not planned. For a description of previous findings, see Chapter 4 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a) and the subsequent annual technical manual updates (DLM Consortium, 2018a, 2018b).

Table 4.3. Adaptation of Linkage Levels Between First and Second Science Testlets (N = 37,809)

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	43.1	56.9	79.2	20.8	31.1	43.1	25.8	67.6	32.4
6–8	31.5	68.5	66.8	33.2	40.5	39.8	19.7	70.5	29.5
9–12	30.6	69.4	61.3	38.7	43.9	39.0	17.1	80.1	19.9
Biology	27.3	72.7	16.3	83.7	24.1	33.3	42.6	85.7	14.3

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.

4.2.3. Administration Incidents

As in all previous operational years, testlet assignment during the spring 2019 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 2019 science 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 2019 operational implementation of the DLM Science alternate assessment. The categories of evidence include survey data relating to user experience and accessibility.

4.3.1. User Experience with the DLM System

User experience with the 2018–2019 assessments was evaluated through the spring 2019 survey, which was disseminated to teachers who had administered a DLM assessment during the spring window. In 2019, the survey was distributed to teachers in Kite Student Portal, 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 Student Portal, the relationship of assessment content to instruction by subject, and score reports.

A total of 10,897 teachers from states participating in DLM science assessments responded to the survey (with a response rate of 78%) for 23,977 students.

Participating teachers responded to surveys for a median of two students. Teachers reported having an average of 10 years of experience in science and with students with significant cognitive disabilities. The median response to the number of years of experience in science was 8 years, and the median experience with students with significant cognitive disabilities was 7 years. Approximately 33% indicated they had experience administering the DLM science assessment in all four operational years.

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

4.3.1.1. Educator Experience

Survey respondents were asked to reflect on their own experience with the assessments as well as their comfort level and knowledge administering them. Most of the questions required teachers to respond on a four-point scale: *strongly disagree*, *disagree*, *agree*, or *strongly agree*. Responses are summarized in Table 4.4.

Nearly all teachers (96%) agreed or strongly agreed that they were confident administering DLM

testlets. Most respondents (89%) agreed or strongly agreed that the required test administrator training prepared them for their responsibilities as test administrators. Most teachers also responded that they had access to curriculum aligned with the content that was measured by the assessments (86%) and that they used the manuals and the Educator Resources page (92%).

Table 4.4. Teacher Responses Regarding Test Administration

	SD		D		A		SA		A+SA	
Statement	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
I was confident in my ability to deliver DLM testlets.	74	1.0	206	2.8	3,068	41.7	4,018	54.5	7,086	96.2
Required test administrator training prepared me for the responsibilities of a test administrator.	194	2.6	583	7.9	3,701	50.3	2,875	39.1	6,576	89.4
I have access to curriculum aligned with the content measured by DLM assessments.	202	2.7	803	10.9	3,798	51.6	2,553	34.7	6,351	86.3
I used manuals and/or the DLM Educator Resource Page materials.	123	1.7	506	6.9	4,052	55.0	2,690	36.5	6,742	91.5

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 Student Portal and Educator Portal.

The software used for the administration of DLM testlets is Kite Student Portal. Teachers were asked to consider their experiences with Kite Student Portal and respond to each question on a four-point scale: *very hard*, *somewhat hard*, *somewhat easy*, or *very easy*. Table 4.5 summarizes teacher responses to these questions.

Respondents found it to be either *somewhat easy* or *very easy* to log in to the system (93%), to navigate within a testlet (94%), to record a response (96%), to submit a completed testlet (97%), and to administer testlets on various devices (92%). Open-ended survey response feedback indicated testlets were easy to administer and that technology had improved compared to previous years.

Table 4.5. Ease of Using Kite Student Portal

	VH		SH		SE		VE		SE+VE	
Statement	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Enter the site	80	1.4	325	5.6	2,009	34.7	3,370	58.3	5,379	93.0
Navigate within a testlet	55	1.0	260	4.5	1,895	32.8	3,561	61.7	5,456	94.5
Record a response	39	0.7	176	3.1	1,690	29.4	3,848	66.9	5,538	96.3
Submit a completed testlet	43	0.7	141	2.5	1,625	28.3	3,941	68.5	5,566	96.8
Administer testlets on various devices	95	1.7	364	6.4	2,078	36.3	3,194	55.7	5,272	92.0

Note: VH = very hard; SH = somewhat hard; 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. To address teachers' feedback from prior administrations, the appearance and functionality of Educator Portal was updated during the summer of 2018. The update focused on the improvement of user experience, accessibility, and a general improvement to the look, feel, and functionality of Educator Portal without causing undue disruption to how educators use the application. Updates made to Educator Portal during the summer of 2018 include: updating the user interface to be more intuitive, have a more logical flow, display auto-populated fields, and restrict users from saving incomplete records; reordering tabs to be more intuitive; updating the color scheme to be consistent across the application; and rewriting data upload error messages in nontechnical language instead of programming language.

Teachers were asked to assess the ease of navigating and using Educator Portal for its intended purposes. The data are summarized in Table 4.6 using the same scale used to rate experiences with Kite Student Portal. Overall, the improvements made to Educator Portal during summer 2018 are reflected in the respondents' favorable feedback. The percentage of teachers rating *somewhat easy* or *very easy* increased over last year (DLM Consortium, 2018a). A majority of teachers found it to be either *somewhat easy* or *very easy* to navigate the site (87%), enter PNP and First Contact information (91%), manage student data (88%), manage their accounts (90%), or manage tests (89%).

Table 4.6. Ease of Using Educator Portal

Statement	VH		SH		SE		VE		SE+VE	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
Navigate the site	93	2.1	482	10.7	1,831	40.8	2,085	46.4	3,916	87.2
Enter Access Profile and First Contact information	63	1.4	334	7.5	1,877	41.9	2,206	49.2	4,083	91.1
Manage student data	90	2.0	431	9.6	1,956	43.8	1,992	44.6	3,948	88.4
Manage my account	75	1.7	370	8.3	1,977	44.2	2,054	45.9	4,031	90.1
Manage tests	87	1.9	421	9.4	1,869	41.7	2,102	46.9	3,971	88.6

Note: VH = very hard; SH = somewhat hard; SE = somewhat easy; VE = very easy; SE+VE = somewhat easy and very easy.

Finally, respondents were asked to rate their overall experience with Kite Student Portal and Educator Portal on a four-point scale: *poor, fair, good,* and *excellent*. Results are summarized in Table 4.7. The majority of respondents reported a positive experience with Kite Student Portal. A total of 89% of respondents rated their Kite Student Portal experience as *good* or *excellent*, while 81% rated their overall experience with Educator Portal as *good* or *excellent*.

Table 4.7. Overall Experience With Kite Student Portal and Educator Portal

Statement	Poor		Fair		Good		Excellent		Good + Excellent	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
Student Portal	106	1.8	539	9.3	2,757	47.5	2,397	41.3	5,154	88.8
Educator Portal	216	3.7	878	15.1	3,005	51.7	1,711	29.4	4,716	81.1

Overall, feedback from teachers indicated that Kite Student Portal and Educator Portal was easy to navigate and user friendly.

4.3.2. Accessibility

Accessibility supports provided in 2018–2019 were the same as those available in previous years. The *DLM Accessibility Manual* DLM Consortium (2017b), distinguishes among accessibility supports that are provided in Kite Student Portal via the Personal Needs and Preferences Profile, require additional tools or materials, or are provided by the test administrator outside the system.

Table 4.8 shows selection rates for the three categories of accessibility supports. The most commonly selected supports were human read aloud, test administrator enters responses for student, and individualized manipulatives. For a complete description of the available accessibility supports, see Chapter 4 in the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a).

Table 4.8. Accessibility Supports Selected for Students (N = 30,134)

Support	*n*	%
Supports provided in Kite Student Portal via Access Profile		
Individualized manipulatives	12,235	40.6
Calculator	9,913	32.9
Single-switch system	996	3.3
Alternate form - visual impairment	742	2.5
Two-switch system	450	1.5
Supports requiring additional tools/materials		
Uncontracted braille	11	0.0
Human read aloud	26,097	86.6
Test administrator enters responses for student	15,799	52.4
Partner assisted scanning	2,361	7.8
Language translation of text	537	1.8
Sign interpretation of text	430	1.4
Supports provided outside the system		
Spoken audio	4,443	14.7
Magnification	3,158	10.5
Color contrast	2,581	8.6
Overlay color	1,469	4.9
Invert color choice	1,013	3.4

Table 4.9 describes teacher responses to survey items about the accessibility supports used during administration. Teachers were asked whether the student was able to effectively use available accessibility supports and whether the accessibility supports were similar to the ones used for instruction. The majority of teachers agreed that students were able to effectively use accessibility supports (94%), while responses to whether the accessibility supports were similar to ones students used for instruction were mixed (60%). While states and districts have differing policies for whether to include accessibility supports on the student's IEP, most (65%) indicated supports were included.

Table 4.9. Teacher Report of Student Accessibility Experience

	Agree		Disagree	
Statement	***n***	**%**	***n***	**%**
Student was able to effectively use accessibility features.	5976	93.5	416	6.5
Accessibility features were similar to ones student uses for instruction.	244	60.2	161	39.8

Of the teachers who reported that their student was unable to effectively use the accessibility

supports (6%), the most commonly reported reason was that the student could not provide a response even with the support provided (66%).

Table 4.10. Reason Student was Unable to Effectively Use Available Accessibility Supports

Reason	*n*	%
Student could not provide a response even with support	254	65.5
Student refused support during testing	83	21.4
Student needed a support which was not available or allowed	81	20.9
Student was unfamiliar with support	70	18.0
Technology problem	18	4.6

Teachers have several allowable options for flexibility while assessing students. Of these options for flexibility, teachers most frequently reported using breaks (64%), reinforcement (40%), or individualized student response mode (32%). Additionally, 32% of teachers reported adapting or substituting materials.

Table 4.11. Options for Flexibility Teachers Reported Utilizing for a Student

Option	*n*	%
Breaks	3,961	64.5
Use of reinforcement	2,429	39.5
Individualized student response mode	1,982	32.3
Blank paper	1,380	22.5
None of these	1,005	16.4
Navigation across screens	877	14.3
Alternate representation of answer options	874	14.2
Generic definitions	661	10.8
Special equipment for positioning	409	6.7
Display testlet on interactive whiteboard	299	4.9
Graphic organizer	260	4.2

While overall these data support the conclusion that the accessibility supports of the DLM alternate assessment were effectively used by students, additional data will be collected during the spring 2020 to determine whether additional improvements can be made to ensure all students can access DLM assessments.

4.4. Conclusion

During the 2018–2019 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 Profile selections. Results from the teacher survey indicated that teachers felt confident administering testlets in the system, that Kite Student Portal was easy to use, and that Educator Portal 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 2018–2019 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 2018–2019 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. In addition to the item parameters, the psychometric model also includes a structural parameter, which defines the base rate of mastery for each linkage level. A summary of the operational parameters used to score the 2018–2019 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 2019 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%) performed as expected. Additionally, 98% of linkage levels (n = 100) had a conditional probability of masters providing a correct response over .6.

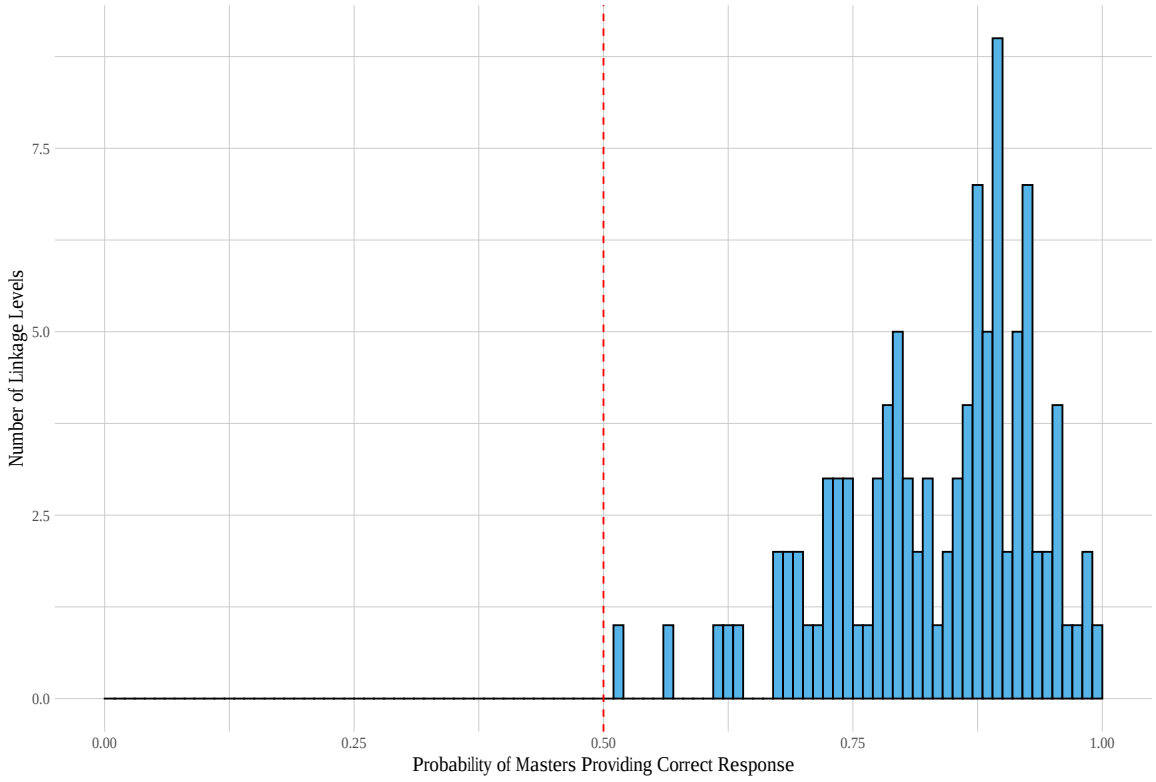


Figure 5.1. Probability of masters providing a correct response to items measuring each linkage level. *Note*. Histogram bins are shown in increments of .01. Reference line indicates .5.

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%) performed as expected, non-masters sometimes had a greater than chance (> .5) likelihood of providing a correct response to items measuring the linkage level. Although most linkage levels (n = 56, 55%) have a conditional probability of non-masters providing a correct response less than .4, 3 (3%) have a conditional

probability for non-masters providing a correct response greater than .6, indicating there are many linkage levels non-masters are more likely than not to provide a correct response. 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.

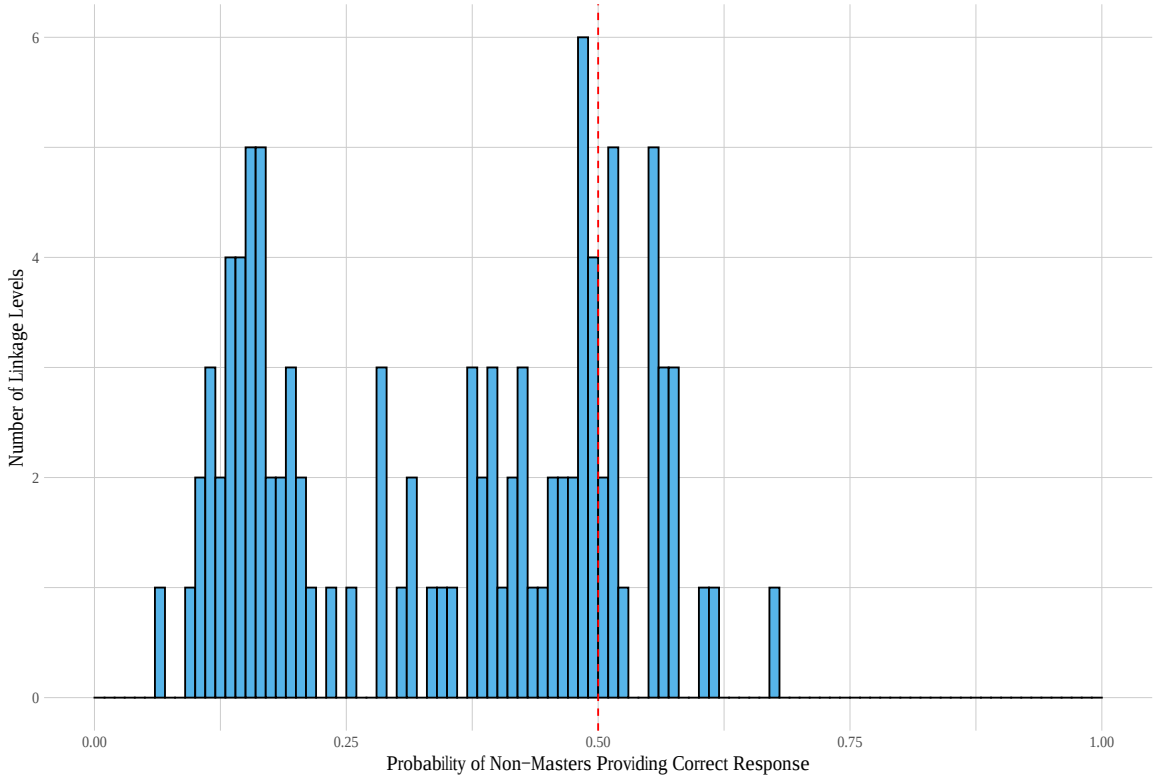


Figure 5.2. Probability of non-masters providing a correct response to items measuring each linkage level. *Note*. Histogram bins are in increments of .01. Reference line indicates .5.

5.2.3. Item Discrimination

The discrimination of a linkage level represents how well the items are able to differentiate masters and non-masters. For diagnostic models, this is assessed by comparing the conditional probabilities of masters and non-masters providing a correct response. Linkage levels that are highly discriminating will have a large difference between the conditional probabilities, with a maximum value of 1.0 (i.e., masters have a 100% chance of providing a correct response and non-masters a 0% chance). Figure 5.3 shows the distribution of linkage level discrimination values. Overall, 69% of linkage levels (n = 70) have a discrimination greater than .4, indicating a large difference between the conditional probabilities (e.g., .75 to .35, .9 to .5, etc.). However, there were 2 linkage levels (2%) with a discrimination of less than .1, indicating that masters and non-masters tend to perform similarly on items measuring these linkage levels.

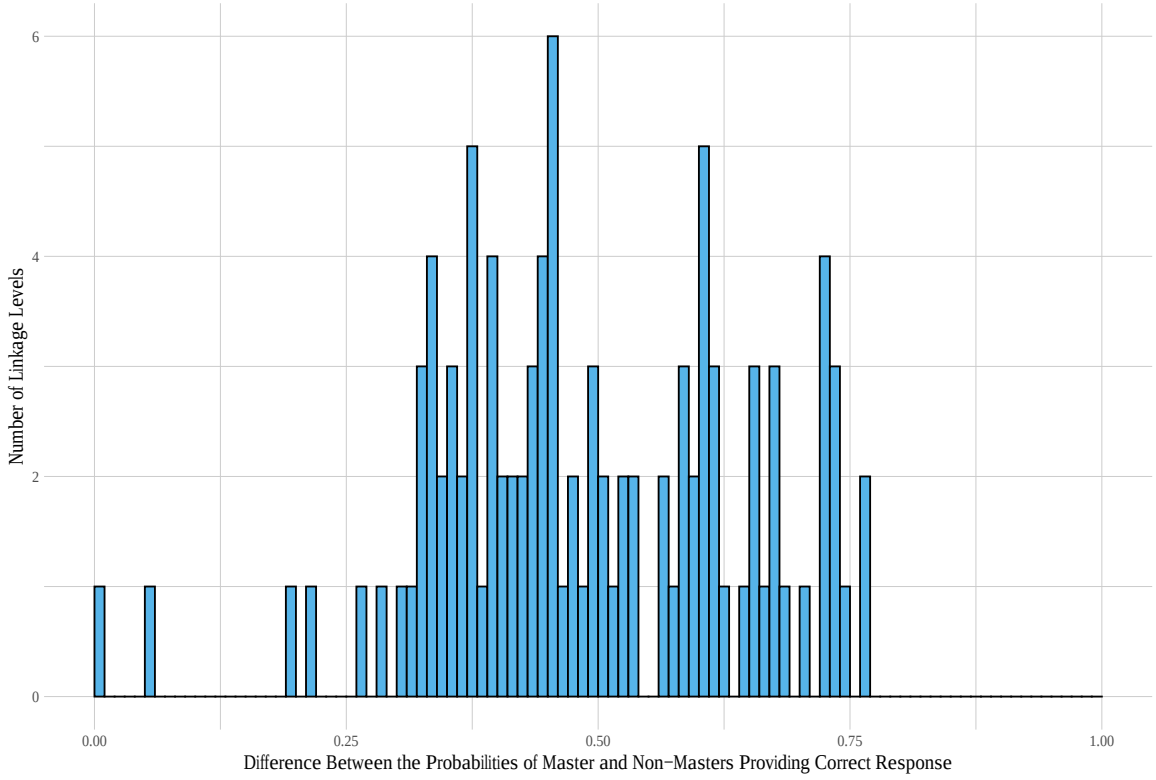


Figure 5.3. Difference between masters' and non-masters' probability of providing a correct response to items measuring each linkage level. *Note*. Histogram bins are in increments of .01. Reference line indicates .5.

5.2.4. Base Rate Probability of Mastery

The DLM assessments are designed to maximize the match of student knowledge and skill to the appropriate linkage level content. The base rate of mastery represents the estimated proportion of masters among students assessed on an EE and linkage level. A base rate of mastery close to .5 indicates that students assessed on a given linkage level are equally likely to be a master or non-master. Conversely a high base rate of mastery would indicate that nearly all students testing on a linkage level are classified as masters. Figure 5.4 depicts the distribution of the base rate of mastery probabilities. Overall, 72% of linkage levels (n = 73) had a base rate of mastery between .25 and .75. This indicates that most linkage levels are performing as expected. On the edges of the distribution, 14 linkage levels (14%) had a base rate of mastery less than .25, and 15 linkage levels (15%) had a base rate of mastery higher than .75. This indicates that students are more likely be assessed on linkage levels they have mastered than those they have not mastered.

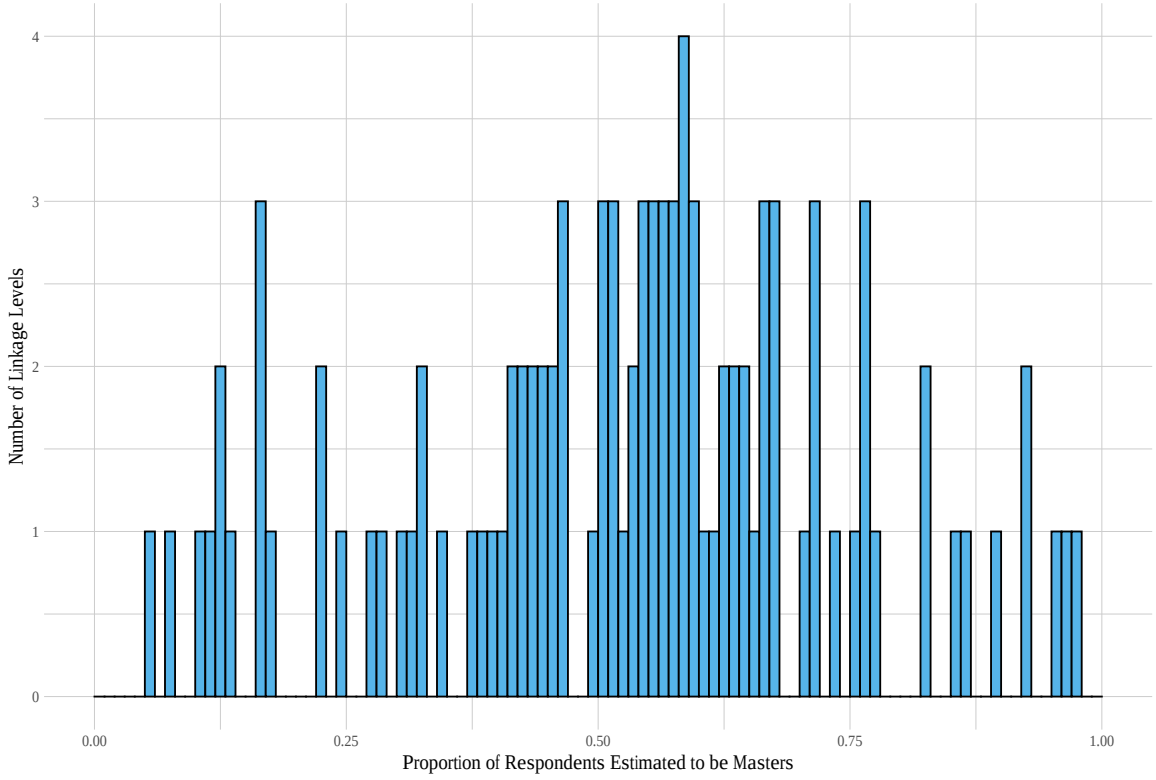


Figure 5.4. Base rate of linkage level mastery. *Note.* Histogram bins are shown in increments of .01.

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 2018–2019 administration of DLM assessments, the percentage of mastery statuses obtained by each scoring rule was calculated, as shown in Figure 5.5. 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 77% 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 12% to 18% 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 2018–2019 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 68% 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).

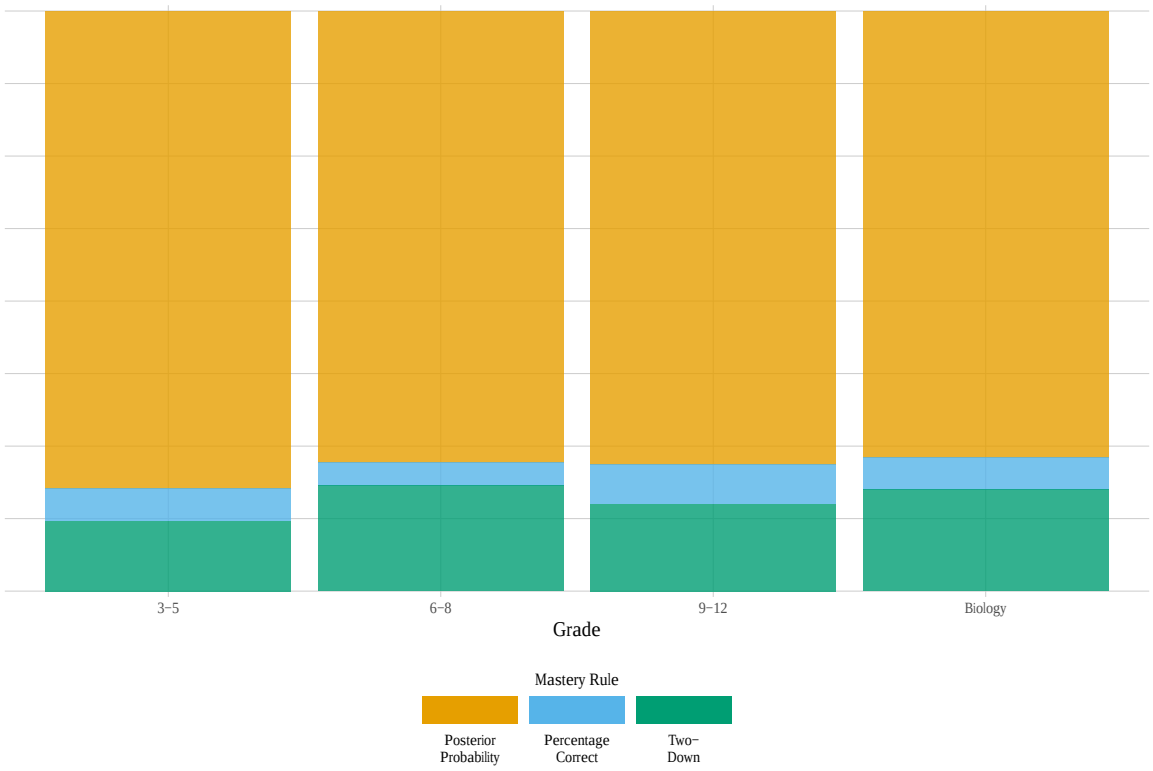


Figure 5.5. Linkage level mastery assignment by mastery rule for each grade band and course.

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 and 2018–2019 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 2018–2019 year, the modeling subcommittee reviewed research related to Bayesian methods for assessing model and item-level fit using posterior predictive model checks (Gelman & Hill, 2006; Gelman et al., 1996), the effect of partial equivalency constraints on model parameters, and new methods for model comparisons (e.g., Vehtari et al., 2017).

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 initial science standard-setting process conducted in 2016 for the Dynamic Learning Maps® (DLM®) Alternate Assessment System derived cut points for then-tested grades 4, 5, 6, 8, and high school. The process specified cuts for describing student achievement relative to four performance levels: Emerging, Approaching the Target, At Target, and Advanced. Because DLM assessments are scored using a diagnostic model to produce mastery determinations for each assessed Essential Element (EE), the standard-setting method used a profile-based method for specifying cuts between total linkage levels mastered (A. K. Clark et al., 2017). For a description of the process, including the development of policy for performance-level descriptors, the three-day science 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).

After a new state that assesses students in science in all grades joined the consortium, additional cut points were needed for grades 3 and 7. The standard-setting process for grades 3 and 7 used existing cuts to determine cut point values and consisted of a virtual panelist meeting (for grade 3 standard setting only), a review by the Technical Advisory Committee (TAC), and a state partner evaluation of the results. This chapter provides a brief description of the procedures and results of establishing grade 3 and 7 cut points. A more detailed description of the standard-setting activities and results can be found in *2019 Science Standard Setting: Grades 3 and 7* (Nash et al., 2019).

6.1. Standard Setting Grade 3

The grade 3 standard setting modified the methodology and used cuts determined in the original standard setting implemented in 2016 for grades 4, 5, 6, and 8. The consortium aimed to establish cut points in grades 3 and 7 without affecting the existing cut points. Range finding and pinpointing rating exercises were originally used to determine cut points (see 2016 Standard Setting: Science (Nash et al., 2016) for more details). The standard-setting process for 2019, however, prompted panelists to base their determinations on existing cut points in the grade band (i.e., the grade band containing grades 4 and 5) and impact data.

6.1.1. Panelists

DLM staff recruited standard-setting panelists from a database of volunteer educators spanning all DLM consortium states. Panelists' eligibility was determined by their experience educating students with significant cognitive disabilities and/or teaching science in elementary grade levels. The nine panelists who participated in the grade 3 standard setting represented four different states. Table 6.1 and Table 6.2 summarize their demographic information. Panelists held between 6 and 27 years of experience teaching science and 3 and 22 years teaching students with significant cognitive disabilities.

Table 6.1. Demographic Characteristics of Panelists

Demographics Category	Count
Gender	
Female	9
Male	0
Race	
African American	1
American Indian/Alaska Native	0
Asian	0
Hispanic/Latino	0
Native Hawaiian/Pacific Islander	0
White	8
State	
Arkansas	4
Iowa	2
Missouri	2
Rhode Island	1

Table 6.2. Panelists' Years of Experience

	Mean	Median	Min	Max
Students with the most significant cognitive disabilities	10.9	8	3	22
Science	13.7	12	6	27

6.1.2. Training

Panelists were provided with training both before and during the standard-setting workshop, which required approximately three to four hours of panelists' time. Advance training was available online on-demand in the 15 days prior to the standard-setting workshop. The advance training addressed the following topics:

1. Characteristics of students who take the DLM assessments
2. Content of the assessment system, including EEs for science, domains and topics, linkage levels, and alignment
3. Accessibility by design, including the framework for the DLM Alternate Assessment System's cognitive taxonomy and strategies for maximizing accessibility of the content; the use of the Access (Personal Needs and Preferences) Profile (PNP) to provide accessibility supports during the assessment; and the use of the First Contact survey to determine linkage level assignment
4. Assessment design, including item types, testlet design, and sample items from various linkage levels in science
5. An overview of the assessment model, including test blueprints and the timing and selection of testlets administered

6. A high-level introduction to two topics that would be covered in more detail during onsite training: the DLM approach to scoring and reporting and the steps in the standard-setting process.

Panelists responded to survey questions upon completion of the online advance training. The questions asked panelists to report (1) their perceived level of preparedness for the virtual meeting, (2) if they deemed their level of understanding of the DLM system to be sufficient to allow them to make decisions about student achievement, and (3) any questions that need clarification before the meeting. In response to the survey, one-third of the panelists reported they were somewhat prepared, while two-thirds reported they were very prepared for the virtual meeting. All panelists responded that they had an appropriate level of knowledge of the DLM system to allow them to review cut points, and none of the panelists asked any clarifying questions.

Supplementary panelist training took place during the virtual standard-setting event where panelists received (1) a general review of information (due to their report of preparedness and lack of requests for specific clarification), (2) an overview of DLM assessment scoring and reporting, and (3) a description of the standard-setting methodology and what was expected of them during the meeting.

6.1.3. Procedures

The existing cut points set during the original science standard-setting event (see 2016 Standard Setting: Science (Nash et al., 2016)) were used as a basis for grade 3 science standard-setting methodology. Mastery profiles were used in both standard-setting approaches, but range finding and pinpointing rating exercises were only used in the 2016 standard setting. For grade 3 standard setting, DLM staff used existing cut points in grades 4 and 5 and impact data to determine proposed cut points for grade 3. The proposed cut points were designed to produce similar percentages of students achieving at each performance level in grade 3 as in grades 4 and 5 (the other grades in the elementary grade band). They also were designed to ensure that cut point values were not duplicated and did not exceed the existing values in grade 4 or 5. DLM staff used DLM science assessment impact data from 2016 through May 8, 2019, to propose appropriate starting cut points.

After reviewing example student mastery profiles for the proposed cut points and adjacent values, the educator panel presented their recommendations for cut points.

6.1.3.1. Panel Profile Review and Discussion

Panelists first reviewed and discussed policy performance level descriptors (PLDs) for the four performance levels, which are as follows:

- 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.

Panelists were instructed to examine the skills in the mastery profiles (also referred to as the Learning Profile) to determine the three performance level cuts that distinguish the four performance levels. Panelists were given a sample profile without mastery shading (example profile shown below in Figure 6.1) and used linkage level statements and available resources to explore the skills described in each cell of the profile. Panelists also studied the grade-specific PLDs for each grade. For these activities, panelists were provided with policy performance level descriptions, specific science performance level descriptions for grade 4, extended linkage level descriptors, blueprints of science EEs, and a glossary of relevant terms.

End of Year Learning Profile

SUBJECT: Science
MODEL: Year-End

GRADE: Elementary science
PROFILE ID: 0

YEAR: 2018-19
TOTAL LL: 0

Essential Element	Level Mastery 1	2	3 (Target)
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
SCI.5.PS.2.1	Recognize the direction objects go when dropped	Predict the direction objects go when dropped	Demonstrate that gravity is directed down
SCI.5.PS.3.1	Identify models that show plants need sunlight to grow	Model plants capturing energy from sunlight	Model energy in food coming from the Sun
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.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

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

Page 1 of 1

Figure 6.1. Example blank learning profile

Panelists set the cut points distinguishing each level starting with Approaching the Target/At Target, then for At Target/Advanced, and finally, for Emerging/Approaching the Target. Panelists set each cut point using the same procedures.

To set each cut point, panelists first examined profiles based on the proposed cut point and profiles at one and two linkage levels down from the proposed cut. Panelists privately reported if they agreed or disagreed with the proposed cut point. A group discussion then took place, where panelists either supported the proposed cut point or an alternate cut point using content-based rationales until

consensus was reached.

6.1.3.2. Panel-Recommended Cut Points and Impact Data

Staff provided the panelists with impact data associated with the cut points determined by group consensus. The impact data provided panelists with the percentage of students achieving at each performance level based on their recommended cut point values. The panel discussed the set of results and indicated one final time if they were in agreement or disagreement with the panel-recommended set of cut points.

6.1.3.3. Standard-Setting Evaluation

At the end of the standard-setting process, all panelists responded to a survey evaluation concerning the training, the panel process, and the resulting cut points by indicating their level of agreement with specific statements.

6.1.4. Results

This section summarizes the panel-recommended cut points, the impact data, and the evaluation results.

6.1.4.1. Panel-Recommended and Proposed Cut Points

Table 6.3 provides the grade 3 cut points recommended by panelists during the standard-setting process and the original proposed cut points for grade 3. The existing grades 4 and 5 cut points are also presented. In all cases, panelists recommended cut points that were lower than the cut points based strictly on impact data.

Table 6.3. Panel-Recommended and Proposed Third-Grade and Existing Fourth- and Fifth-Grade Cut Points

	Performance Level		
Grade	**Emerging/Approaching**	**Approaching/Target**	**Target/Advanced**
3	7 (8)	13 (14)	18 (20)
4	9	15	21
5	10	17	25

Note: Maximum number of linkage levels is 27.

Table 6.4 provides the associated impact data for the panel-recommended grade 3 cuts and existing cuts for grades 4 and 5. The impact data consists only of data based on the current operational administration for spring 2019 (March 11 through May 30, 2019). DLM staff did not include data for students with any untested grade-relevant science EEs.

Table 6.4. Percentage of Students Achieving at Each Science Performance Level Based on Panel-Recommended Third-Grade Cut Points

Performance Level	Grade		
	3 (n = 607)	4 (n = 1,199)	5 (n = 7,057)
Emerging	54.2	63.2	64.7
Approaching	27.4	21.9	21.0
Target	8.9	10.7	13.3
Advanced	9.6	4.2	1.1
Target and Advanced	18.5	14.9	14.4

6.1.4.2. Evaluations of Standard-Setting Process and Results

Panelists responded to a questionnaire after the meeting concluded. The DLM TAC then presented an evaluation of the standard-setting process and results during a conference call.

6.1.5. Panelists Evaluations of Cut Points

Panelists provided diverse ratings during their first review of the proposed cut point values but were able to agree following the group discussion. Panelists had difficulty agreeing to proposed cuts for grade 3 due to concerns that some of the grade-banded EEs were not included in third grade curriculums. Panelists were also mindful of the distance between cut points within and across grades and considered the content complexity of some of the Target level skills assessed in the elementary grade band and students' overall opportunity to learn that content. Panelists ultimately addressed these concerns by lowering all three cut points by either one or two linkage levels.

6.1.6. Panelists Evaluation of Meeting

Appendix C of Nash et al., 2019 contains a summary of the panelist post-meeting questionnaire responses. In the responses, they provided their level of agreement with statements about the standard-setting meeting organization, training, process, and the results. Panelists also were prompted to provide comments to accompany their responses and to discuss any likes or dislikes regarding the meeting. Panelists reported an overall positive experience, agreeing or strongly agreeing with positive statements concerning the meeting and the overall evaluation of the standard-setting process. Panelists reported an increase in understanding of DLM assessments, and that they appreciated the input from DLM staff to further their understanding of the standard-setting process. Panelists also appreciated the ability to use voice or text during the meeting, the ability to share all opinions, as well as the organization and pacing of the meeting.

6.1.7. Technical Advisory Committee Member Observation

The DLM TAC member who observed the grade 3 standard-setting meeting reported a minor concern that the nature of a virtual meeting may have potentially caused a reduction of full conversations and engagement of the panelists. The observer was also concerned that since two of the panelists were from the same school, there was a reduction in member state representation, which

may have had a potential impact on process. Through discussion, the TAC confirmed that the constraints of a virtual meeting were necessary in this application and were consistent with previous recommendations for how to conduct the standard-setting process. They also deemed the standard-setting process itself to be reasonable.

6.2. Standard Setting Grade 7

Because the operational cut points in grades 6 and 8 were either consecutive numbers or only one number apart, the DLM TAC recommended that cut points for grade 7 could be determined without a panel process.

6.2.1. Procedures

DLM staff reviewed existing cut points and identified proposed cut point values that were either the midpoint between the two adjacent grade cut points or equal to the cut at the grade level below, when necessary. The impact data were based on data from 2019 only, collected from March 11 through May 30, 2019. The data were filtered to include only students who did not have any untested grade-relevant science EEs.

6.2.2. Results

This section summarizes the grade 7 cut points and associated impact data.

6.2.2.1. Cut Points and Impact Data for Grade 7

For the Emerging/Approaching the Target and Approaching the Target/At Target cut points, the cuts for the grade level below were chosen to allow students more opportunity to achieve a higher performance level. Table 6.5 displays the cut points for grade 7. Table 6.6 displays the impact data associated with those cut points.

Table 6.5. Seventh-Grade and Adjacent Grade-Band Cut Points

	Performance Level		
Grade	**Emerging/Approaching**	**Approaching/Target**	**Target/Advanced**
6	9	15	21
7	9	15	22
8	10	16	23

Table 6.6. Percentage of Students Achieving at Each Science Performance Level Based on Seventh-Grade Cut Points

Performance Level	Grade		
	6 (*n* = 694)	7 (*n* = 650)	8 (*n* = 7,503)
Emerging	54.2	63.2	64.7
Approaching	27.4	21.9	21.0
Target	8.9	10.7	13.3
Advanced	9.6	4.2	1.1
Target and Advanced	18.5	14.9	14.4

6.3. Review of Results and Final Acceptance

DLM staff gathered grade 3 and 7 cut point recommendations, panelist evaluation responses, and impact data for the DLM TAC to review. State partners then reviewed the results and the TAC feedback.

Following a period of internal state education agency review in May 2019, state partners voted on acceptance of cut points for the consortium. This step did not imply state adoption of the cuts; DLM member states are free to use their own cut points or those adopted by the consortium. States completed their own procedures for formally adopting the cuts. The TAC voted to approve the memorandum summarizing the methods used in the grade 3 standard setting and provided commentary on the standard-setting process (see Appendix D of *Technical Report 19-02* for details).

6.4. Future Steps

DLM staff will create PLDs specific to grades 3 and 7 based on the existing PLDs in the adjacent grades while taking into consideration remarks made by the grade 3 panelists during the standard-setting event concerning the critical skills and understandings needed for each performance level. Using the same procedure for developing other grade-level PLDs, the test development team will draft grade 3 and grade 7 PLDs. The drafts will be finalized after incorporating feedback from reviews as well as input from the partner states.

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 2018–2019 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 2019. Finally, this chapter describes updates made to score reports and data files during spring 2019. 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 2019, science assessments were administered to 37,819 students in 16 states and one Bureau of Indian Education (BIE) school. Counts of students assessed in each state and BIE are displayed in Table 7.1. The assessments were administered by 15,414 educators in 9,065 schools and 3,371 school districts.

Table 7.1. Student Participation by State (N = 37,819)

State	Students (*n*)
Alaska	227
Arkansas	3,849
Delaware	472
District of Columbia	184
Illinois	4,873
Iowa	980
Kansas	1,172
Maryland	2,449
Miccosukee Indian School	8
Missouri	2,858
New Hampshire	356
New Jersey	4,554
New York	9,127
Oklahoma	2,417
Rhode Island	418
West Virginia	737
Wisconsin	3,138

Table 7.2 summarizes the number of students assessed in each grade and course. More than 12,000 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) almost 13,400 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 = 37,819)

Grade	Students (*n*)
3	650
4	4,191
5	7,391
6	748
7	718
8	10,748
9	4,385
10	1,697
11	6,793
12	297
Biology	201

Table 7.3 summarizes the demographic characteristics of the students who participated in the spring 2019 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).

Table 7.3. Demographic Characteristics of Participants (N = 37,819)

Subgroup	*n*	%
Gender		
Male	25,157	66.5
Female	12,662	33.5
Race		
White	22,722	60.1
African American	8,875	23.5
Two or more races	3,330	8.8
Asian	1,736	4.6
American Indian	882	2.3
Native Hawaiian or Pacific Islander	193	0.5
Alaska Native	81	0.2
Hispanic ethnicity		
No	30,832	81.5
Yes	6,987	18.5
English learner (EL) participation		
Not EL eligible or monitored	35,453	93.7
EL eligible or monitored	2,366	6.3

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 7,098 students took at least one instructionally embedded testlet during the 2018–2019 academic year.

Table 7.4. Students Completing Instructionally Embedded Science Testlets by State (N = 7,098)

State	*n*
Arkansas	2,629
Delaware	14
Illinois	4
Iowa	761
Kansas	1,229
Missouri	2,391
New York	18
Oklahoma	51
West Virginia	1

Table 7.5 summarizes the number of instructionally embedded test sessions taken in science. Across

all states, students took 57,026 total testlets during the instructionally embedded window.

Table 7.5. Number of Instructionally Embedded Science Test Sessions, by Grade or Course (N = 57,026)

Grade	*n*
3	3,104
4	3,032
5	12,053
6	3,645
7	3,715
8	11,990
9	3,411
10	5,869
11	8,250
12	1,949
Biology	8

7.2. Student Performance

Student performance on DLM assessments is interpreted using cut points, determined during standard setting, which separate student results 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 2019 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 2019 administration for science.

The spring 2019 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 students who achieved at the At Target or Advanced levels

ranged from approximately 14% to 24%; in middle school the range was 22% to 26%; and in high school and end-of-instruction biology, the percentages ranged from 5% to 25%.

Table 7.6. Percentage of Students by Grade and Performance Level

Grade	Emerging (%)	Approaching (%)	Target (%)	Advanced (%)	Target+ Advanced (%)
3 (n = 650)	56.3	26.2	8.3	9.2	17.5
4 (n = 4,191)	57.0	19.4	15.4	8.2	23.6
5 (n = 7,391)	65.7	20.4	12.8	1.0	13.9
6 (n = 748)	54.0	21.9	18.6	5.5	24.1
7 (n = 718)	51.5	22.8	20.8	4.9	25.6
8 (n = 10,748)	53.2	24.6	19.5	2.7	22.2
9 (n = 4,385)	47.5	27.2	18.9	6.4	25.4
10 (n = 1,697)	57.8	27.0	13.3	1.9	15.2
11 (n = 6,793)	54.2	27.5	14.4	3.8	18.2
12 (n = 297)	79.8	15.2	4.0	1.0	5.1
Biology (n = 201)	62.2	19.4	13.9	4.5	18.4

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.

Table 7.7. Performance Level Distributions, by Demographic Subgroup (N = 37,819)

Subgroup	Emerging		Approaching		Target		Advanced	
	n	%	*n*	%	*n*	%	*n*	%
Gender								
Male	13,907	55.3	5,993	23.8	4,227	16.8	1,030	4.1
Female	7,300	57.7	3,077	24.3	1,882	14.9	403	3.2
Race								
White	12,657	55.7	5,541	24.4	3,724	16.4	800	3.5
African American	4,937	55.6	2,130	24.0	1,412	15.9	396	4.5
Two or more races	1,957	58.8	787	23.6	486	14.6	100	3.0
Asian	1,117	64.3	343	19.8	219	12.6	57	3.3
American Indian	389	44.1	216	24.5	210	23.8	67	7.6
Native Hawaiian or Pacific Islander	93	48.2	42	21.8	46	23.8	12	6.2
Alaska Native	57	70.4	11	13.6	12	14.8	1	1.2
Hispanic ethnicity								
No	17,336	56.2	7,427	24.1	4,947	16.0	1,122	3.6
Yes	3,871	55.4	1,643	23.5	1,162	16.6	311	4.5
English learner (EL) participation								
Not EL eligible or monitored	20,047	56.5	8,524	24.0	5,599	15.8	1,283	3.6
EL eligible or monitored	1,160	49.0	546	23.1	510	21.6	150	6.3

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 35% of the time. The percentage of students who mastered as high as the Target linkage level ranged from approximately 16% in grade 12 to 46% in grade 9.

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

	Linkage Level			
Grade	No evidence (%)	Initial (%)	Precursor (%)	Target (%)
3 (n = 650)	9.5	35.2	19.5	35.7
4 (n = 4,191)	5.3	34.1	16.0	44.6
5 (n = 7,391)	5.4	37.0	18.0	39.6
6 (n = 748)	9.1	17.9	32.1	40.9
7 (n = 718)	12.3	18.0	26.6	43.2
8 (n = 10,748)	5.5	15.9	33.7	44.9
9 (n = 4,385)	6.3	20.8	27.3	45.7
10 (n = 1,697)	7.5	25.4	33.9	33.2
11 (n = 6,793)	5.5	24.9	31.5	38.1
12 (n = 297)	26.3	34.3	22.9	16.5
Biology (n = 201)	3.5	38.3	25.4	32.8

7.3. Data Files

Data files were made available to DLM state partners following the spring 2019 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 included all students who exited at any point during the academic year. In the event of observed incidents during assessment delivery, state partners are provided with an Incident File describing students impacted.

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 via Educator Portal. The final GRF was used to generate score reports.

In addition to the GRF and its supplemental files, states were provided with two additional de-identified data files: a teacher survey data file and a test administration observations data file. The teacher survey file provided state-specific teacher survey responses, with all identifying information about the student and educator removed. The test administration observations file provided test administration observation responses with any identifying information removed. For more information regarding teacher survey content and response rates, see Chapter 4 of this manual. For more information about test administration observation results, see Chapter 9 of this manual.

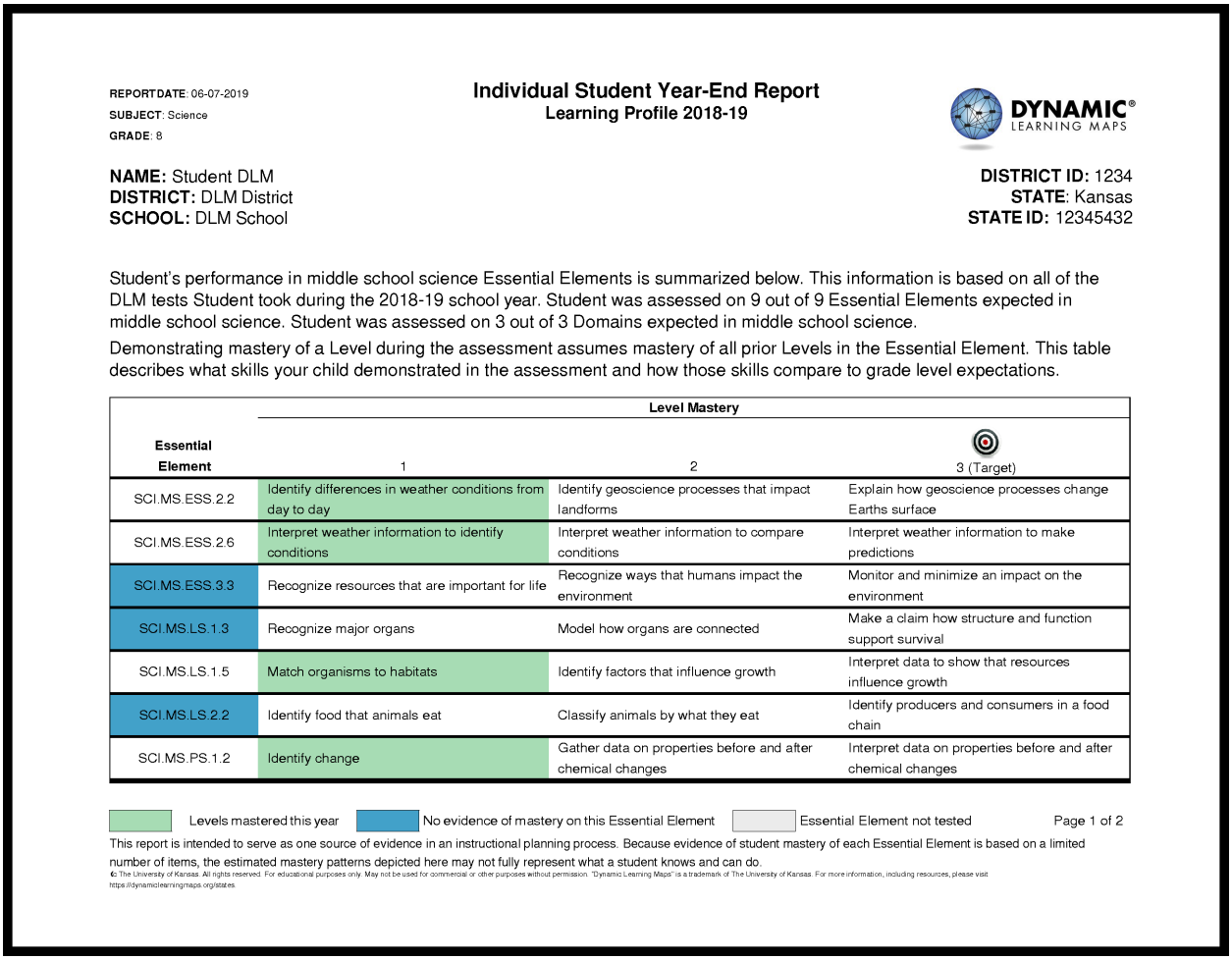
7.4. Score Reports

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 2019. 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 2018–2019 year, minor changes were made to the Individual Student Score Reports. A website was added to the footnote of the report which linked to additional resources related to the DLM assessment and understanding student results. On the Performance Profile portion of the report, a text description of the bar graphs was added to aid in interpretation. On the Learning Profile portion of the report, a cautionary statement was added to the footer to also aid in interpretation of results.

A sample Learning Profile reflecting the 2019 changes is provided in Figure 7.1. A sample Performance Profile portion of the report reflecting the 2019 changes is provided in Figure 7.2.



REPORTDATE: 06-07-2019
SUBJECT: Science
GRADE: 8

Individual Student Year-End Report
Learning Profile 2018-19

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

DISTRICT ID: 1234
STATE: Kansas
STATE ID: 12345432

Student's performance in middle school science Essential Elements is summarized below. This information is based on all of the DLM tests Student took during the 2018-19 school year. Student was assessed on 9 out of 9 Essential Elements expected in middle school science. Student was assessed on 3 out of 3 Domains expected in middle school science.

Demonstrating mastery of a Level during the assessment assumes mastery of all prior Levels 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.MS.ESS.2.2	Identify differences in weather conditions from day to day	Identify geoscience processes that impact landforms	Explain how geoscience processes change Earths surface
SCI.MS.ESS.2.6	Interpret weather information to identify conditions	Interpret weather information to compare conditions	Interpret weather information to make predictions
SCI.MS.ESS.3.3	Recognize resources that are important for life	Recognize ways that humans impact the environment	Monitor and minimize an impact on the environment
SCI.MS.LS.1.3	Recognize major organs	Model how organs are connected	Make a claim how structure and function support survival
SCI.MS.LS.1.5	Match organisms to habitats	Identify factors that influence growth	Interpret data to show that resources influence growth
SCI.MS.LS.2.2	Identify food that animals eat	Classify animals by what they eat	Identify producers and consumers in a food chain
SCI.MS.PS.1.2	Identify change	Gather data on properties before and after chemical changes	Interpret data on properties before and after chemical changes

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

Page 1 of 2

This report is intended to serve as one source of evidence in an instructional planning process. Because evidence of student mastery of each Essential Element is based on a limited number of items, the estimated mastery patterns depicted here may not fully represent what a student knows and can do.

Figure 7.1. Example page of the Learning Profile for spring 2019.

REPORT DATE: 06-07-2019
SUBJECT: Science
GRADE: 8

Individual Student Year-End Report
Performance Profile 2018-19

DYNAMIC® LEARNING MAPS

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

DISTRICT ID: 1234
STATE: DLM State
STATE ID: 12345432

Performance Profile, continued

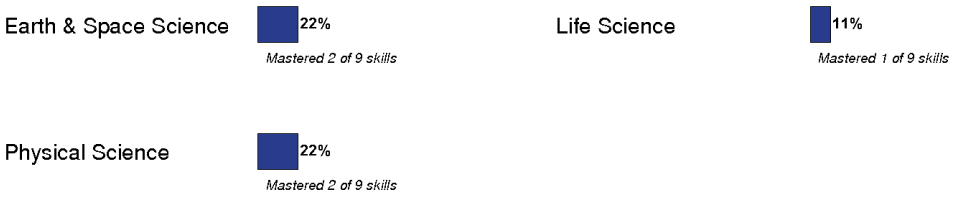
- identify foods that animals eat

In earth and space science, the student can

- interpret basic weather symbols
- compare differences in basic weather conditions

Domain

Bar graphs summarize the percent of skills mastered by domain. Not all students test on all skills due to availability of content at different levels per standard.



More information about Student's performance on each of the Essential Elements that make up the Domains is located in the Learning Profile.

For more information, including resources, please visit https://dynamiclearningmaps.org/states.

Page 2 of 2

Figure 7.2. Example page of the Performance Profile for spring 2019.

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 2019. 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 2019 administration, five data files were delivered to state partners: GRF, special circumstance code file, exited students file, teacher survey data file, and test administration observations file. Overall, between 5% and 26% 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 2019 administration, so an incident file was not needed. Minor changes were made to score reports to assist in the interpretation of results.

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 2018–2019 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 et al. [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 assessment data to generate simulated examinees. This process guarantees that the simulation takes on characteristics of the DLM operational assessment 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 assessment 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[3] for the items of the testlet, conditional on the profile of linkage level mastery or non-mastery for the student.
3. Score the simulated item responses using the operational DLM scoring procedure, estimating linkage level mastery or non-mastery for the simulated student. See Chapter 5 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017a) for more information.[4]
4. Compare the estimated linkage level mastery or non-mastery to the known values from Step 2 for all linkage levels at which the student was administered items.
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.

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

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

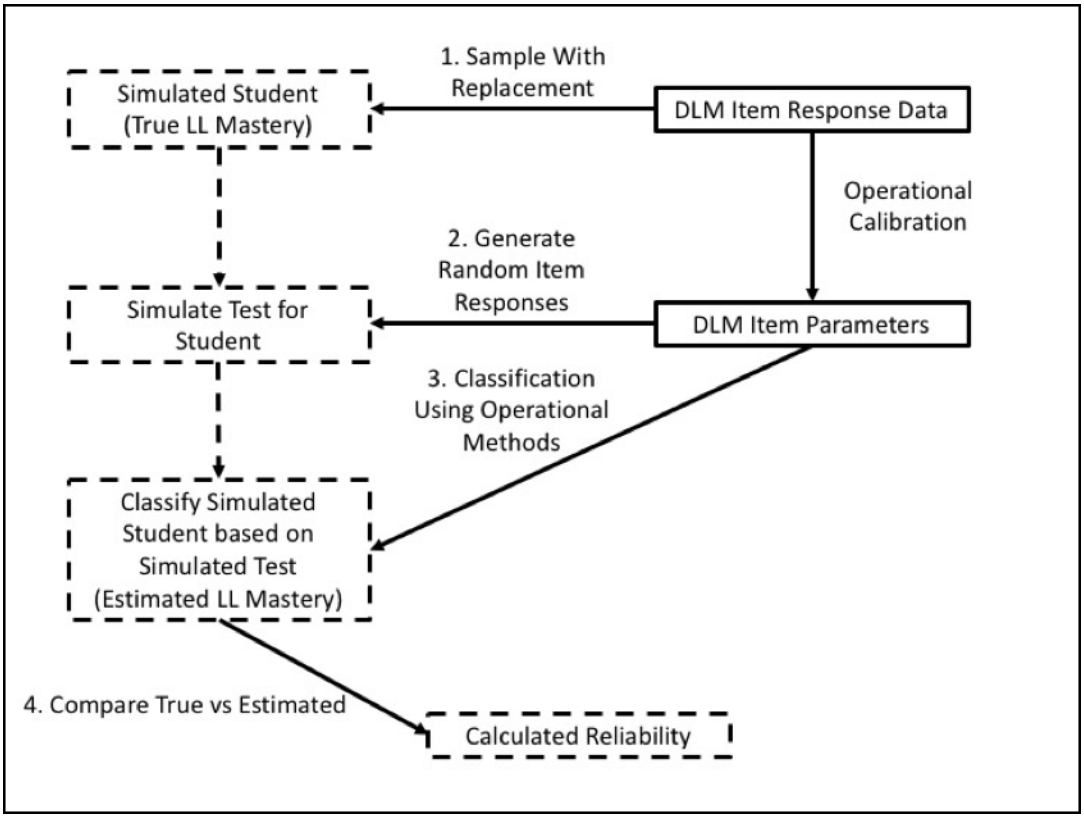


Figure 8.1. Simulation process for creating reliability evidence. *Note.* LL = linkage level.

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[5] 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 .97. Correct classification rates range from .77 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.

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

Table 8.1. Summary of Performance Level Reliability Evidence

Grade	Polychoric correlation	Correct classification rate	Cohen's kappa
3	.965	.849	.886
4	.965	.808	.888
5	.965	.866	.871
6	.939	.768	.846
7	.936	.785	.838
8	.931	.786	.831
9	.965	.815	.879
10	.960	.844	.864
11	.961	.830	.870
12	.971	.900	.874
Biology	.973	.865	.912

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 total linkage levels mastered.

Table 8.2. Summary of Subject Reliability Evidence

Grade	Linkage levels mastered correlation	Average student correct classification	Average student Cohen's kappa
3	.942	.980	.956
4	.949	.976	.947
5	.943	.977	.949
6	.925	.973	.945
7	.930	.972	.945
8	.919	.971	.942
9	.955	.981	.962
10	.946	.984	.970
11	.950	.983	.967
12	.950	.990	.982
Biology	.955	.981	.959

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 .70 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	.776	.994	.991
3	LS	.727	.997	.997
3	PS	.923	.993	.990
4	ESS	.797	.993	.989
4	LS	.695	.997	.996
4	PS	.931	.993	.990
5	ESS	.787	.993	.990
5	LS	.697	.997	.996
5	PS	.928	.993	.990
6	ESS	.769	.993	.990
6	LS	.841	.993	.990
6	PS	.837	.994	.991
7	ESS	.748	.992	.989
7	LS	.846	.993	.990
7	PS	.824	.993	.990
8	ESS	.760	.993	.990
8	LS	.838	.993	.991
8	PS	.823	.993	.991
9	ESS	.851	.994	.991
9	LS	.820	.994	.991
9	PS	.912	.996	.995
10	ESS	.836	.995	.993
10	LS	.804	.995	.993
10	PS	.897	.996	.995
11	ESS	.847	.995	.993
11	LS	.811	.994	.992
11	PS	.903	.996	.996
12	ESS	.826	.996	.995
12	LS	.821	.996	.995
12	PS	.884	.997	.996
Biology	LS1.A	.847	.995	.993
Biology	LS1.B	1.000	.999	.999
Biology	LS2.A	.724	.996	.996
Biology	LS3.B	1.000	.999	.999
Biology	LS4.C	.885	.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 show strong evidence for reliability for the number of linkage levels mastered within EEs.

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	.206	.294	.235	.059
Correct classification rate	<.001	<.001	<.001	<.001	<.001	.176	.588	.206	.029
Cohen's kappa	.029	.088	.059	.118	.324	.118	.206	.059	<.001

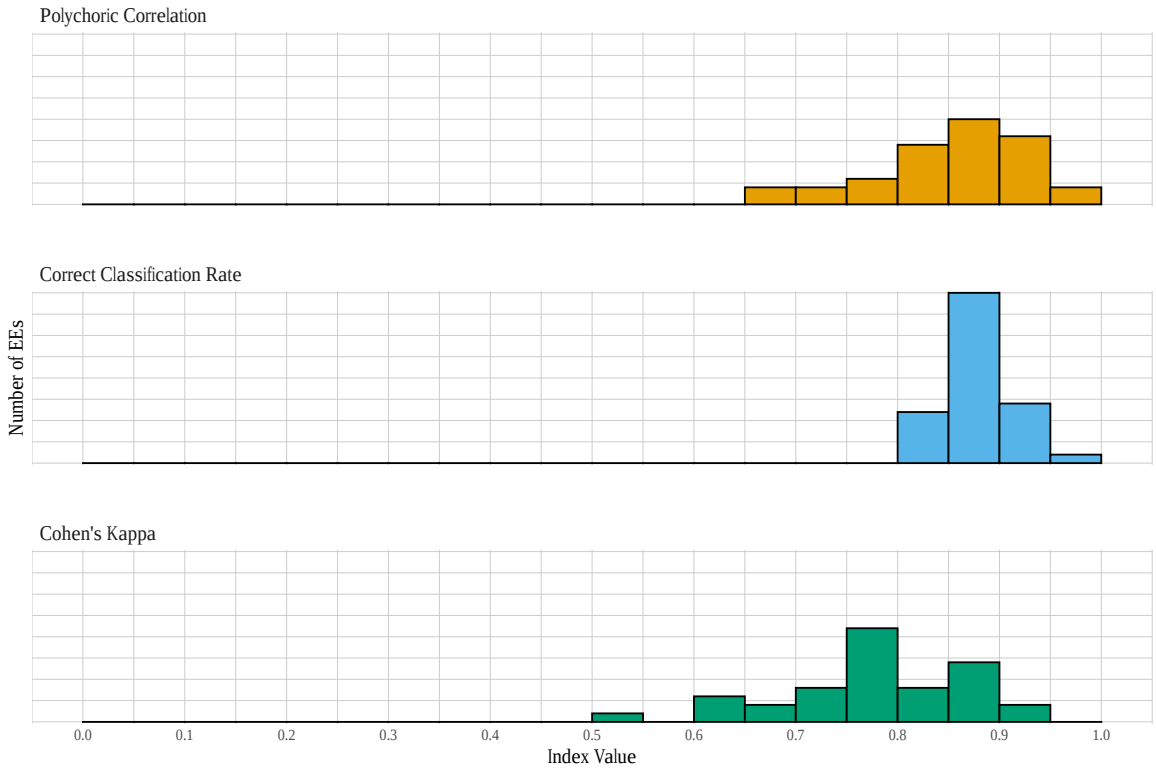


Figure 8.2. Number of linkage levels mastered within EE reliability summaries.

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

For each statistic, figures are given comparing the results of all 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.

[6]The linkage level reliability evidence presented here focuses on consistency of measurement given student responses to items. For more information on how students were assigned linkage levels during assessment, see Chapter 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 12 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	<.001	.029	.010	.029	.098	.196	.608
Correct classification rate	<.001	<.001	<.001	<.001	<.001	.020	.147	.559	.275
Cohen's kappa	.118	.059	.078	.118	.147	.216	.157	.088	.020

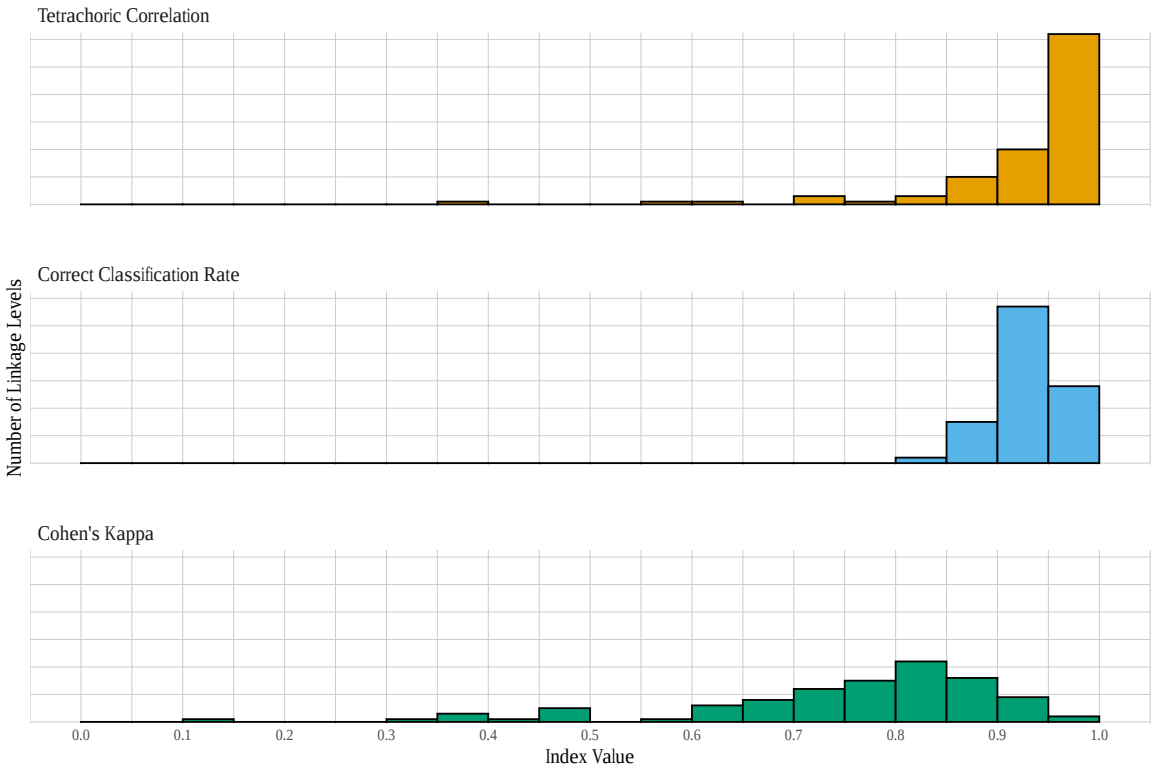


Figure 8.3. Summaries of linkage level reliability.

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 pre-specified ranges of values for the reliability summary statistics. The correlations and correct classification rates generally indicate that all three linkage levels provide reliable classifications of student mastery, with the Initial level demonstrating the most internal consistency across the three reported metrics. Because results were more variable for the Precursor and Target levels, the test development team will evaluate the items

at these linkage levels to determine if changes are needed.

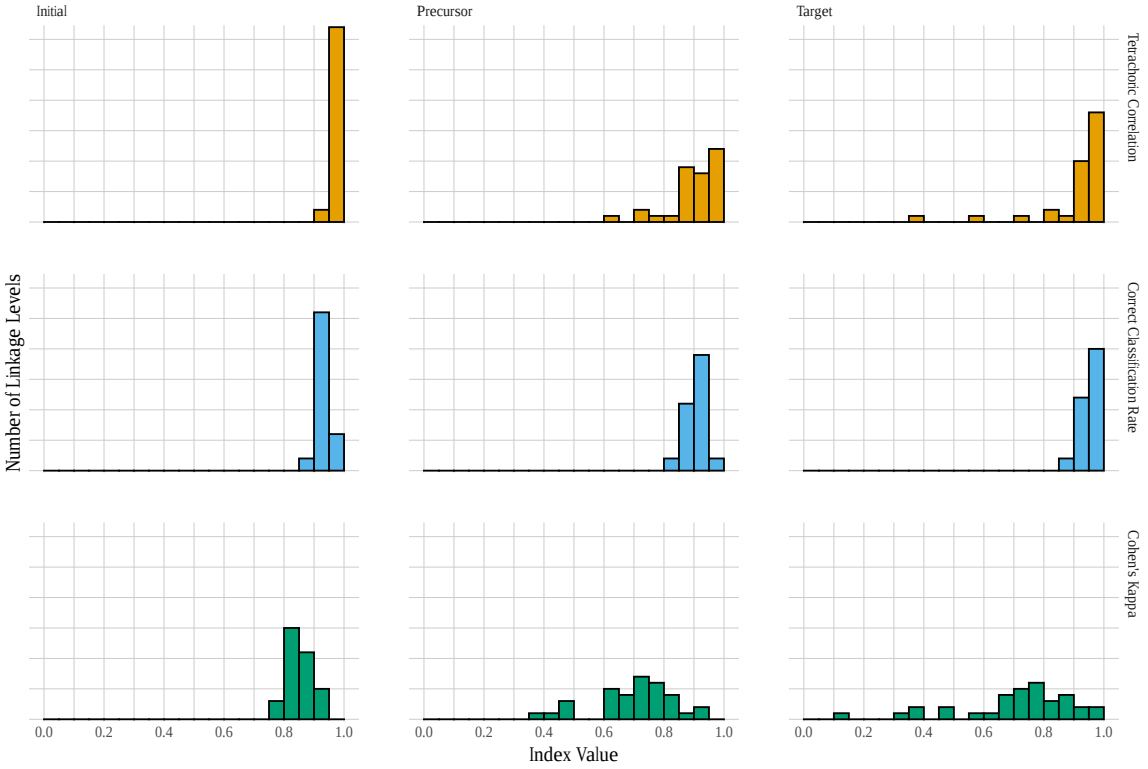


Figure 8.4. Conditional reliability evidence summarized by linkage level.

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. This chapter presents additional evidence collected during 2018–2019 for four of 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, 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, 2018b).

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 2018–2019 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 administration of the spring 2019 operational assessments, teachers were invited to complete a survey about the assessment (see Chapter 4 of this manual for more information on recruitment and response rates). The survey included three blocks of items. The first and third blocks were fixed forms assigned to all teachers. For the second block, teachers received one randomly assigned section.

The first block of the survey served several purposes.[7] One item provided information about the relationship between students' learning opportunities before testing and the test content (i.e., testlets) they encountered on the assessment. The survey asked teachers to indicate the extent to which they judged test content to align with their instruction across all testlets; Table 9.1 reports the results. Approximately 55% of responses (n = 12,534) 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*	%
1,880	8.3	6,490	28.6	7,974	35.1	4,560	20.1	1,815	8.0

The second block of the survey was randomly spiraled so that teachers received one randomly assigned section. In one of the randomly assigned sections, a subset of teachers were asked to indicate the approximate number of hours they spent instructing students on each of the DLM

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

science core ideas and in the science and engineering practices. Teachers responded using a five-point scale: *0-5 hours*, *6-10 hours*, *11-15 hours*, *16-20 hours*, or *more than 20 hours*. Table 9.2 and Table 9.3 indicate the amount of instructional time spent on 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 *0-5 hours*.

Table 9.2. Instructional Time Spent on Science Core Ideas

	Number of hours									
	0-5		6-10		11-15		16-20		>20	
Core Idea	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Physical Science										
Matter and its interactions	1,000	44.5	488	21.7	310	13.8	212	9.4	239	10.6
Motion and stability: Forces and interactions	1,118	49.9	464	20.7	289	12.9	178	7.9	192	8.6
Energy	996	44.7	514	23.1	311	14.0	187	8.4	218	9.8
Life Science										
From molecules to organisms: Structures and processes	1,169	52.6	443	19.9	255	11.5	176	7.9	178	8.0
Ecosystems: Interactions, energy, and dynamics	893	40.0	498	22.3	347	15.5	235	10.5	261	11.7
Heredity: Inheritance and variation of traits	1,319	59.2	391	17.5	224	10.0	143	6.4	152	6.8
Biological evolution: Unity and diversity	1,242	56.0	427	19.3	239	10.8	151	6.8	159	7.2
Earth and Space Science										
Earth's place in the universe	1,037	46.4	471	21.1	334	15.0	187	8.4	204	9.1
Earth's systems	1,034	46.3	471	21.1	333	14.9	188	8.4	205	9.2
Earth and human activity	939	42.0	507	22.7	356	15.9	209	9.4	223	10.0

Table 9.3. Instructional Time Spent on Science and Engineering Practices

	Number of hours									
	0-5		6-10		11-15		16-20		>20	
Science and engineering practice	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Developing and using models	1,088	48.5	516	23.0	277	12.3	185	8.2	178	7.9
Planning and carrying out investigations	965	43.1	555	24.8	305	13.6	217	9.7	196	8.8
Analyzing and interpreting data	871	38.9	529	23.6	345	15.4	233	10.4	262	11.7
Using mathematics and computational thinking	814	36.4	502	22.4	312	13.9	251	11.2	358	16.0
Constructing explanations and designing solutions	1,063	47.6	495	22.2	303	13.6	181	8.1	192	8.6
Engaging in argument from evidence	1,209	54.2	441	19.8	255	11.4	163	7.3	161	7.2
Obtaining, evaluating, and communicating information	886	39.6	504	22.6	328	14.7	244	10.9	273	12.2

Results from the teacher survey were also correlated with total linkage levels mastered by grade band. The median of instructional time was calculated for each student across from teacher responses at the core idea level. While a direct relationship between amount of instructional time and the total number of linkage levels mastered 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 would also have spent more time in instruction. More evidence is needed to evaluate this assumption.

Table 9.4 summarizes the Spearman rank-order correlations between instructional time and the total number linkage levels mastered, by grade band and course. Correlations ranged from .17 to .19. Based on guidelines from Cohen (1988), the observed correlations were small.

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

Grade Band	Correlation with instructional time
Elementary	0.18
Middle School	0.17
High School	0.18
Biology	0.19

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 2019 regarding students' ability to respond to testlets and test administration observation data collected during 2018–2019. 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 2019, teachers provided feedback via a teacher survey. Survey data that inform evaluations of assumptions regarding response processes include teacher perceptions of students' ability to respond as intended, free of barriers, and with necessary supports available.[8]

One of the fixed-form sections of the spring 2019 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 (85% or greater) 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.

[8]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*	%
This student responded to the items on this assessment to the best of his or her knowledge and ability.	820	3.6	1,450	6.3	11,706	51.2	8,887	38.9	20,593	90.1
This student was able to respond to items regardless of his or her disability, behavior, or health concerns.	1,469	6.4	1,954	8.5	11,499	50.2	7,982	34.8	19,481	85.0
This student had access to all necessary supports in order to participate in the assessment.	551	2.4	715	3.1	10,955	47.8	10,676	46.6	21,631	94.4

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 2018–2019 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.

During the 2018–2019 academic year, the DLM Consortium added a new option for states to use when collecting test administration observation data. In previous years, the DLM Consortium collected observations using paper forms, which were submitted via mail or email, or Qualtrics® surveys completed in a web browser. In 2018–2019, the DLM Consortium also collected observations in a new mobile application, Kite Collector. The application allows state and local education agency staff to collect observation data electronically using smart phones and tablets.

The Kite Collector mobile application allows observers to collect data offline without internet access in a testing location. Observers can then later upload their observations using the mobile application when they regain internet access.

In 2018–2019 the total number of observations increased to 140 observations collected by 7 states, a 500% increase compared with the 28 total observations collected by 3 states in 2017–2018.

Table 9.6 shows the number of observations collected by state. Of the observations, 87 (62%) were of computer-delivered assessments and 53 (38%) were of teacher-administered testlets.

Table 9.6. Teacher Observations by State (N = 140)

State	*n*	%
Arkansas	88	62.9
Iowa	7	5.0
Kansas	14	10.0
Missouri	9	6.4
New York	2	1.4
West Virginia	14	10.0
Wisconsin	6	4.3

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 9.7; behaviors were identified as supporting, neutral, or nonsupporting. For example, clarifying directions (73% 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, 59% of observed behaviors were classified as supporting, with 1% of observed behaviors reflecting nonsupporting actions.

Table 9.7. Test Administrator Actions During Computer-Delivered Testlets (n = 87)

Action	*n*	%
Supporting		
Read one or more screens aloud to the student	58	76.3
Clarified directions or expectations for the student	49	73.1
Navigated one or more screens for the student	36	59.0
Repeated question(s) before student responded	32	48.5
Neutral		
Used pointing or gestures to direct student attention or engagement	38	64.4
Used verbal prompts to direct the student's attention or engagement (e.g. "look at this.")	37	62.7
Asked the student to clarify or confirm one or more responses	12	21.1
Used materials or manipulatives during the administration process	19	32.8
Allowed student to take a break during the testlet	6	10.3
Repeated question(s) after student responded (gave a second trial at the same item)	6	10.3
Nonsupporting		
Physically guided the student to a response	1	1.8
Reduced the number of answer choices available to the student	1	1.8

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 59% 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 9.8. Independent response selection was observed in 93% 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 Student Portal was seen in 9% 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 9.8. Student Actions During Computer-Delivered Testlets (n = 87)

Action	*n*	%
Selected answers independently	67	93.1
Navigated screens independently	46	62.2
Selected answers after verbal prompts	36	59.0
Navigated screens after verbal prompts	29	45.3
Navigated screens after TA pointed or gestured	24	38.7
Revisited one or more questions after verbal prompt(s)	6	10.3
Used materials outside of Kite Student Portal to indicate responses to testlet items	5	8.9
Independently revisited a question after answering it	4	7.1
Skipped one or more items	2	3.5

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 53 observations of teacher-administered testlets, observers noted difficulty in four cases (8%). For computer-delivered testlets, evidence to evaluate the assumption was collected by noting students indicating responses to items using varied response modes such as eye gaze (2%) and using manipulatives or materials outside of Kite Student Portal (9%). Additional evidence for this assumption was gathered by observing whether students were able to complete testlets. Of the 140 test administration observations collected, students completed the testlet in 138 cases (99%).

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 9.9 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 9.9. Primary Response Mode for Teacher-Administered Testlets (n = 53)

Response mode	*n*	%
Gestured to indicate response to TA who selected answers	13	24.5
Used computer/device to respond independently	11	20.8
Verbally indicated response to TA who selected answers	8	15.1
Used switch system to respond independently	1	1.9
Eye-gaze system indication to TA who selected answers	0	0.0
No response	24	45.3

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 24 of 87 (28%) observations of computer-delivered testlets, the test administrator entered responses on the student's behalf. In 23 (96%) of those cases, observers indicated that the entered response matched the student's response, while one observer left the item blank.

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 2019 followed the same procedure used in previous years, including data from 2015–2016 through 2017–2018 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, 551 (98%) items on science testlets were selected. In total, 157 were evaluated in the elementary school grade band, 169 items in the middle school grade band, 164

items in the high school grade band, and 61 items in the biology end-of-instruction assessment. Item sample sizes ranged from 269 to 12,826.

Of the 13 items that were not included in the DIF analysis, 11 (85%) had a focal group sample size of less than 100 and 2 (15%) had an item *p*-value greater than .95. Table 9.10 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 9.10. 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	4	36.4	0	0.0	0	0.0
Precursor	7	63.6	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 non-uniform 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.

9.3.1.2. Results

9.3.1.2.1. Uniform DIF Model

A total of 73 items were flagged for evidence of uniform DIF when comparing M_0 to M_1. Table 9.11 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 3% to 18%.

Table 9.11. 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	20	157	12.7	0
Middle	30	169	17.8	0
High	21	164	12.8	0
Biology	2	61	3.3	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 113 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 9.12 summarizes the number of items flagged by grade band or course. The percentage of items flagged for each grade band or course ranged from 13% to 24%.

Table 9.12. 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	38	157	24.2	0
Middle	38	169	22.5	0
High	29	164	17.7	0
Biology	8	61	13.1	1

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, one item had a moderate change in effect size, zero had a large change in effect size, and the remaining 112 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 9.13. 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 9.13. 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*
50061	Biology	HS.LS.2.1	12.97	<.01	1.34	-0.34	.04	A	B

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 A. Clark et al., 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 moderate 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. Chapter 5 also provides a description of the linkage level discrimination (i.e., the ability to differentiate between masters and non-masters).

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%) linkage levels, masters had a greater than .5 chance of providing a correct response to items. Additionally, for 100 (98%) linkage levels, masters had a greater than .6 chance of providing a correct response, compared to only 0 (<1%) linkage levels where masters had a less than .4 chance of providing a correct response. Similarly, for 80 (78%) linkage levels, non-masters had a less than .5 chance of providing a correct response to items. For most linkage levels (n = 56; 55%) non-masters had a less than .4 chance of providing a correct response; however, for 3 (3%) linkage levels, non-masters had a greater than .6 chance of providing a correct response. Finally, 70 (69%) linkage levels had discrimination index of greater than .4, indicating that linkage levels are largely able to discriminate between master and non-masters.

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, 564 (100%) 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 Relation to Other Variables

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

9.4.1. Teacher Ratings on First Contact Survey

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

For each academic item on the First Contact survey, test development teams reviewed the science linkage levels to identify linkage levels that measured the same skills as the academic items. A summary of the First Contact academic items and the number of linkage levels is provided in Table 9.14. Overall, the test development team identified between 2 and 35 linkage levels that measured the same skill as each academic item from the First Contact survey. Table 9.15 shows the number and percentage of linkage levels from each grade band and course that were identified by the test development team as measuring the same skill as at least one academic item. The percentage of linkage levels on the blueprint that were also measured by the First Contact survey ranged from 73% in biology to 78% in middle school.

[9]More information on the First Contact survey, including calculation of complexity band, can be found in Chapter 3 of DLM Consortium (2017a).

Table 9.14. First Contact Items With Linkage Levels Identified

First Contact Item	Number of linkage levels
Sorts objects or materials by common properties (e.g., color, size, shape)	2
Identifies similarities and differences	18
Recognizes patterns	8
Compares initial and final conditions to determine if something changed	22
Uses data to answer questions	35
Identifies evidence that supports a claim	19
Identifies cause and effect relationships	24
Uses diagrams to explain phenomena	18

Table 9.15. Linkage Levels Measuring the Same Skills as First Contact Survey

Grade Band or Course	Measured Linkage Levels	Total Linkage Levels	\%
Elementary	20	27	74.1
Middle School	21	27	77.8
High School	20	27	74.1
Biology	22	30	73.3

Note.
High School and Biology share nine linkage levels.

9.4.1.1. Relationship Between Mastery and First Contact Ratings

For each linkage level identified by the test development team, a data set was created that included student mastery of the EE and linkage level, as well as First Contact survey responses[10]. Science First Contact items asked teachers to use a four-point scale to indicate how consistently students demonstrated each skill: *almost never* (0–20% of the time), *occasionally* (21–50% of the time), *frequently* (51–80% of the time), or *consistently* (81–100% of the time).

Polychoric correlations were calculated to determine the relationship between the teachers' First Contact ratings and the students' mastery of the linkage levels associated with the First Contact items.

Moderate but positive correlations were expected between First Contact ratings and student mastery of the linkage level for several reasons. The First Contact items were not originally designed to align directly with assessment items. Also, teachers are required to complete the First Contact survey

[10]Students who demonstrated mastery via the two-down rule were not included. See Chapter 5 in this manual for a complete description of the scoring rules

before testlet administration; some teachers complete it at the beginning of the school year. Teachers may choose to update survey responses during the year but do not have to. Therefore, First Contact ratings may reflect student knowledge or understandings before instruction, while linkage level mastery represents end-of-year performance. However, in general, higher First Contact ratings were expected to be associated with student mastery of the linkage level measuring the same skill.

Correlations for First Contact items with linkage level mastery are summarized in Table 9.16. A total of six correlations could not be calculated due to a lack of variance in either the linkage level mastery status for students or First Contact responses from teachers. Across all First Contact academic items, most correlations (>70%) differed significantly from 0.

Table 9.16. Correlations of First Contact Item Response to Linkage Level Mastery

First Contact Item	Linkage Levels (*n*)	*r*			*SE*			% significant
		Min	Max	Median	Min	Max	Median	
Sorts objects or materials by common properties (e.g., color, size, shape)	2	0.06	0.24	0.15	0.02	0.02	0.02	100
Identifies similarities and differences	18	-0.24	0.56	0.24	0.02	0.21	0.03	72
Recognizes patterns	7	-0.05	0.22	0.15	0.02	0.04	0.03	86
Compares initial and final conditions to determine if something changed	21	-0.03	0.51	0.24	0.02	0.21	0.02	76
Uses data to answer questions	34	-0.24	0.44	0.19	0.02	0.24	0.03	62
Identifies evidence that supports a claim	18	-0.19	0.23	0.15	0.02	0.22	0.03	67
Identifies cause and effect relationships	23	-0.19	0.29	0.15	0.02	0.19	0.03	74
Uses diagrams to explain phenomena	17	0.04	0.26	0.11	0.02	0.25	0.03	71

The majority of correlations were based on sample sizes greater than 1,000 (n = 112, 80%). However, there were 27 correlations (19%) that were based on a sample size of less than 200. These correlations were all evaluating the relationship between First Contact items and linkage levels associated with EEs that are only assessed on the end-of-course Biology assessment. The Biology assessment is administered to fewer students relative to the general high school science assessment, which accounts for the smaller observed sample sizes[11]. Small sample size is associated with increased standard errors (Moinester & Gottfried, 2014), which were also observed for these correlations. Furthermore, a

[11]For a description of participation in the 2018–2019 assessment, see Chapter 7 of this manual

negative relationship was observed in 16 instances. Of these, eight were Biology linkage levels with a low sample size, and only two were significantly different from zero. In total, there were two negative correlations that were significantly different than zero with a large sample size. Both correlations were evaluating the Target level of EE SCI.MS.PS.1.2. The linkage level statement for this EE and linkage level is "Interpret data on properties before and after chemical changes." The two First Contact items that were negatively associated with this linkage level were "Identifies similarities and differences" (r = -.24, p = .026) and "Recognize patterns (r = -.05, p = .034). However, test development teams identified five First Contact items that aligned to this linkage level statement, more than any other EE and linkage level. Thus, this linkage level likely utilizes multiple skills, contributing to the negative relationship observed for these First Contact items when examined in isolation.

Overall, 89% (n = 124) of the correlations were positive and 71% (n = 99) were significantly different from 0, indicating generally positive associations between linkage level mastery and First Contact ratings. Results for all correlations are summarized in Figure 9.1.

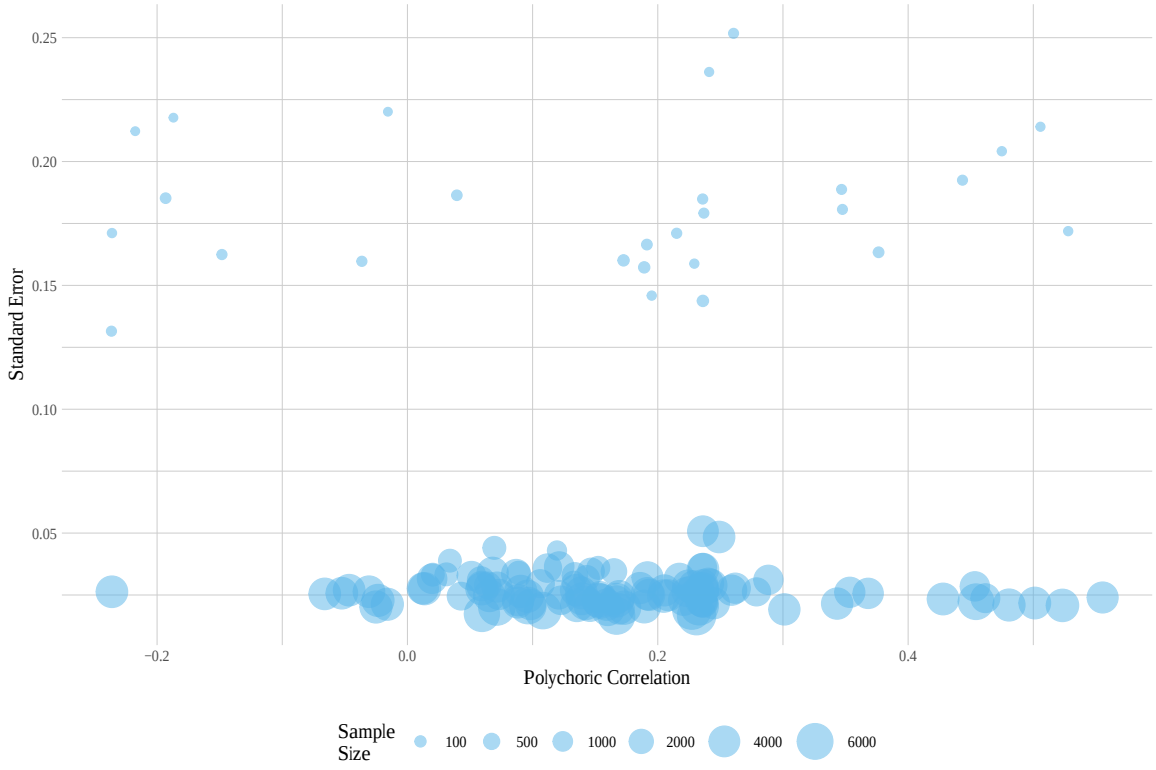


Figure 9.1. Relationship of First Contact responses to linkage level mastery.

This study provides preliminary evidence of the relationship between a portion of the science blueprints with external variables, as indicated by teacher ratings on First Contact academic items. Because the grain size of linkage level statements varies by level and grade band or course, the relationship of linkage level mastery to First Contact rating was expected to be stronger in some areas than in others.

Because this study examined only the subset of EEs and linkage levels associated with First Contact items in the First Contact survey, evidence of the relationship to external variables is available for a portion of the blueprint. An additional study is planned for spring 2020 to begin collecting evidence regarding the relationship between performance and external data for the complete blueprint. See Chapter 11 of this manual for more information.

9.5. 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 2019 via teacher survey responses regarding teacher perceptions of assessment content.

9.5.1. Teacher Perception of Assessment Content

On the spring 2019 survey,[12] teachers were asked two questions about their perceptions of assessment content: whether the content measured important academic skills and knowledge and whether the content reflected high expectations. Table 9.17 summarizes their responses. Teachers generally agreed or strongly agreed that content reflected high expectations for their students (86%) and measured important academic skills (75%).

While the majority of teachers agreed with these statements, 14%-25% 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 et al., 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.

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

Table 9.17. Teacher Perceptions of Assessment Content

Statement	Strongly Disagree		Disagree		Agree		Strongly Agree		Agree + Strongly Agree	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
The content of the assessments measured important academic skills and knowledge for this student.	2,193	9.5	3,540	15.4	13,069	56.8	4,193	18.2	17,262	75.0
The content of the assessments reflected high expectations for this student.	1,067	4.7	2,079	9.1	13,314	58.2	6,411	28.0	19,725	86.2

9.6. Conclusion

This chapter presents additional studies as evidence for the overall validity argument for the DLM Alternate Assessment System. The studies are organized into categories, where available (content, response process, internal structure, 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, 2018b), 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 2018–2019.

11. Conclusion and Discussion

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

The DLM system completed its fourth operational administration year in 2018–2019. This technical manual update provides updated evidence from the 2018–2019 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, 2018b). 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 2018–2019 year
2	Not updated for 2018–2019
3, 4	Provides evidence collected during 2018–2019 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	Provides a brief description of the procedures and results of establishing new grade 3 and 7 cut points for 2018–2019
7, 8	Describes results and analyses from the fourth operational administration, evaluating how students performed on the assessment, the distributions of those results, including aggregated and disaggregated results, and analysis of the consistency of student responses
9	Provides additional studies from 2018–2019 focused on specific topics related to validity
10	Not updated for 2018–2019

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 2018–2019 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 2018–2019

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	9.4	7.1, 7.2, 9.5
2. Achievement level descriptors provide useful information about student achievement.	7.1, 7.2		8.1		7.1, 7.2, 9.5
3. Inferences regarding student achievement can be drawn at the conceptual area level.	7.2, 9.1		8.1		7.2, 9.5
4. Assessment scores provide useful information to guide instructional decisions.					9.5

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 Relation to Other Variables
9.5	9	Evidence Based on Consequences of Testing

11.2. Continuous Improvement

11.2.1. Operational Assessment

As noted previously in this manual, 2018–2019 was the fourth year the DLM Alternate Assessment System was operational. While the 2018–2019 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 2019–2020. This section describes significant changes from the third to fourth year of operational administration, as well as examples of improvements to be made during the 2019–2020 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 2018–2019.

Based on an ongoing effort to improve Kite® system functionality, several changes were implemented during 2018–2019. Educator Portal was enhanced to improve the usability of the online platform. Additionally, a new system was implemented for the collection of test administration observations,

resulting in a larger, more robust sample of observations for evaluating the administration of testlets.

The validity evidence collected in 2018–2019 expands upon the data compiled in the first three operational years for four of the critical sources of evidence as described in *Standards for Educational and Psychological Testing* (AERA et al., 2014): evidence based on test content, internal structure, response process, relation to other variables, 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. Evaluation of item-level bias via differential item functioning analysis, along with item-pool statistics and model parameters, provided additional evidence collected based on internal structure. An analysis of the relationship between First Contact survey items measuring academic skills and linkage level mastery provided evidence based on the relation to other external variables. Teacher-survey responses also provided evidence based on consequences of testing. Studies planned for 2019–2020 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 2019–2020 and beyond. The manual identifies some areas for further investigation.

DLM staff members are planning several studies for spring 2020 to collect data from teachers in the DLM Consortium states. The teacher survey will include a new spiraled block to collect additional information 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 2020 to obtain the fourth year of data for longitudinal survey items as further validity evidence. 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 ethnicity subgroups.

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.

Bechard, S., Clark, A., Swinburne Romine, R., Karvonen, M., Kingston, N., & Erickson, K. (2019). Use of evidence-centered design to develop learning maps-based assessments. *International Journal of Testing*, *19*, 188–205. https://doi.org/10.1080/15305058.2018.1543310[13]

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.

Clark, A. K., Nash, B., Karvonen, M., & Kingston, N. (2017). Condensed mastery profile method for setting standards for diagnostic assessment systems. *Educational Measurement: Issues and Practice*, *36*(4), 5–15. https://doi.org/10.1111/emip.12162[14]

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

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

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

Dynamic Learning Maps Consortium. (2017a). *2015–2016 Technical Manual—Science* (tech. rep.). 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* (tech. rep.). 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* (tech. rep.). University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS). Lawrence, KS.

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

Dynamic Learning Maps Consortium. (2018c). *Educator Portal User Guide* (tech. rep.). University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS). Lawrence, KS.

Dynamic Learning Maps Consortium. (2018d). *Test Administration Manual 2018–2019* (tech. rep.). University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS). Lawrence, KS.

Dynamic Learning Maps Consortium. (2019a). *2018–2019 Technical Manual Update—Integrated Model* (tech. rep.). University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS). Lawrence, KS.

Dynamic Learning Maps Consortium. (2019b). *2018–2019 Technical Manual Update—Year-End Model* (tech. rep.). University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS). Lawrence, KS.

[13] https://doi.org/10.1080/15305058.2018.1543310
[14] https://doi.org/10.1111/emip.12162

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].

Moinester, M., & Gottfried, R. (2014). Sample size estimation for correlations with pre-specified confidence interval. *The Quantitative Methods for Psychology*, *10*, 124–130. https://doi.org/10.20982/tqmp.10.2.p124[15]

Nash, B. Et al. (2016). *2016 Standard Setting: Science* (tech. rep. No. 16-03). University of Kansas, Center for Educational Testing and Evaluation. Lawrence, KS.

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.

Nash, B., Kavitsky, E., & Clark, A. (2019). *Standard setting technical report science: Grades 3 and 7* (tech. rep. No. 19-02). University of Kansas, Accessible Teaching, Learning, and Assessment Systems (ATLAS). 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*, 361–370.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*, 251–275. https://doi.org/10.1007/s00357-013-9129-4[16]

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4[17]

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

[15] https://doi.org/10.20982/tqmp.10.2.p124
[16] https://doi.org/10.1007/s00357-013-9129-4
[17] https://doi.org/10.1007/s11222-016-9696-4

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

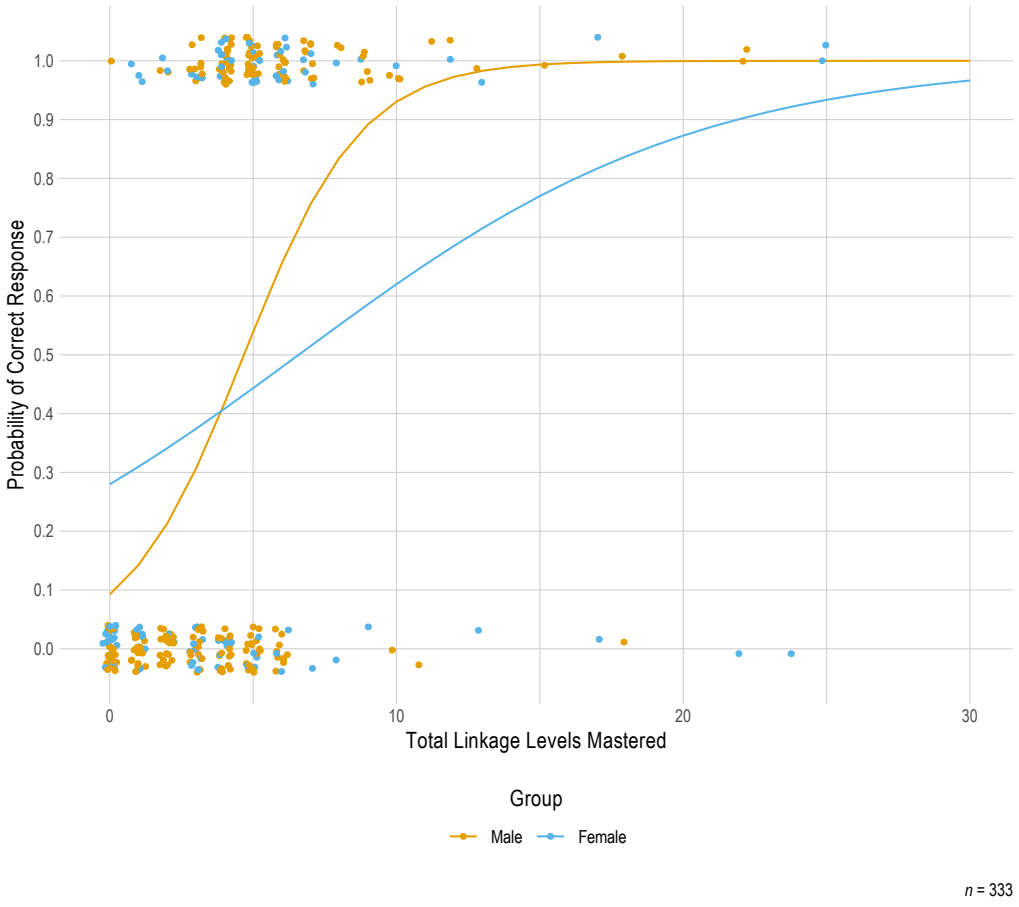
No items had a non-negligible effect-size change when comparing equation (9.2) to equation (9.1). In this model, the probability of a correct response was modeled as a function of ability and gender.

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 50061

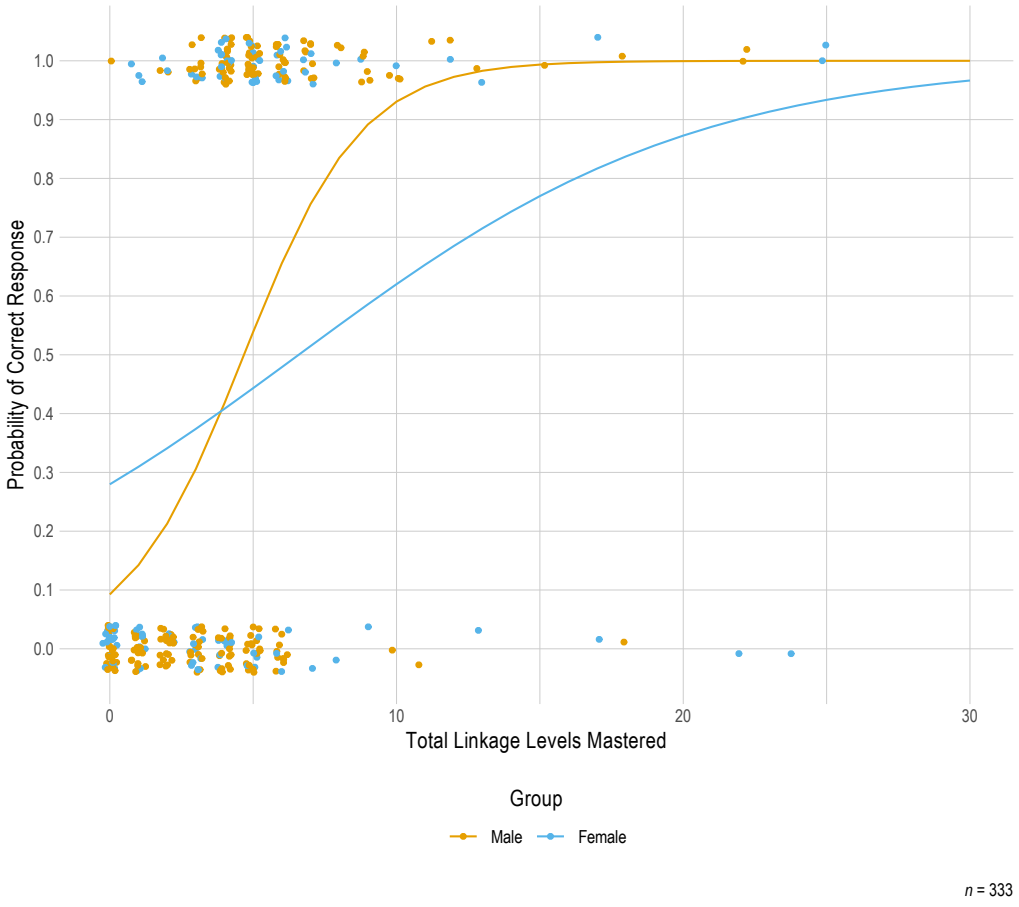
χ^2 = 12.97, p = 0.0015; Nagelkerke's R^2 = 0.04, Zumbo & Thomas: *negligible*, Jodoin & Gierl: *moderate*



n = 333

Item 50061

χ^2 = 12.97, p = 0.0015; Nagelkerke's R^2 = 0.04, Zumbo & Thomas: *negligible*, Jodoin & Gierl: *moderate*



n = 333