



**DYNAMIC**<sup>®</sup>  
LEARNING MAPS

***2021–2022 Technical Manual  
Update***

---

Science

December 2022

**All rights reserved.** Any or all portions of this document may be reproduced and distributed without prior permission provided the source is cited as:

Dynamic Learning Maps Consortium. (2022, December). *2021–2022 Technical Manual Update—Science*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems.

### **Acknowledgements**

The publication of this technical manual update builds on 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.**, *Assistant Director for Psychometrics*

**Amy K. Clark, Ph.D.**, *Associate Director for Operational Research*

**Brooke Nash, Ph.D.**, *Associate Director for Psychometrics*

The authors also wish to acknowledge Ashley Hirt, Jeffrey Hoover, Elizabeth Kavitsky, Jennifer Kobrin, and Noelle Pablo for their role in developing, organizing, and compiling this manual. The authors also wish to acknowledge Brianna Beitling, Amber Cavazos, Alson Cole, Karen Erickson, Zachary Hopper, Sarah Koebley, Jessica Lancaster, Mari Langas, and Delaney Wilson for their contributions to this manual. Finally, the authors wish to thank Kristy Bledsoe, Lucas Cooper, Justin Dean, Aaron Gates, Whitney Lohrenz, and Sara Lundberg for their editing and project management work. 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 on the DLM System. Members of the Technical Advisory Committee during the 2021–2022 operational year include:

**Russell Almond, Ph.D.**, *Florida State University*

**Karla Egan, Ph.D.**, *EdMetric*

**Claudia Flowers, Ph.D.**, *University of North Carolina at Charlotte*

**Robert Henson, Ph.D.**, *University of North Carolina at Greensboro*

**Joan Herman, Ed.D.**, *University of California, Los Angeles*

**James Pellegrino, Ph.D.**, *University of Illinois Chicago*

**Edward Roeber, Ph.D.**, *Michigan Assessment Consortium*

**David Williamson, Ph.D.**, *The College Board*

**Phoebe Winter, Ph.D.**, *Independent Consultant*

# Contents

<b>1 Overview</b>	<b>1</b>
1.1 Current DLM Collaborators for Development and Implementation	1
1.2 Student Population	2
1.3 Assessment	3
1.4 Theory of Action and Interpretive Argument	4
1.5 Key Features	7
1.6 Technical Manual Overview	8
<b>2 Essential Element Development</b>	<b>10</b>
<b>3 Assessment Design and Development</b>	<b>11</b>
3.1 Test Development Procedures	11
3.1.1 Testlet and Item Writing	11
3.1.2 External Reviews	15
3.2 Evidence of Item Quality	20
3.2.1 Field Testing	20
3.2.2 Operational Assessment Items for 2021–2022	24
3.2.3 Evaluation of Item-Level Bias	30
3.3 Conclusion	37
<b>4 Assessment Delivery</b>	<b>38</b>
4.1 Key Features of the Science Assessment Model	38
4.1.1 Assessment Administration Windows	38
4.2 Evidence from the DLM System	39
4.2.1 Administration Time	39
4.2.2 Device Usage	40
4.2.3 Blueprint Coverage	41
4.2.4 Adaptive Delivery	42
4.2.5 Administration Incidents	45
4.2.6 Accessibility Support Selections	45
4.3 Evidence From Monitoring Assessment Administration	46
4.3.1 Test Administration Observations	46
4.3.2 Data Forensics Monitoring	51
4.4 Evidence From Test Administrators	51
4.4.1 User Experience With the DLM System	52
4.4.2 Opportunity to Learn	54
4.5 Conclusion	59
<b>5 Modeling</b>	<b>60</b>
5.1 Psychometric Background	60
5.2 Essential Elements and Linkage Levels	61
5.3 Overview of the DLM Modeling Approach	61
5.3.1 Model Specification	61

5.3.2	Model Calibration .....	63
5.3.3	Estimation of Student Mastery Probabilities .....	65
5.4	Model Evaluation .....	66
5.4.1	Model Fit.....	66
5.4.2	Classification Accuracy .....	67
5.5	Calibrated Parameters.....	68
5.5.1	Probability of Masters Providing Correct Response.....	68
5.5.2	Probability of Nonmasters Providing Correct Response .....	70
5.5.3	Item Discrimination.....	71
5.5.4	Base Rate Probability of Class Membership.....	72
5.6	Conclusion .....	73
<b>6</b>	<b>Standard Setting .....</b>	<b>74</b>
<b>7</b>	<b>Reporting and Results.....</b>	<b>75</b>
7.1	Student Participation.....	75
7.2	Student Performance.....	80
7.2.1	Overall Performance .....	80
7.2.2	Subgroup Performance .....	81
7.3	Mastery Results .....	83
7.3.1	Mastery Status Assignment .....	83
7.3.2	Linkage Level Mastery .....	84
7.4	Data Files.....	85
7.5	Score Reports.....	86
7.5.1	Individual Student Score Reports.....	86
7.6	Quality-Control Procedures for Data Files and Score Reports.....	88
7.7	Conclusion .....	88
<b>8</b>	<b>Reliability .....</b>	<b>90</b>
8.1	Background Information on Reliability Methods .....	90
8.2	Methods of Obtaining Reliability Evidence .....	92
8.2.1	Reliability Sampling Procedure .....	93
8.3	Reliability Evidence.....	94
8.3.1	Linkage Level Reliability Evidence .....	95
8.3.2	Conditional Reliability Evidence by Linkage Level .....	97
8.3.3	Essential Element Reliability Evidence .....	100
8.3.4	Domain and Topic Reliability Evidence .....	102
8.3.5	Subject Reliability Evidence .....	104
8.3.6	Performance Level Reliability Evidence .....	105
8.4	Conclusion .....	106
<b>9</b>	<b>Training and Professional Development .....</b>	<b>108</b>
9.1	Updates to Required Test Administrator Training .....	108
9.2	Instructional Professional Development .....	108
9.2.1	Professional Development Participation and Evaluation.....	109

9.3 Conclusion .....	116
<b>10 Validity Evidence</b> .....	<b>117</b>
10.1 Validity Evidence Summary .....	118
10.2 Continuous Improvement.....	119
10.2.1 Improvements to the Assessment System.....	120
10.2.2 Future Research .....	120
<b>11 References</b> .....	<b>121</b>
<b>A Supplemental Information About Assessment Design and Development</b> .....	<b>126</b>
A.1 Differential Item Functioning Plots .....	126
A.1.1 Uniform Model.....	126
A.1.2 Combined Model .....	127

## List of Tables

3.1	Item Writers' Years of Teaching Experience.....	12
3.2	Item Writers' Grade-Level Teaching Experience .....	12
3.3	Item Writers' Level of Degree.....	12
3.4	Item Writers' Degree Type.....	13
3.5	Item Writers' Experience with Disability Categories .....	13
3.6	Professional Roles of Item Writers.....	14
3.7	Population Density for Schools of Item Writers .....	14
3.8	Demographics of the Item Writers.....	14
3.9	External Reviewers' Years of Teaching Experience .....	16
3.10	External Reviewers' Grade-Level Teaching Experience .....	16
3.11	External Reviewers' Level of Degree .....	16
3.12	External Reviewers' Degree Type.....	17
3.13	External Reviewers' Experience with Disability Categories .....	17
3.14	Professional Roles of External Reviewers .....	18
3.15	Population Density of School of Content Panelists .....	18
3.16	Demographics of the External Reviewers .....	19
3.17	Spring 2022 Field-Test Testlets .....	21
3.18	2021–2022 Operational Testlets, by Grade Band .....	25
3.19	Educator Perceptions of Assessment Content.....	26
3.20	Number of Items Evaluated for Each Race.....	32
3.21	Comparisons Not Included in Differential Item Functioning Analysis for Gender, by Linkage Level.....	32
3.22	Comparisons Not Included in Differential Item Functioning Analysis for Race, by Linkage Level.....	32
3.23	Combinations Flagged for Evidence of Uniform Differential Item Functioning for Gender...	33
3.24	Combinations Flagged for Evidence of Uniform Differential Item Functioning for Race .....	33
3.25	Combinations Flagged for Uniform DIF With Moderate or Large Effect Size.....	34
3.26	Items Flagged for Evidence of Differential Item Functioning for the Combined Model for Gender .....	34
3.27	Items Flagged for Evidence of Differential Item Functioning for the Combined Model for Race .....	35
3.28	Combinations Flagged for DIF With Moderate or Large Effect Size for the Combined Model .....	36
4.1	Distribution of Response Times per Testlet in Minutes .....	40
4.2	Essential Elements Required for Blueprint Coverage .....	41
4.3	Student Blueprint Coverage by Complexity Band .....	42
4.4	Correspondence of Complexity Bands and Linkage Levels.....	42
4.5	Adaptation of Linkage Levels Between First and Second Science Testlets .....	44
4.6	Distribution of Linkage Levels Assigned for Assessment.....	45
4.7	Accessibility Supports Selected for Students .....	46
4.8	DLM Resources for Test Administration Monitoring Efforts.....	47

4.9	Educator Observations by State .....	48
4.10	Test Administrator Actions During Computer-Delivered Testlets .....	49
4.11	Student Actions During Computer-Delivered Testlets .....	50
4.12	Primary Response Mode for Educator-Administered Testlets.....	51
4.13	Test Administrator Responses Regarding Test Administration.....	53
4.14	Test Administrator Perceptions of Student Experience with Testlets.....	54
4.15	Educator Ratings of Portion of Testlets That Matched Instruction .....	55
4.16	Instructional Time Spent on Science Core Ideas .....	56
4.17	Instructional Time Spent on Science and Engineering Practices.....	57
4.18	Correlation Between Instruction Time in Science Linkage Levels Mastered.....	58
4.19	Student Attention Levels During Instruction .....	58
5.1	Depiction of Fungible Item Parameters for Items Measuring a Single Linkage Level.....	62
5.2	Percentage of Models With Acceptable Model Fit ( $ppp > .05$ ).....	67
5.3	Estimated Classification Accuracy by Linkage Level .....	68
7.1	Student Participation by State .....	76
7.2	Student Participation by Grade or Course .....	77
7.3	Demographic Characteristics of Participants .....	78
7.4	Students Completing Instructionally Embedded Science Testlets by State .....	79
7.5	Number of Instructionally Embedded Science Testlets by Grade .....	79
7.6	Percentage of Students by Grade and Performance Level.....	80
7.7	Science Performance Level Distributions by Demographic Subgroup.....	82
8.1	Reliability Summaries Across All Linkage Levels: Proportion of Linkage Levels Falling Within a Specified Index Range .....	96
8.2	Reliability Summaries Across All Linkage Levels: Proportion of Linkage Levels Falling Within a Specified Index Range .....	99
8.3	Reliability Summaries Across All Essential Elements: Proportion of Essential Elements Falling Within a Specified Index Range .....	101
8.4	Summary of Domain and Topic Reliability Evidence.....	103
8.5	Summary of Subject Reliability Evidence.....	105
8.6	Summary of Performance Level Reliability Evidence .....	106
9.1	Number of Professional Development Modules Completed as Part of the Required Test Administrator Training .....	110
9.2	Professional Development Modules Selected for Inclusion in Required Test Administrator Training.....	111
9.3	Number of Self-Directed Modules Completed in 2021–2022 by Educators in DLM States and Other Localities .....	112
9.4	Response Rates and Rate of <i>Agree</i> or <i>Strongly Agree</i> on 2021–2022 Self-Directed Module Evaluation Questions .....	114
10.1	Review of Technical Manual Update Contents.....	117
10.2	DLM Alternate Assessment System Claims and Sources of Updated Evidence for 2021–2022 .....	118
10.3	Evidence Sources Cited in Table 10.2 .....	119

## List of Figures

1.1	Design of the DLM Science Assessment .....	4
1.2	Dynamic Learning Maps Theory of Action .....	6
3.1	<i>p</i> -values for Science Field-Test Items .....	23
3.2	Standardized Difference Z-Scores for Science Field-Test Items.....	24
3.3	<i>p</i> -values for Science 2022 Operational Items .....	27
3.4	Standardized Difference Z-Scores for Science 2021–2022 Operational Items.....	28
3.5	Standardized Difference Z-Scores for 2021–2022 Operational Items by Linkage Level .....	29
4.1	Distribution of Devices Used for Completed Testlets .....	41
5.1	Probability of Masters Providing a Correct Response to Items Measuring Each Linkage Level.....	69
5.2	Probability of Nonmasters Providing a Correct Response to Items Measuring Each Linkage Level.....	71
5.3	Difference Between Masters’ and Nonmasters’ Probability of Providing a Correct Response to Items Measuring Each Linkage Level .....	72
5.4	Base Rate of Linkage Level Mastery .....	73
7.1	Linkage Level Mastery Assignment by Mastery Rule for Each Grade Band or Course .....	84
7.2	Students’ Highest Linkage Level Mastered Across Science Essential Elements by Grade.	85
7.3	Example Page of the Performance Profile With Cautionary Statement for 2021–2022.....	87
7.4	Example Page of the Learning Profile With Cautionary Statement for 2021–2022. ....	88
8.1	Simulation Process for Creating Reliability Evidence .....	94
8.2	Summaries of Linkage Level Reliability .....	97
8.3	Conditional Reliability Evidence Summarized by Linkage Level.....	100
8.4	Number of Linkage Levels Mastered Within Essential Element Reliability Summaries .....	102



## 1. Overview

The Dynamic Learning Maps® (DLM®) Alternate Assessment System assesses student achievement in English language arts (ELA), mathematics, 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], 2022a, 2022b). The purpose of the system is to improve academic experiences and outcomes for students with the most significant cognitive disabilities by setting high and actionable academic expectations and providing appropriate and effective supports to educators. Results from the DLM alternate assessment are intended to support interpretations about what students know and are able to do and to support inferences about student achievement in the given subject. Results provide information that can guide instructional decisions as well as information for use with state accountability programs.

The DLM System is developed and administered by Accessible Teaching, Learning, and Assessment Systems (ATLAS), a research center within the University of Kansas's Achievement and Assessment Institute. The DLM System is based on the core belief that all students should have access to challenging, grade-level or grade-band 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 assessments cannot.

A complete technical manual was created after the first operational administration in 2015–2016. After each annual administration, a technical manual update is provided to summarize updated information. The current technical manual provides updates for the 2021–2022 administration. Only sections with updated information are included in this manual. For a complete description of the DLM assessment system, refer to previous technical manuals, including the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

### 1.1. Current DLM Collaborators for Development and Implementation

The DLM System was initially developed by a consortium of state education agencies (SEAs) beginning in 2010 and expanding over the years, with a focus on ELA and mathematics. The development of a DLM science assessment began with a subset of the participating SEAs in 2014. Due to the differences in the development timelines, separate technical manuals are prepared for ELA and mathematics and science. During the 2021–2022 academic year, DLM assessments were available to students in 21 states: Alaska, Arkansas, Colorado, Delaware, District of Columbia, Illinois, Iowa, Kansas, Maryland, Missouri, New Hampshire, New Jersey, New Mexico, New York, North Dakota, Oklahoma, Pennsylvania, Rhode Island, Utah, West Virginia, and Wisconsin. One SEA partner, Colorado, only administered assessments in ELA and mathematics; one SEA partner, District of Columbia, only administered assessments in science. The DLM Governance Board is comprised of two representatives from the SEAs of each member state. Representatives have expertise in special education and state assessment administration. The DLM Governance Board advises on the administration, maintenance, and enhancement of the DLM System.

In addition to ATLAS and governance board states, other key partners include the Center for Literacy and Disability Studies at the University of North Carolina at Chapel Hill and Agile Technology Solutions at the University of Kansas.

The DLM System is also supported by a Technical Advisory Committee (TAC). DLM TAC members possess decades of expertise, including in large-scale assessments, accessibility for alternate assessments, diagnostic classification modeling, and assessment validation. The DLM TAC provides advice and guidance on technical adequacy of the DLM assessments.

## 1.2. Student Population

The DLM System serves students with the most significant cognitive disabilities, sometimes referred to as students with extensive support needs, who are eligible to take their state’s alternate assessment based on alternate academic achievement standards. This population is, by nature, diverse in learning style, communication mode, support needs, and demographics.

Students with the most significant cognitive disabilities have a disability or multiple disabilities that significantly impact intellectual functioning and adaptive behavior. When adaptive behaviors are significantly impacted, the individual is unlikely to develop the skills to live independently and function safely in daily life. In other words, significant cognitive disabilities impact students in and out of the classroom and across life domains, not just in academic settings. The DLM System is designed for students with these significant instruction and support needs.

The DLM System provides the opportunity for students with the most significant cognitive disabilities to show what they know, rather than focusing on deficits (Nitsch, 2013). These are students for whom general education assessments, even with accommodations, are not appropriate. These students learn academic content aligned to grade-level content standards, but at reduced depth, breadth, and complexity. The content standards are derived from the *Framework for K–12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (Framework, National Research Council, 2012) and the Next Generation Science Standards (NGSS, NGSS Lead States [NGSS], 2013) and are called Essential Elements (EEs). The EEs are the learning targets for elementary, middle school, and high school grade bands and high school Biology. Chapter 2 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017) provides a complete description of the content structures for the DLM assessment, including the EEs.

While all states provide additional interpretation and guidance to their districts, three general participation guidelines are considered for a student to be eligible for the DLM alternate assessment.

1. The student has a significant cognitive disability, as evident from a review of the student records that indicates a disability or multiple disabilities that significantly impact intellectual functioning and adaptive behavior.
2. The student is primarily being instructed (or taught) using the DLM EEs as content standards, as evident by the goals and instruction listed in the IEP for this student that are linked to the enrolled grade level or grade band DLM EEs and address knowledge and skills that are appropriate and challenging for this student.
3. The student requires extensive direct individualized instruction and substantial supports to achieve measurable gains in the grade-and age-appropriate curriculum. The student (a) requires extensive, repeated, individualized instruction and support that is not of a temporary or transient nature and (b) uses substantially adapted materials and individualized methods of accessing information in alternative ways to acquire, maintain, generalize, demonstrate and transfer skills across multiple settings.

The DLM System eligibility criteria also provide guidance on specific considerations that are not acceptable for determining student participation in the alternate assessment:

- a disability category or label
- poor attendance or extended absences
- native language, social, cultural, or economic differences
- expected poor performance on the general education assessment
- receipt of academic or other services
- educational environment or instructional setting
- percent of time receiving special education
- English Language Learner status
- low reading or achievement level
- anticipated disruptive behavior
- impact of student scores on accountability system
- administrator decision
- anticipated emotional duress
- need for accessibility supports (e.g., assistive technology) to participate in assessment

### **1.3. Assessment**

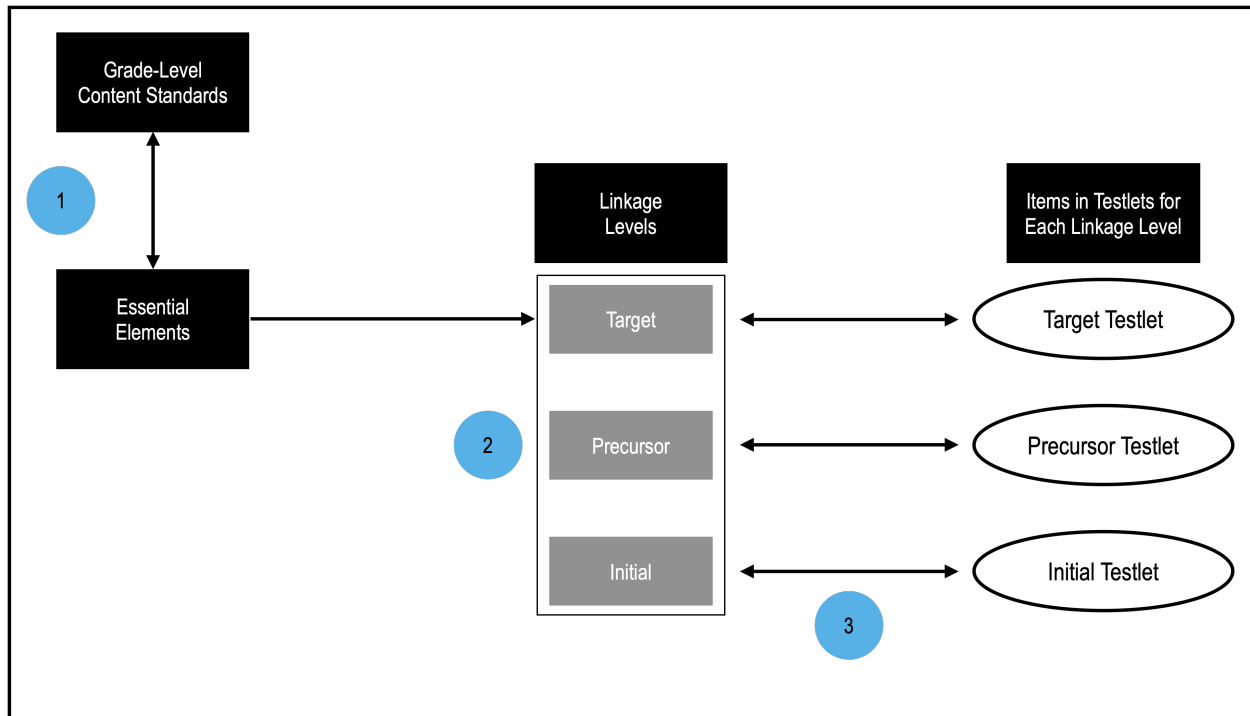
The DLM science assessment is based on EEs for science. The EEs are based on the general education grade-banded content standards but exhibit reduced depth, breadth, and complexity. They link the general education content standards to grade band expectations that are at an appropriate level of rigor and challenge for students with significant cognitive disabilities. The EEs specify the academic content standards and delineate three levels of cognitive complexity: Initial, Precursor, and Target. These levels represent knowledge, skills, and understandings in science that support a progression toward mastery associated with the grade band content standards. Assessment design is based on three key relationships between system elements (see Figure 1.1):

1. Content standards (*Framework*, NGSS) and the DLM science EEs for each grade band
2. An EE and its associated linkage levels
3. Linkage levels and assessment items.

These relationships are further explained in Chapter 3 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

**Figure 1.1**

*Design of the DLM Science Assessment*



For all aspects of the DLM System, our overarching goal is to align with the latest research from a full range of accessibility lenses (e.g., universal design of assessment, physical and sensory disabilities, special education) to ensure the assessments are accessible for the widest range of students who will be interacting with the content. In order to exhibit the assessed skills, students must be able to interact with the assessment in the means most appropriate for them. Thus, the DLM assessments provide different means of student interaction and ensure those means can be used while maintaining the validity of the inferences from and intended uses of the DLM System. These pathways begin in the earliest stages of assessment and content development, from item writing to assessment administration. We seek both content adherence and accessibility by maximizing the quality of the assessment process while preserving evidence of the targeted cognition. Ensuring accessibility for all students supports the validity of the intended uses. The overarching goal of accessible content is reflected in the Theory of Action for the DLM System, which is described in the following section.

### 1.4. Theory of Action and Interpretive Argument

The Theory of Action that guided the design of the DLM System for science was similar to the Theory of Action for the ELA and mathematics assessments, which was formulated in 2011, revised in December 2013, and revised again in 2019. It expresses the belief that high expectations for students with the most significant cognitive disabilities, combined with appropriate educational supports and diagnostic tools for educators, results in improved academic experiences and outcomes for students and educators.

The process of articulating the Theory of Action started with identifying critical problems that characterize

large-scale assessment of students with the most significant cognitive disabilities so that the DLM System design could alleviate these problems. For example, traditional assessment models treat knowledge as unidimensional and are independent of teaching and learning, yet teaching and learning are multidimensional activities and are central to strong educational systems. Also, traditional assessments focus on standardized methods and do not allow various, non-linear approaches for demonstrating learning even though students learn in various and non-linear ways. In addition, using assessments for accountability pressures educators to use assessments as models for instruction with assessment preparation replacing best-practice instruction. Furthermore, traditional assessment systems often emphasize objectivity and reliability over fairness and validity. Finally, negative, unintended consequences for students must be addressed and eradicated.

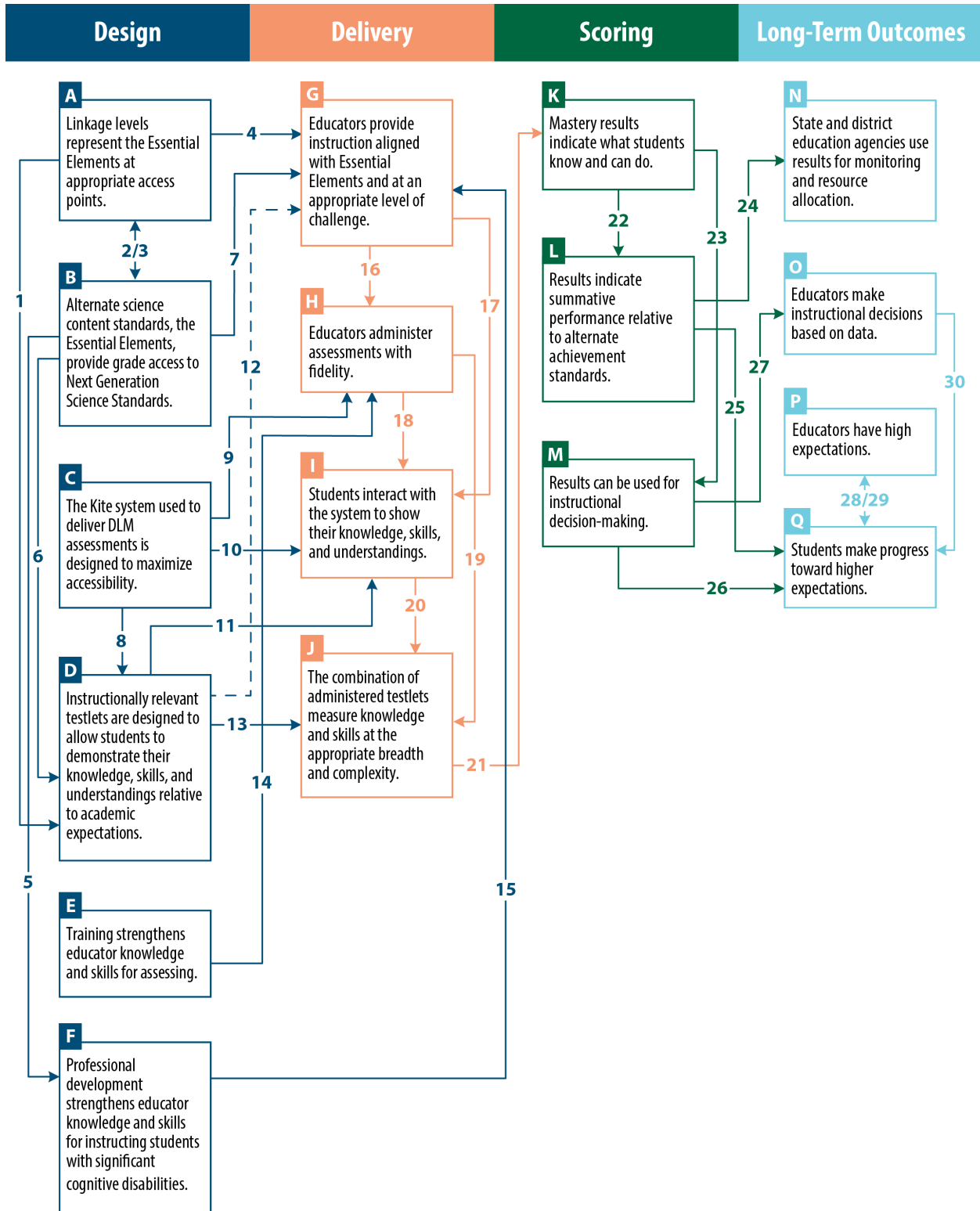
The DLM Theory of Action expresses a commitment to provide students with the most significant cognitive disabilities access to an assessment system that is capable of validly and reliably evaluating their achievement. Ultimately, students will make progress toward higher expectations, educators will make instructional decisions based on data, educators will hold higher expectations of students, and state and district education agencies will use results for monitoring and resource allocation.

The DLM Governance Board adopted an argument-based approach to assessment validation. The validation process began in 2013 by defining with governance board members the policy uses of DLM assessment results. We followed this with a three-tiered approach, which included specification of 1) a Theory of Action defining statements in the validity argument that must be in place to achieve the goals of the system; 2) an interpretive argument defining propositions that must be evaluated to support each statement in the Theory of Action; and 3) validity studies to evaluate each proposition in the interpretive argument.

After identifying these overall guiding principles and anticipated outcomes, specific elements of the DLM Theory of Action were articulated to inform assessment design and to highlight the associated validity arguments. The Theory of Action includes the assessment's intended effects (long-term outcomes), statements related to design, delivery and scoring, and action mechanisms (i.e., connections between the statements; see Figure 1.2). The chain of reasoning in the Theory of Action is demonstrated broadly by the order of the four sections from left to right. Dashed lines represent connections that are present when the optional instructionally embedded assessments are utilized. Design statements serve as inputs to delivery, which informs scoring and reporting, which collectively lead to the long-term outcomes for various stakeholders. The chain of reasoning is made explicit by the numbered arrows between the statements.

**Figure 1.2**

*Dynamic Learning Maps Theory of Action*



## 1.5. Key Features

Consistent with the Theory of Action, key elements were identified to guide the design of the DLM science alternate assessment. The list of key elements below mirrors the organization of this manual and provides chapter references.

1. **A set of particularly important learning targets most frequently addressed in DLM science states that serve as grade band content standards for students with significant cognitive disabilities and provide an organizational structure for educators.**

The selection of learning targets is crucial to instruction and assessment development; educators must be able to build the knowledge, skills, and understandings required to achieve the content standard expectations for each grade band and subject. This forms a local learning progression toward a specific learning target. The process for selecting learning targets and developing EEs with three linkage levels for assessment are described in Chapter 2 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

2. **Instructionally relevant testlets that engage the student in science tasks and reinforce learning.**

Instructionally relevant assessments consist of activities an educator could use as a springboard for designing instructional activities combined with the systematic gathering and analysis of data. These assessments necessarily take different forms depending on the population of students and the concepts being taught. The development of an instructionally relevant assessment begins by creating items using principles of evidence-centered design and Universal Design for Learning (UDL), then linking related items together into meaningful groups, which the DLM System calls testlets. Item and testlet design are described in Chapter 3 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

3. **Adaptive assessments that reinforce academic expectations.**

The DLM science alternate assessment is designed as an adaptive, computer-delivered assessment that is intended to measure knowledge, skills, and understandings at appropriate levels of complexity for the content. It consists of an end-of-year assessment that meets the requirements of accountability systems and provides detailed descriptions of what students know and can do. Assessment administration is described in Chapter 4 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

4. **Accessibility by design and alternate testlets.**

Accessibility is a prerequisite to validity or the degree to which an assessment score interpretation is justifiable for a particular purpose and supported by evidence and theory (American Educational Research Association et al. [AERA et al.], 2014). Therefore, throughout all phases of development, the DLM System was designed with accessibility in mind to provide multiple means of representation, expression, action, and engagement. Students must understand what is being asked in an item or task and have the tools to respond in order to demonstrate what they know and can do (Karvonen et al., 2015). The DLM alternate assessment provides accessible content, accessible delivery via technology, and adaptive routing. Since all students taking an alternate assessment based on alternate academic achievement standards are students with SCD, accessibility supports are universally available. The emphasis is on selecting the appropriate accessibility features and tools for each individual student. Accessibility considerations are described in Chapter 2 (linkage levels), Chapter 3 (testlet development), and Chapter 4 (accessibility during assessment administration) of



the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

**5. Status and score reporting that is readily actionable.**

Due to the unique characteristics of a mastery-based system, DLM assessments require new approaches to psychometric analysis and modeling, with the goal of assuring accurate inferences about student performance relative to the content as it is organized in the EEs and linkage levels. Each EE is designed to address three levels of complexity, called linkage levels. Diagnostic classification modeling is used to determine a student’s likelihood of mastering assessed linkage levels associated with each EE. Providing student mastery information at the linkage level allows for instructional next steps to be readily derived. A student’s overall performance level in the subject is determined by aggregating linkage level mastery information across EEs. This scoring model supports reports that can be immediately used to guide instruction and describe levels of mastery. The DLM modeling approach is described in Chapter 5 of this manual, and score report design is described in Chapter 7 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

## **1.6. Technical Manual Overview**

This manual provides evidence collected during the 2021–2022 administration of science assessments.

Chapter 1 provides an overview of the theoretical underpinnings of the DLM science assessment, including a description of the DLM collaborators, the target student population, an overview of the assessment, an introduction to the Theory of Action and interpretive argument, and a summary of contents of the remaining chapters.

Chapter 2 was not updated for 2021–2022. For a full description of the process by which the EEs were developed, including the intended coverage with the *Framework* (National Research Council, 2012) and the NGSS (NGSS, 2013), see the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

Chapter 3 outlines assessment design and development, including a description of test development activities and external review of content. The chapter then presents evidence of item quality including field test items, operational item data, and differential item functioning.

Chapter 4 describes assessment delivery including updated procedures and data collected in 2021–2022. The chapter presents evidence from the DLM System, including administration time, device usage, adaptive routing, and accessibility support selections; evidence from monitoring assessment administration, including test administration observations and data forensics monitoring; and evidence from test administrators, including user experience with the DLM System and student opportunity to learn.

Chapter 5 describes the updated psychometric model which was implemented in 2021–2022. The chapter demonstrates how the DLM project draws upon a well-established research base in cognition and learning theory and uses psychometric methods that provide feedback about student progress and learning acquisition. This chapter describes the psychometric model that underlies the DLM project and describes the process used to estimate item and student parameters from student test data and evaluate model fit.

Chapter 6 was not updated for 2021–2022; no changes were made to the cut points used in scoring DLM assessments. See the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017) for a description of the methods, preparations, procedures, and results of the original standard-setting meeting and the follow-up evaluation of the impact data. For a description 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 *2018–2019 Technical Manual Update—Science* (DLM Consortium, 2019).

Chapter 7 reports the 2021–2022 operational results, including student participation data. The chapter details the percentage of students achieving 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 2021–2022 administration.

Chapter 8 focuses on reliability evidence and describes the updated simulated retest method for assessing consistency in student results. This includes a description of the methods used to evaluate assessment reliability and a summary of results by the linkage level, EE, domain or topic, and subject (overall performance).

Chapter 9 describes updates to the professional development offered across states administering DLM assessments in 2021–2022, including participation rates and evaluation results.

Chapter 10 synthesizes the evidence provided in the previous chapters. It evaluates how the evidence supports the claims in the Theory of Action as well as the long-term outcomes. The chapter ends with a description of our future research and ongoing initiatives for continuous improvement.

## 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, NGSS, 2013), see Chapter 2 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

## 3. Assessment Design and Development

Chapter 3 of the Dynamic Learning Maps® (DLM®) Alternate Assessment System *2015–2016 Technical Manual—Science* (DLM Consortium, 2017) describes assessment design and development procedures. This chapter provides an overview of updates to item and test development for the 2021–2022 academic year. The first portion of the chapter provides an overview of 2021–2022 item writers' characteristics, followed by the 2021–2022 external review of items, testlets, and texts based on criteria for content, bias, and accessibility. The next portion of the chapter describes field test testlets available for administration during 2021–2022, changes to the pool of operational items, and an evaluation of differential item functioning.

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 used to guide test development, see the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

### 3.1. Test Development Procedures

This section describes information pertaining to item writing and item writer demographics for the 2021–2022 year. For a complete summary of item and testlet development procedures that were developed and implemented in 2015–2016 and continue to be used in 2021–2022, see Chapter 3 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

#### 3.1.1. Testlet and Item Writing

Item development for 2021–2022 was reduced in scope to support an initiative to review and refresh resources that guide the development of items and testlets.

##### 3.1.1.1. Participants

Item writers were selected from the ATLAS MemberClicks database based on predetermined qualifications such as special education teaching experience and previous DLM item-writing experience. The database is a profile-based recruitment tool hosted in MemberClicks, a membership management software. Individuals create their participant profile, which can be accessed and updated at any time. We encourage individuals to update their profile information annually or at the time of event recruitment for reporting accuracy.

Participant profiles include attribute fields to capture demographic, education, and work experience data. Item writers were assigned to a subject area based on their qualifications and previous DLM item writing experience. In total, 37 science item writers contributed to writing testlets during the 2021–2022 year.

The median and range of years of item writers' teaching experience is shown in Table 3.1. Of the item writers who responded to the question, the median years of experience was 13 years for item writers of science testlets in pre-K–12 and 10 years of special education experience.

**Table 3.1**

*Item Writers' Years of Teaching Experience*

Teaching Experience	<i>n</i>	Median	Range
Science	29	12.0	4–35
Pre-K–12	33	13.0	4–38
Special education	30	10.0	1–34

\* The *n* column indicates the number of nonmissing responses to the survey question

High school was most commonly taught by item writers (*n* = 38; 28%). See Table 3.2 for a summary.

**Table 3.2**

*Item Writers' Grade-Level Teaching Experience*

Grade level	<i>n</i>	%
Grade 3	12	8.9
Grade 4	15	11.1
Grade 5	15	11.1
Grade 6	16	11.9
Grade 7	19	14.1
Grade 8	20	14.8
High school	38	28.1

The level and most common types of degrees held by item writers are shown in Table 3.3 and Table 3.4, respectively. All science item writers held at least a bachelor's degree. The majority of the science item writers (*n* = 31; 84%) also held a master's degree, for which the most common field of study was special education (*n* = 10; 32%).

**Table 3.3**

*Item Writers' Level of Degree*

Degree	<i>n</i>	%
Bachelor's	5	13.5
Master's	31	83.8
Other	1	2.7

**Table 3.4**

*Item Writers' Degree Type (N = 37)*

Degree	<i>n</i>
Bachelor's degree	
Education	9
Content specific	4
Special education	8
Other	8
Missing	7
Master's degree	
Education	8
Content specific	3
Special education	10
Other	8
Missing	2

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

**Table 3.5**

*Item Writers' Experience with Disability Categories*

Disability category	<i>n</i>	%
Blind/low vision	14	5.6
Deaf/hard of hearing	16	6.4
Emotional disability	23	9.2
Mild cognitive disability	24	9.6
Multiple disabilities	29	11.6
Orthopedic impairment	19	7.6
Other health impairment	30	12.0
Significant cognitive disability	30	12.0
Specific learning disability	25	10.0
Speech impairment	25	10.0
Traumatic brain injury	16	6.4

The professional roles reported by the 2021–2022 item writers are shown in Table 3.6. Roles included educators, instructional coaches, district staff, and other (i.e., university staff, program coordinators, supervisors of instruction).

**Table 3.6**

*Professional Roles of Item Writers*

Role	<i>n</i>	%
Classroom educator	23	62.2
District staff	2	5.4
Instructional coach	3	8.1
Other	9	24.3

Science item writers were from 14 different states. Population density of schools in which item writers taught or held a position is reported in Table 3.7. Within the survey, rural was defined as a population living outside settlements of 1,000 or fewer inhabitants, suburban was defined as an outlying residential area of a city of 2,000-49,000 or more inhabitants, and urban was defined as a city of 50,000 inhabitants or more. The demographics for the item writers are presented in Table 3.8.

**Table 3.7**

*Population Density for Schools of Item Writers*

Population density	<i>n</i>	%
Rural	22	59.5
Suburban	7	18.9
Urban	8	21.6

**Table 3.8**

*Demographics of the Item Writers*

	<i>n</i>	%
Gender		
Female	33	89.2
Male	4	10.8
Race		
White	31	83.8
Other	3	8.1
African American	2	5.4
Chose not to disclose	1	2.7
Hispanic ethnicity		
Non-Hispanic	32	86.5
Hispanic	1	2.7
Chose not to disclose	4	10.8

### **3.1.1.2. Item Writing Process**

The selected item writers completed independent asynchronous advanced training and later participated in a 2-day virtual item-writing event that was held on January 25–26, 2022. Item writer training included instruction on the item-writing process and peer review process. During the event, item-writing pairs collaborated and began to develop testlets. Following the virtual event, item writers continued producing and peer reviewing testlets virtually via a secure online platform through June 2022. A total of 270 testlets were written for science.

### **3.1.2. External Reviews**

The purpose of external reviews of items and testlets is to evaluate whether the items and testlets measure the intended content, are accessible, and are free of biased or sensitive content. Panelists use external review criteria established for DLM alternate assessments to recommend items be accepted, revised, or rejected. Panelists also provide recommendations for revisions or an explanation for a “reject” rating. The test development team uses the collective feedback from the panelists to inform decisions about items and testlets before they are field-tested.

External review for 2021–2022 was held as a 2-day virtual event. Materials were updated to meet the needs of virtual panel meetings, including the advance training and facilitator and co-facilitator training. When held in person, one facilitator led the feedback discussion for each panel. This year, a facilitator and co-facilitator led the feedback discussions and recorded decisions for each panel meeting.

#### **3.1.2.1. Review Recruitment, Assignments, and Training**

Panelists were selected from the ATLAS MemberClicks database based on predetermined qualifications for each panel type. The ATLAS MemberClicks database is populated using a profile creation survey that captures demographic, education, and work experience of candidates from DLM partner states. Panelists were assigned to content, accessibility, or bias and sensitivity panels based on their qualifications.

There were 39 science reviewers: 16 on accessibility panels, 8 on content panels, and 15 on bias and sensitivity panels.

Prior to participating in the virtual panel meetings, panelists completed an advance training course that included an External Review Procedures module and a module for their assigned panel type. The content modules were subject specific, while the accessibility and bias and sensitivity modules were universal for all subjects. After each module, panelists completed a posttest and were required to score 80% or higher to pass advance training.

After completing the modules and corresponding posttests, panelists completed a practice activity that simulated the external review process for each panel type. Panelists used the criteria for their assigned panel type to complete this external review.

Following the completion of advance training, panelists completed asynchronous reviews on two or three collections of testlets dependent upon panel type. Collections had between 34 and 54 testlets, dependent on the panel type. Content panels had fewer testlets per collection, and bias and sensitivity and accessibility panels had more testlets per collection. Ratings from the asynchronous reviews were sorted and new collections were created containing items and testlets with discrepant panel ratings. Dependent on the subject, there were two to four virtual panel meetings led by facilitators and co-facilitators to obtain

collective feedback about the items and testlets.

The median and range of years of teaching experience is shown in Table 3.9. The median years of experience for external reviewers was 15 years in pre-K–12 and 10 years in science.

**Table 3.9**

*External Reviewers' Years of Teaching Experience*

Teaching experience	Median	Range
Pre-K–12	15.0	5–35
Science	10.0	1–35

High school was most commonly taught by the external reviewers ( $n = 32$ ; 28%). See Table 3.10 for a summary.

**Table 3.10**

*External Reviewers' Grade-Level Teaching Experience*

Grade level	$n$	%
Grade 3	10	25.6
Grade 4	11	28.2
Grade 5	12	30.8
Grade 6	15	38.5
Grade 7	18	46.2
Grade 8	17	43.6
High school	32	82.1

*Note.* Reviewers could indicate multiple grade levels.

The 39 external reviewers represented a highly qualified group of professionals. The level and most common types of degrees held by external reviewers are shown in Table 3.11 and Table 3.12, respectively. A majority ( $n = 35$ ; 90%) also held a master's degree, for which the most common field of study was special education ( $n = 13$ ; 33%).

**Table 3.11**

*External Reviewers' Level of Degree*

Degree	$n$	%
Bachelor's	4	10.3
Master's	35	89.7



**Table 3.12**

*External Reviewers' Degree Type*

Degree	<i>n</i>	%
<b>Bachelor's degree</b>		
Education	12	30.8
Content specific	1	2.6
Special education	10	25.6
Other	15	38.5
Missing	1	2.6
<b>Master's degree</b>		
Education	10	28.6
Content specific	2	5.7
Special education	14	40.0
Other	9	25.7

Most external reviewers had experience working with students with disabilities (77%), and 90% had experience with the administration of alternate assessments. The variation in percentages suggest some item writers may have had experience with administration of alternate assessments but perhaps did not regularly work with students with disabilities.

External reviewers reported a range of experience working with students with different disabilities, as summarized in Table 3.13. External reviewers collectively had the most experience working with students with a significant cognitive disability, multiple disabilities, or other health impairments.

**Table 3.13**

*External Reviewers' Experience with Disability Categories*

Disability category	<i>n</i>	%
Blind/low vision	16	41.0
Deaf/hard of hearing	13	33.3
Emotional disability	21	53.8
Mild cognitive disability	23	59.0
Multiple disabilities	26	66.7
Orthopedic impairment	15	38.5
Other health impairment	24	61.5
Significant cognitive disability	26	66.7
Specific learning disability	23	59.0
Speech impairment	19	48.7
Traumatic brain injury	14	35.9

*Note.* Reviewers could select multiple categories.

Panelists had varying experience teaching students with the most significant cognitive disabilities. Science panelists had a median of 5.5 years of experience teaching students with the most significant cognitive disabilities, with a minimum of 3 years and a maximum of 10 years of experience.

The professional roles reported by the 2021–2022 reviewers are shown in Table 3.14. Roles included educators, instructional coaches, state education agency staff, and other (i.e., university staff, program coordinators, supervisors of instruction).

**Table 3.14**

*Professional Roles of External Reviewers*

Role	<i>n</i>	%
Instructional coach	1	2.6
Other	4	10.3
State education agency staff	1	2.6
Not specified	1	2.6
	32	82.1

Science panelists' were from five different states. Population density of schools in which reviewers taught or held a position is reported in Table 3.15. Within the survey, rural was defined as a population living outside settlements of 1,000 or fewer inhabitants, suburban was defined as an outlying residential area of a city of 2,000-49,000 or more inhabitants, and urban was defined as a city of 50,000 inhabitants or more. The demographics for the external reviewers are presented in Table 3.16.

**Table 3.15**

*Population Density of School of Content Panelists*

Population density	<i>n</i>	%
Rural	19	48.7
Suburban	5	12.8
Urban	15	38.5

**Table 3.16**

*Demographics of the External Reviewers*

	<i>n</i>	%
Gender		
Female	32	82.1
Male	7	17.9
Race		
White	31	79.5
African American	4	10.3
Chose not to disclose	2	5.1
American Indian	1	2.6
Native Hawaiian or Pacific Islander	1	2.6
Hispanic ethnicity		
Non-Hispanic	36	92.3
Hispanic	2	5.1
Chose not to disclose	1	2.6

Prior to attending the on-site external review event, panelists completed an advance training course. The course included five modules that all panelists had to complete, allowing the panelists to gain familiarity with all areas being covered. All panelists completed the following modules: DLM Overview and External Review Process, Accessibility, Bias and Sensitivity, and Content. Each content module was subject-specific, while the bias and sensitivity and accessibility modules were universal for all subjects. After each module, the panelists had to complete a posttest and receive a score of at least 80% to continue to the next module. Panelists were required to complete advance training prior to reviewing any testlets at the event.

Review of testlets was completed only during the two days of the on-site event. Due to technical issues, not all panelists had time to review as many testlets as in prior years. As a consequence, some subject rooms saw fewer testlets reviewed than others. Panelists reviewed each testlet on their own and then as a group. Each group came to a consensus for each item and testlet, and the facilitator recorded the recommendation.

Panelists recommended most content be accepted. For science, the percentage of items and testlets rated as “accept” ranged from 44% to 91% and 50% to 96%, respectively. The percentage of items and testlets rated as “revise” ranged from 9% to 54% and 4% to 47% respectively. The rate at which both items and testlets were recommended for rejection ranged from 0% to 3% across grades, pools, and rounds of review.

### **3.1.2.2. Item and Testlet 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. Once the test development team views each item and testlet and considers the feedback provided by the panelists, it assigns a decision to each one.

The science test development team accepted as is, 56% of testlets and 30% of items. Of the items and testlets that were revised, most required major changes (e.g., stem or response option replaced) as opposed to minor changes (e.g., minor rewording but concept remained unchanged). The science test development team made 31 minor revisions to items, 250 major revisions to items, and rejected 26 testlets. Most of the content reviewed during this external review is scheduled for the spring 2023 window.

## **3.2. Evidence of Item Quality**

Testlets are the fundamental unit of the DLM alternate assessments. Each year, testlets are added to and removed from the operational pool to maintain a pool of high-quality testlets. The following sections describe evidence of item quality, including evidence supporting field-test testlets available for administration, a summary of the operational pool, and evidence of differential item functioning (DIF).

### **3.2.1. Field Testing**

During the 2021–2022 academic year, DLM field-test testlets were administered to evaluate item quality for EEs assessed at each grade level for science. Field testing is conducted to deepen operational pools so that multiple testlets are available in the spring assessment window, including making more content available at EEs and linkage levels that educators administer to students the most. By deepening the operational pool, testlets can also be evaluated for retirement in instances where other testlets perform better.

In this section we describe the field-test testlets administered in 2021–2022 and the associated review activities. A summary of prior field test events can be found in Chapter 3 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017) and subsequent annual DLM technical manual updates.

#### **3.2.1.1. Description of Field Tests Administered in 2021–2022**

Testlets were made available for field testing based on the availability of field-test content for each EE and linkage level.

During spring assessment window, field-test testlets were administered after completion of the operational assessment. A field-test testlet was assigned for an EE that was assessed during the operational assessment at a linkage level equal or adjacent to the linkage level of the operational testlet.

Table 3.17 summarizes the number of field-test testlets available during 2022. A total of 124 were available across grades.

**Table 3.17**

*Spring 2022 Field-Test Testlets*

Grade	<i>n</i>
Elementary	43
Middle school	33
High school	33
Biology	15

Participation in field testing was not required, but educators were encouraged to administer all available testlets to their students. In total, 26,926 (61%) students completed at least one field-test testlet. In the spring assessment window, 88% of field-test testlets had a sample size of at least 20 students (i.e., the threshold for item review).

### 3.2.1.2. Field-Test Data Review

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

Review of field-test data occurs annually during February and March. This includes data from the previous spring assessment window. That is, the review in February and March of 2022 includes field-test data collected during the 2021 spring assessment window. Data that were collected during the 2022 spring assessment window will be reviewed in February and March of 2023, with results included in the 2022–2023 technical manual update.

Test development teams for each subject make four types of item-level decisions as they review field-test items flagged for either a *p*-value or a standardized difference value beyond the threshold:

1. No changes made to item. Test development team decided item can go forward to operational assessment.
2. Test development team identified concerns that required modifications. Modifications were clearly identifiable and were likely to improve item performance.
3. Test development team identified concerns that required modifications. The content was worth preserving rather than rejecting. Item review may not have clearly pointed to specific edits that were likely to improve the item.
4. Rejected item. Test development team determined the item was not worth revising.

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

Common reasons for flagging an item for modification included items that were misaligned to the linkage level, distractors that could be argued as partially correct, or unnecessary complexity in the language of the stem. After reviewing flagged items, the reviewers looked at all items rated as three or four within the testlet to help determine whether to retain or reject the testlet. Here, the test development team could elect to keep the testlet (with or without revision) or reject it. If a revision was needed, it was assumed the testlet needed field testing again. The entire testlet was rejected if the test development team determined the flagged items could not be adequately revised.

### **3.2.1.3. Results of Item Analysis**

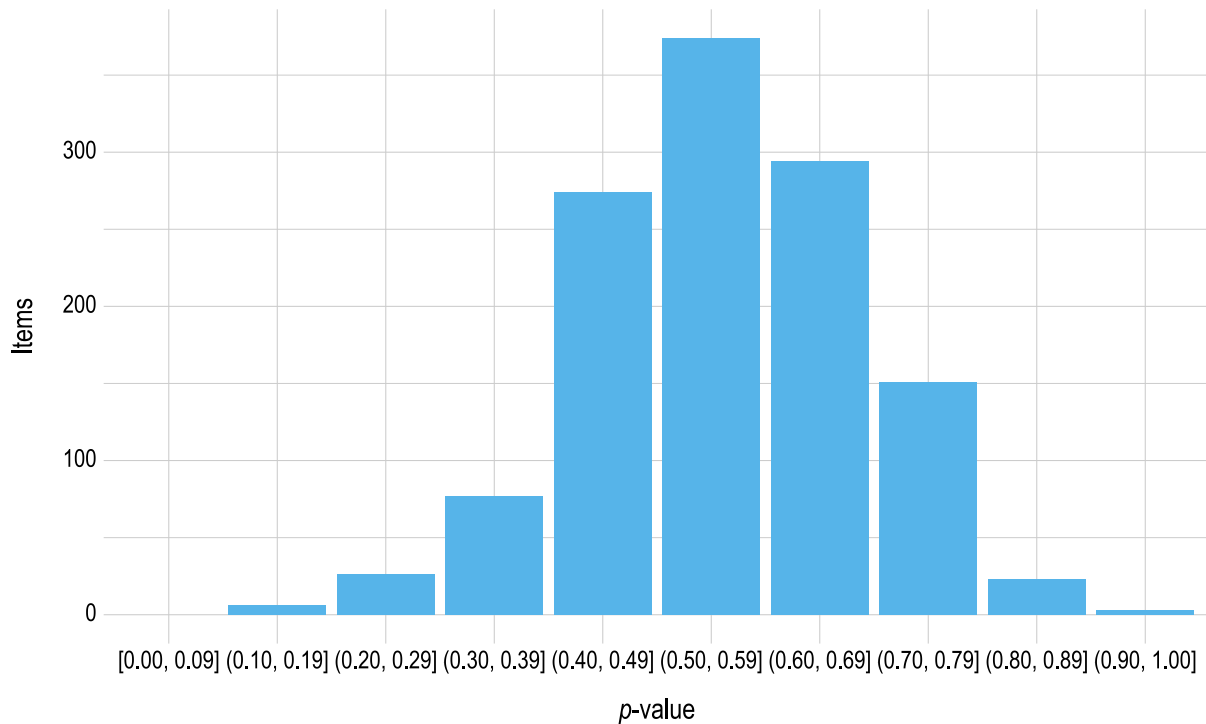
Criteria used for item flagging during previous field test events were retained for 2021–2022. Items were flagged for review by test development teams if they met either of the following statistical criteria:

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

Figure 3.1 summarizes the  $p$ -values for items that met the minimum sample size threshold of 20. Most items fell above the .35 threshold for flagging. Test development teams for each subject reviewed items below the threshold.

**Figure 3.1**

*p-values for Science Field-Test Items*



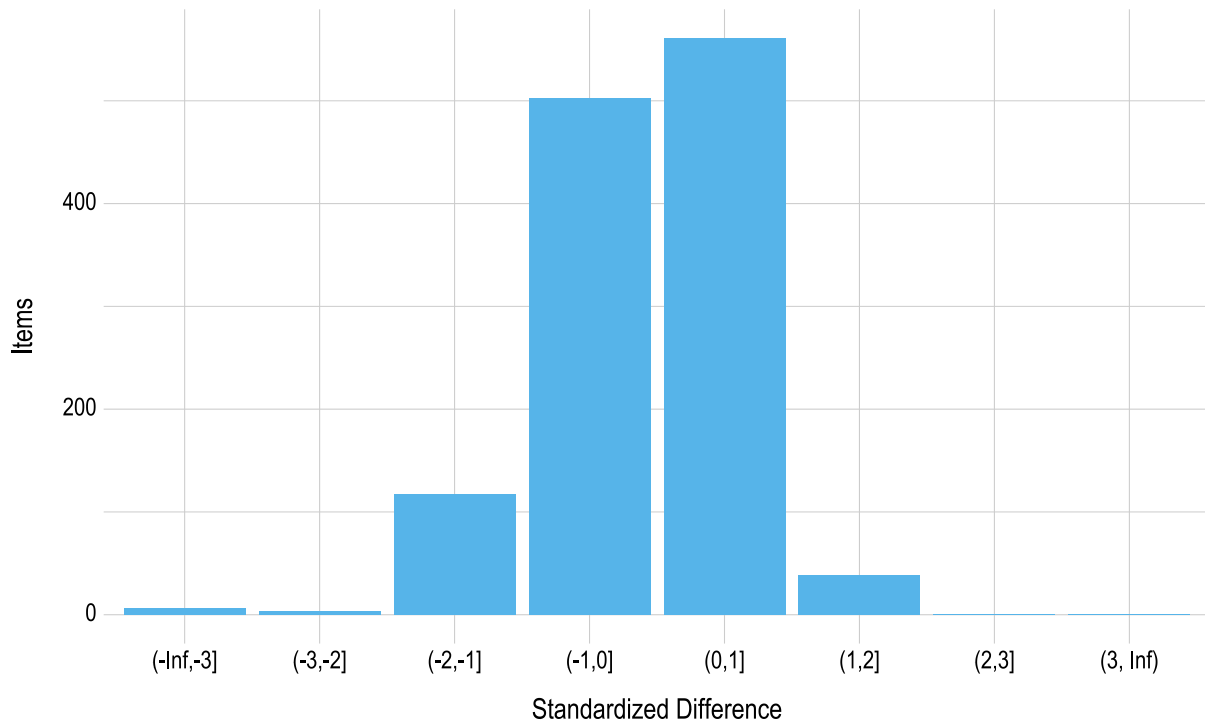
*N* = 1,228

Items with a sample size of less than 20 were omitted.

Figure 3.2 summarizes the standardized difference values for items field tested during the spring assessment window for science. Most items fell within two standard deviations of the mean for the EE and linkage level. Items beyond the threshold were reviewed by test development teams for each subject.

**Figure 3.2**

*Standardized Difference Z-Scores for Science Field-Test Items*



*N* = 1,228

Items with a sample size of less than 20 were omitted.

A total of 10 science testlets (14%) had at least one item flagged due to their *p*-value and/or standardized difference value. Test development teams reviewed all flagged items and their context within the testlet to identify possible reasons for the flag and to determine whether an edit was likely to resolve the issue.

Of the 64 science testlets that were not flagged, four (6%) were edited and reassigned to the field-test pool, 45 (70%) were promoted to the operational pool to maintain pool depth given content-based testlet retirement, seven (11%) were sent back to the field-test pool with no edits for additional data collection to get estimates of item difficulty that are based on larger samples, and eight (12%) were rejected and retired. Of the 10 science testlets that were flagged, three (30%) were edited and reassigned to the field-test pool, one (10%) was promoted to the operational pool to maintain pool depth given content-based testlet retirement, two (20%) were sent back to the field-test pool with no edits for additional data collection to get estimates of item difficulty that are based on larger samples, and four (40%) were rejected and retired.

### **3.2.2. Operational Assessment Items for 2021–2022**

The DLM assessments include a total of 152 operational testlets. Because the operational pool needs to be deepened, particularly for content at the Essential Elements (EEs) and linkage levels that are administered to students the most, updates are made to the operational pool each year. The primary updates are promoting testlets to the operational pool and removing testlets from the operational pool.



Testlets are promoted to the operational pool via field testing, with students who completed the operational assessment in the spring. Field-test testlets are eligible for review after they have been completed by at least 20 students. As mentioned in the field testing section above (section 3.2.1), there are multiple item quality indicators that are considered when deciding whether to promote an item to the operational pool. Statistically, items are expected to be appropriately difficult and to function similarly to items measuring the same EE and linkage level. To review these statistical item quality indicators, the difficulty and internal consistency of items on field-test testlets are evaluated. Items are also expected to be consistent with DLM item-writing guidelines and aligned with the assessed linkage level. To review these content-based item quality indicators, the quality of the eligible items on the field-test testlets is evaluated, and the test development team makes decisions of whether to accept or reject the items on the field-test testlets. For a full description of field testing, see above in section 3.2.1.

Testlets are removed from the operational pool via retirement based on item quality standards. There are several processes that can lead an item or testlet to be prioritized for retirement. Items are evaluated for evidence of model fit, and the results of these evaluations may be used to prioritize items and testlets for retirement. Items are also evaluated for evidence of DIF, and these results may be used to prioritize items and testlets for retirement. This process is described in section 3.2.3. Finally, the test development team periodically reviews the content pool and prioritizes testlets for retirement. These reviews refresh the operational pool by removing older content when newer content is available.

For 2021–2022, 46 science testlets were promoted to the operational pool from field testing in 2020–2021.

Testlets were made available for operational testing in 2021–2022 based on the 2020–2021 operational pool and the promotion of testlets field-tested during 2020–2021 to the operational pool following their review. Table 3.18 summarizes the total number of operational testlets for 2021–2022. In total, there were 152 operational testlets available. This total included 36 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 3.18**

*2021–2022 Operational Testlets, by Grade Band (N = 152)*

Grade	<i>n</i>
Elementary	39
Middle school	41
High school	41
Biology	31

*Note:* Three Essential Elements are shared across the high school and Biology assessments.

### 3.2.2.1. Educator Perception of Assessment Content

Each year, test administrators are asked two questions about their perceptions of the assessment content;<sup>1</sup> Table 3.19 describes their responses in 2021–2022. Questions pertained to whether the DLM assessments measured important academic skills and reflected high expectations for their students.

<sup>1</sup> Participation in the test administrator survey is described in Chapter 4 of this manual.

Test administrators generally responded that content reflected high expectations for their students (86% agreed or strongly agreed) and measured important academic skills (77% agreed or strongly agreed). While the majority of test administrators agreed with these statements, 14%–23% 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), test administrators' responses may reflect awareness that DLM assessments contain challenging content. However, test administrators were divided on its importance in the educational programs of students with the most significant cognitive disabilities. Feedback from focus groups with educators focusing on score reports reflected similar variability in educator perceptions of assessment content (Clark et al., 2018, 2022).

**Table 3.19**

*Educator Perceptions of Assessment Content*

Statement	Strongly disagree		Disagree		Agree		Strongly agree	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Content measured important academic skills and knowledge for this student.	2,262	9.1	3,535	14.2	14,655	58.7	4,510	18.1
Content reflected high expectations for this student.	1,193	4.8	2,378	9.6	14,712	59.4	6,500	26.2

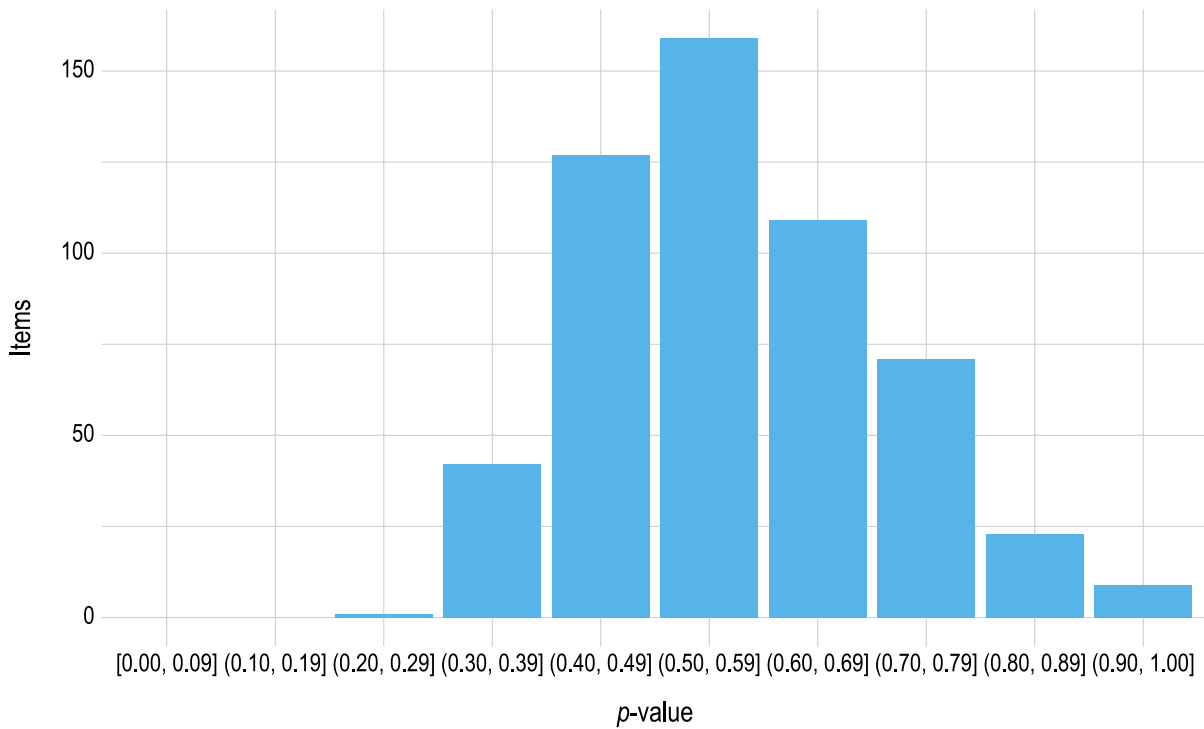
**3.2.2.2. Psychometric Properties of Operational Assessment Items for 2021–2022**

The proportion correct (*p*-value) was calculated for all operational items to summarize information about item difficulty.

Figure 3.3 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. In total, zero items (<1% of all items) were excluded due to small sample size. The *p*-values for most science items were between .4 and .7.

**Figure 3.3**

*p-values for Science 2022 Operational Items*



*N* = 541

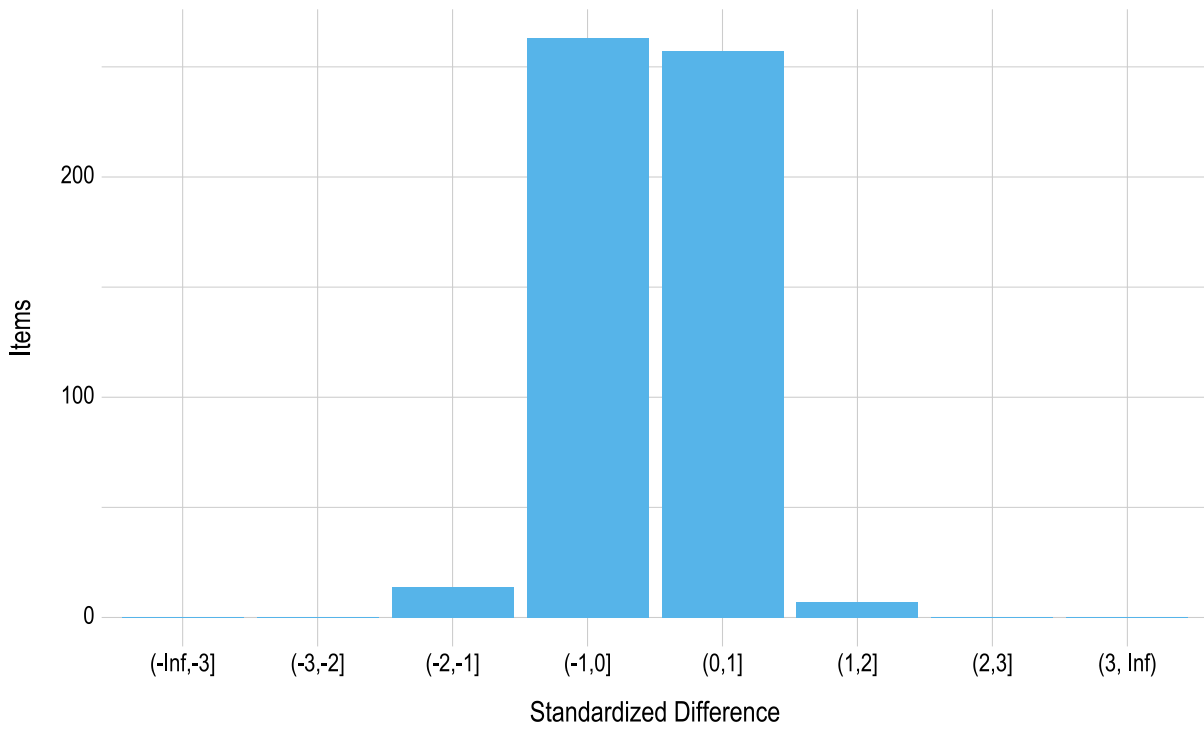
*Note.* Items with a sample size of less than 20 were omitted.

Items in the DLM assessments are fungible (i.e., interchangeable) within each EE and linkage level, meaning that the items are expected to function identically to the other items measuring the same EE and linkage level. To evaluate the fungibility assumption, 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. If an item is fungible with the other items measuring the same EE and linkage level, the item is expected to have a nonsignificant standardized difference value. The standardized difference values provide one source of evidence of internal consistency.

Figure 3.4 summarizes the standardized difference values for operational items for science. Of all items measuring the EE and linkage level, 100% of items fell within two standard deviations of the mean. As additional data are collected and decisions are made regarding item pool replenishment, test development teams will consider item standardized difference values, along with item misfit analyses, when determining which items and testlets are recommended for retirement.

**Figure 3.4**

*Standardized Difference Z-Scores for Science 2021–2022 Operational Items*



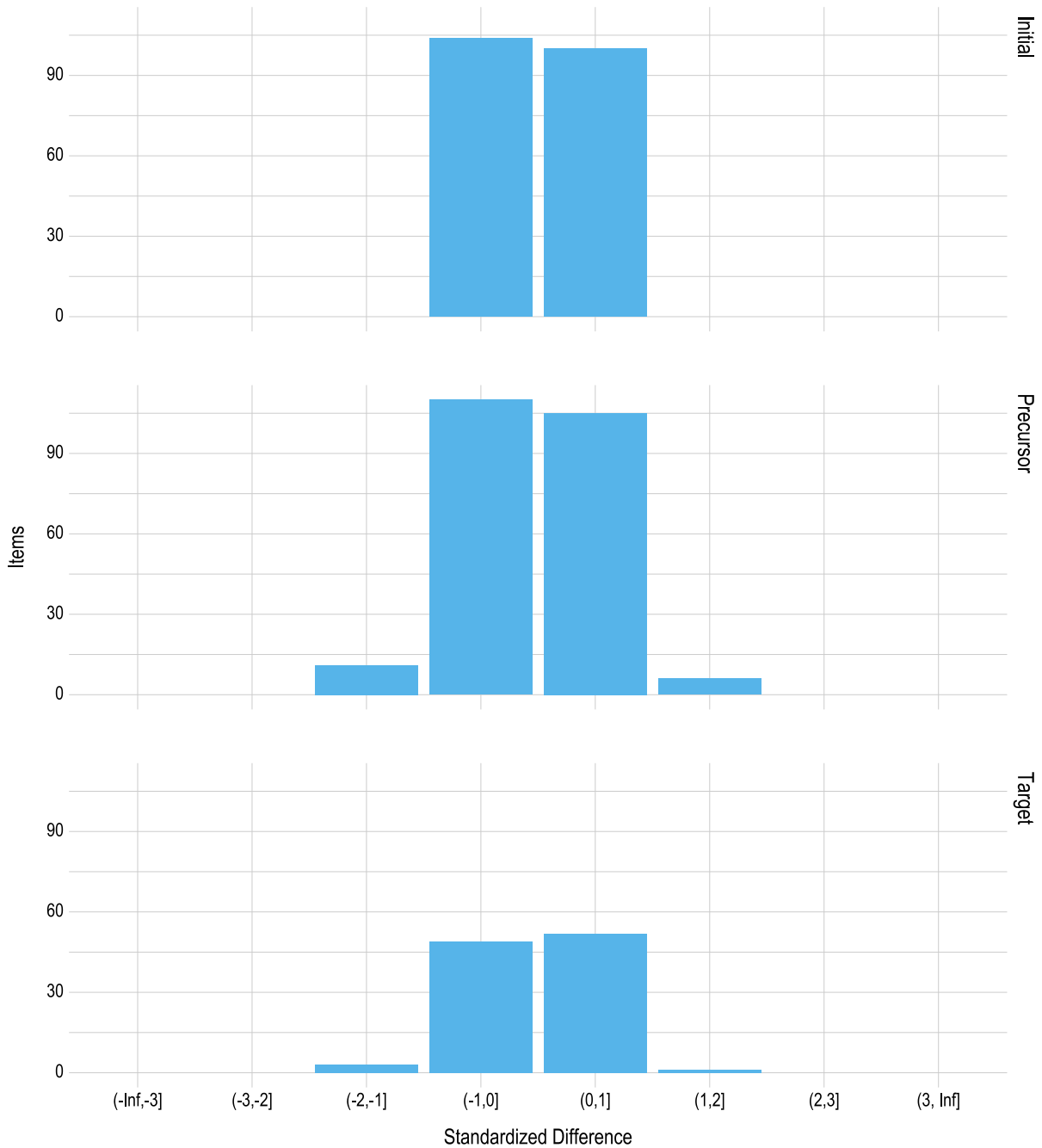
N = 541

*Note.* Items with a sample size of less than 20 were omitted.

Figure 3.5 summarizes the standardized difference values for operational items by linkage level. Most items fell within two standard deviations of the mean of all items measuring the respective EE and linkage level, and the distributions are consistent across linkage levels.

**Figure 3.5**

*Standardized Difference Z-Scores for 2021–2022 Operational Items by Linkage Level*



N = 541

*Note.* Items with a sample size of less than 20 were omitted.

### **3.2.3. Evaluation of Item-Level Bias**

DIF addresses the challenges created when some test items are more difficult for some groups of examinees despite these examinees having knowledge and understanding of the assessed concepts (Camilli & Shepard, 1994). DIF analyses can uncover internal inconsistency if particular items are functioning differently in a systematic way for identifiable subgroups of students (AERA et al., 2014). While identification of DIF does not always indicate a weakness in the test item, it can point to construct-irrelevant variance, posing considerations for validity and fairness.

#### **3.2.3.1. Method**

DIF analyses examined race in addition to gender. Analyses included data from 2015–2016 through 2020–2021<sup>2</sup> 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) and for race subgroups: white, African American, Asian, American Indian, Native Hawaiian or Pacific Islander, Alaska Native, and multiple races.

The DLM student population is unbalanced in both gender and race. The number of female students responding to items is smaller than the number of male students by a ratio of approximately 1:2. Similarly, the number of nonwhite students responding to items is smaller than the number of white students by a ratio of approximately 1:2. Therefore, on advice from the DLM Technical Advisory Committee, the threshold for item inclusion requires that the focal 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.

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 or racial group was greater than .97 or less than .03 were also removed from the analyses.

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 the reference group (i.e., males for gender, white for race) coded as 1 and the focal group (i.e., females for gender; African American, Asian, American Indian, Native Hawaiian or Pacific Islander, Alaska Native, or two or more races for race) coded as 0. An interaction term was included to evaluate whether nonuniform DIF was present for each item (Swaminathan & Rogers, 1990); the presence of nonuniform DIF indicates that the item functions differently because of the interaction between total linkage levels mastered and the student's group (i.e., gender or racial group). When nonuniform DIF is present, the 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:

---

<sup>2</sup> DIF analyses are conducted on the sample of data used to update the model calibration, which uses data through the previous operational assessment. See Chapter 5 of this manual for more information.

$$M_0: \text{logit}(\pi_i) = \beta_0 + \beta_1 X \quad (3.1)$$

$$M_1: \text{logit}(\pi_i) = \beta_0 + \beta_1 X + \beta_2 G \quad (3.2)$$

$$M_2: \text{logit}(\pi_i) = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG \quad (3.3)$$

where  $\pi_i$  is the probability of a correct response to item  $i$ ,  $X$  is the matching criterion,  $G$  is a dummy coded grouping variable (0 = reference group, 1 = focal group),  $\beta_0$  is the intercept,  $\beta_1$  is the slope,  $\beta_2$  is the group-specific parameter, and  $\beta_3$  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 group and interaction terms to the regression model.

For each item, the change in the Nagelkerke pseudo  $R^2$  measure of effect size was captured, from  $M_0$  to  $M_1$  or  $M_2$ , to account for the effect of the addition of the group and interaction terms to the equation. All effect-size values were reported using both the Zumbo and Thomas (1997) and Jodoin and Gierl (2001) indices for reflecting a negligible, moderate, or large effect. The Zumbo and Thomas thresholds for classifying DIF effect size are based on Cohen's (1992) guidelines for identifying a small, medium, or large effect. The thresholds for each level are .13 and .26; values less than .13 have a negligible effect, values between .13 and .26 have a moderate effect, and values of .26 or greater have a large effect. The Jodoin and Gierl thresholds are more stringent, with lower threshold values of .035 and .07 to distinguish between negligible, moderate, and large effects.

### 3.2.3.2. Results

Using the above criteria for inclusion, 471 (87%) items were selected for gender, and 471 (87%) items were selected for at least one racial group comparison. The number of items evaluated by grade in science for gender ranged from 148 in grade 3–5 to 164 in grade 6–8. The number of items evaluated by grade in science for race ranged from 148 in grade 3–5 to 164 in grade 6–8. Because students taking DLM assessments represent seven possible racial groups,<sup>3</sup> there are up to six comparisons that can be made for each item, with the white group as the reference group and each of the other six groups (i.e., African American, Asian, American Indian, Native Hawaiian or Pacific Islander, Alaska Native, two or more races) as the focal group. Across all items, this results in 3,246 possible comparisons. Using the inclusion criteria specified above, 1,869 (58%) item and focal group comparisons were selected for analysis. Overall, five items were evaluated for two racial focal groups, 18 items were evaluated for three racial focal groups, 435 items were evaluated for four racial focal groups, and 13 items were evaluated for five racial focal groups. One racial focal group and the white reference group were used in each comparison. Table 3.20 shows the number of items that were evaluated for each racial focal group. Across all gender and race comparisons, sample sizes for each comparison ranged from 2,929 to 21,257 for gender and from 2,076 to 17,745 for race.

<sup>3</sup> See Chapter 7 of this manual for a summary of participation by race and other demographic variables.

**Table 3.20**

*Number of Items Evaluated for Each Race*

Focal group	Items ( <i>n</i> )
African American	471
American Indian	448
Asian	466
Native Hawaiian or Pacific Islander	13
Two or more races	471

Of the 70 items (13% of the operational item pool) that were not included in the DIF analysis for gender, 70 (100%) had a focal group sample size of less than 100. A total of 70 items were not included in the DIF analysis for race for any of the subgroups. Of the 1,377 item and focal group comparisons that were not included in the DIF analysis for race, 1,361 (99%) had a focal group sample size of less than 100 and 16 (1%) had a subgroup *p*-value greater than .97. Table 3.21 and Table 3.22 show the number and percentage of items that did not meet each inclusion criteria for gender and race, respectively, by the linkage level the items assess.

**Table 3.21**

*Comparisons Not Included in Differential Item Functioning Analysis for Gender, by Linkage Level*

Subject	Sample size		Item proportion correct		Subgroup proportion correct	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Initial	22	31.4	0	0.0	0	0.0
Precursor	25	35.7	0	0.0	0	0.0
Target	23	32.9	0	0.0	0	0.0

**Table 3.22**

*Comparisons Not Included in Differential Item Functioning Analysis for Race, by Linkage Level*

Subject	Sample size		Item proportion correct		Subgroup proportion correct	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Initial	499	36.7	0	0.0	0	0.0
Precursor	569	41.8	0	0.0	7	43.8
Target	293	21.5	0	0.0	9	56.2



### 3.2.3.2.1. Uniform Differential Item Functioning Model

A total of 108 items for gender were flagged for evidence of uniform DIF when comparing  $M_0$  to  $M_1$ . Additionally, 257 item and focal group combinations across 196 items for race were flagged for evidence of uniform DIF. Table 3.23 and Table 3.24 summarize the total number of combinations flagged for evidence of uniform DIF by grade for gender and race, respectively. The percentage of combinations flagged for uniform DIF ranged from 16% to 27% for gender and from 12% to 15% for race.

**Table 3.23**

*Combinations Flagged for Evidence of Uniform Differential Item Functioning for Gender*

Grade	Items flagged ( <i>n</i> )	Total items ( <i>N</i> )	Items flagged (%)	Items with moderate or large effect size ( <i>n</i> )
3–5	23	148	15.5	0
6–8	45	164	27.4	0
9–12	40	159	25.2	0

**Table 3.24**

*Combinations Flagged for Evidence of Uniform Differential Item Functioning for Race*

Grade	Items flagged ( <i>n</i> )	Total items ( <i>N</i> )	Items flagged (%)	Items with moderate or large effect size ( <i>n</i> )
3–5	71	578	12.3	0
6–8	91	649	14.0	1
9–12	95	642	14.8	0

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

The results of the DIF analyses for race were similar to those for gender. When using the Zumbo and Thomas (1997) effect-size classification criteria, all but one combination were found to have a negligible effect-size change after the race term was added to the regression equation. Similarly, when using the Jodoin and Gierl (2001) effect-size classification criteria, all but one combination were found to have a negligible effect-size change after the race term was added to the regression equation.

Table 3.25 provides information about the flagged items with a non-negligible effect-size change after the addition of the group term, as represented by a value of B (moderate) or C (large). The  $\beta_2G$  values in

Table 3.25 indicate which group was favored on the item after accounting for total linkage levels mastered, with positive values indicating that the focal group had a higher probability of success on the item and negative values indicating that the focal group had a lower probability of success on the item. The focal group was favored on one combination.

**Table 3.25**

*Combinations Flagged for Uniform DIF With Moderate or Large Effect Size*

Item ID	Focal	Grade	EE	$\chi^2$	<i>p</i> -value	$\beta_2G$	$R^2$	Z&T*	J&G*
51571	Asian	6–8	SCI.EE.MS.ESS3-3	15.02	< .001	0.31	.901	C	C

Note. EE = Essential Element; Z&T = Zumbo & Thomas; J&G = Jodoin & Gierl.

\* Effect-size measure.

### 3.2.3.2.2. Combined Model

A total of 139 items were flagged for evidence of DIF when both the gender and interaction terms were included in the regression equation, as shown in Equation 3.3. Additionally, 291 item and focal group combinations across 217 items were flagged for evidence of DIF when both the race and interaction terms were included in the regression equation. Table 3.26 and Table 3.27 summarize the number of combinations flagged by grade. The percentage of combinations flagged ranged from 26% to 34% for gender and from 13% to 18% for race.

**Table 3.26**

*Items Flagged for Evidence of Differential Item Functioning for the Combined Model for Gender*

Grade	Items flagged ( <i>n</i> )	Total items ( <i>N</i> )	Items flagged (%)	Items with moderate or large effect size ( <i>n</i> )
3–5	41	148	27.7	0
6–8	56	164	34.1	0
9–12	42	159	26.4	0

**Table 3.27**

*Items Flagged for Evidence of Differential Item Functioning for the Combined Model for Race*

Grade	Items flagged ( <i>n</i> )	Total items ( <i>N</i> )	Items flagged (%)	Items with moderate or large effect size ( <i>n</i> )
3–5	78	578	13.5	0
6–8	95	649	14.6	1
9–12	118	642	18.4	0

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

The results of the DIF analyses for race were similar to those for gender. When using the Zumbo and Thomas (1997) effect-size classification criteria, all but one combination were found to have a negligible effect-size change after the race and interaction terms were added to the regression equation. Similarly, when using the Jodoin and Gierl (2001) effect-size classification criteria, all but one combination were found to have a negligible effect-size change after the race and interaction terms were added to the regression equation.

Information about the flagged items with a non-negligible change in effect size after adding both the group and interaction term is summarized in Table 3.28, where B indicates a moderate effect size, and C a large effect size. In total, one combination had a large effect size. The combination flagged for DIF for the combined model is the same combination flagged for DIF for the uniform model. The  $\beta_3XG$  values in Table 3.28 indicate which group was favored at lower and higher numbers of linkage levels mastered. All combinations favored the focal group higher numbers of total linkage levels mastered and the reference group at lower numbers of total linkage levels mastered.

**Table 3.28**

*Combinations Flagged for DIF With Moderate or Large Effect Size for the Combined Model*

Item ID	Focal	Grade	EE	$\chi^2$	<i>p</i> -value	$\beta_2G$	$\beta_3XG$	$R^2$	Z&T*	J&G*
51571	Asian	6–8	SCI.EE.MS.ESS3-3	15.02	< .001	0.31	0.00	.901	C	C

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

### **3.3. Conclusion**

During the 2021–2022 academic year, the test development teams conducted reduced, virtual events for both item-writing and external review. Overall, 270 testlets were written for science. Additionally, following external review, the science test development team made 31 minor revisions, 250 major revisions to items, and rejected 26 testlets. Of the content already in the operational pool, most items had  $p$ -values within two standard deviations of the mean for the EE and linkage level, and only one item was flagged for non-negligible DIF. Field testing in 2021–2022 focused on collecting data to refresh the operational pool of testlets.

## 4. Assessment Delivery

Chapter 4 of the Dynamic Learning Maps® (DLM®) Alternate Assessment System *2015–2016 Technical Manual—Science* (DLM Consortium, 2017) describes general test administration and monitoring procedures. This chapter describes updated procedures and data collected in 2021–2022, including a summary of administration time, adaptive routing, Personal Needs and Preferences Profile selections, and test administrator survey responses regarding user experience.

Overall, administration features remained consistent with the 2020–2021 intended implementation, including the availability of instructionally embedded testlets, spring operational administration of testlets, the use of adaptive delivery during the spring window, and the availability of accessibility supports.

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

### 4.1. Key Features of the Science Assessment Model

This section describes DLM test administration for 2021–2022. For a complete description of key administration features, including information on assessment delivery, the Kite Suite®, and linkage level assignment, see Chapter 4 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017). Additional information about changes in administration can also be found in the *Test Administration Manual* (DLM Consortium, 2021a) and the *Educator Portal User Guide* (DLM Consortium, 2021d).

#### 4.1.1. Assessment Administration Windows

Assessments are administered in the spring assessment window for operational reporting. Optional assessments are available during the instructionally embedded assessment window for educators to administer for formative information. Additional descriptions of how Essential Elements (EEs) and linkage levels are assigned during the spring assessment window can be found in the Adaptive Delivery section later in this chapter.

##### 4.1.1.1. Instructionally Embedded Assessment Window

During the instructionally embedded assessment window, testlets are optionally available for test administrators to assign to their students. When choosing to administer the optional testlets during the instructionally embedded assessment window, educators decide which EEs and linkage levels to assess for each student. The assessment delivery system recommends a linkage level for each EE based on the educator's responses to the student's First Contact survey, but educators can choose a different linkage level based on their own professional judgment. The dates for the instructionally embedded assessment window are determined by which assessment model each state participates in for English language arts (ELA) and mathematics (i.e., Instructionally Embedded or Year-End). States that only participate in the science assessments follow the dates for the Year-End model. In 2021–2022, the instructionally embedded assessment window occurred between September 13, 2021, and February 23, 2022, for states who participate in the Year-End model and between September 13, 2021, and December 17, 2021, for states who participate in the Instructionally Embedded model. States were given the option of using the entire window or setting their own dates within the larger window. Across all states, the instructionally

embedded assessment window ranged from 4–23 weeks.

#### **4.1.1.2. Spring Assessment Window**

During the spring assessment window, students are assessed on all of the EEs on the assessment blueprint in science. The linkage level for each EE is determined by the system. As with the instructionally embedded assessment window, dates for the spring assessment window are determined by which assessment model is used for ELA and mathematics. In 2021–2022, the spring assessment window occurred between March 14, 2022, and June 10, 2022, for states who participate in the Year-End model and between February 7, 2022, and May 20, 2022, for states who participate in the Instructionally Embedded model. States were given the option of using the entire window or setting their own dates within the larger window. Across all states, the spring assessment window ranged from 6–15 weeks.

### **4.2. Evidence from the DLM System**

This section describes evidence collected by the DLM System during the 2021–2022 operational administration of the DLM alternate assessment. The categories of evidence include data relating to administration time, device usage, adaptive routing, and accessibility support selections.

#### **4.2.1. Administration Time**

Estimated administration time varies by student and subject. Testlets can be administered separately across multiple testing sessions as long as they are all completed within the testing window.

The published estimated total testing time per testlet is around 5–15 minutes. The estimated total testing time is 45–135 minutes per student in the spring assessment window. Published estimates are slightly longer than anticipated real testing times because of the assumption that test administrators need time for setup. Actual testing time per testlet varies depending on each student's unique characteristics.

Kite Student Portal captured start dates, end dates, and time stamps for every testlet. The difference between these start and end times was calculated for each completed testlet. Table 4.1 summarizes the distribution of test times per testlet. The distribution of test times in Table 4.1 is consistent with the distribution observed in prior years. Most testlets took around three minutes or less to complete. Time per testlet may have been impacted by student breaks during the assessment. Testlets with shorter than expected administration times are included in an extract made available to each state. States can use this information to monitor assessment administration and address as necessary. For a description of the administration time monitoring extract, see section 4.3.2 of this chapter.

**Table 4.1**

*Distribution of Response Times per Testlet in Minutes*

Grade	Min	Median	Mean	Max	25Q	75Q	IQR
Elementary	.017	2.17	2.99	90.00	1.33	3.50	2.17
Middle school	.067	1.95	2.71	89.67	1.18	3.17	1.98
High school	.117	2.17	2.97	89.78	1.32	3.45	2.13
Biology	.117	2.13	2.84	88.85	1.33	3.38	2.05

*Note.* Min = minimum, Max = maximum, 25Q = lower quartile, 75Q = upper quartile, IQR = interquartile range.

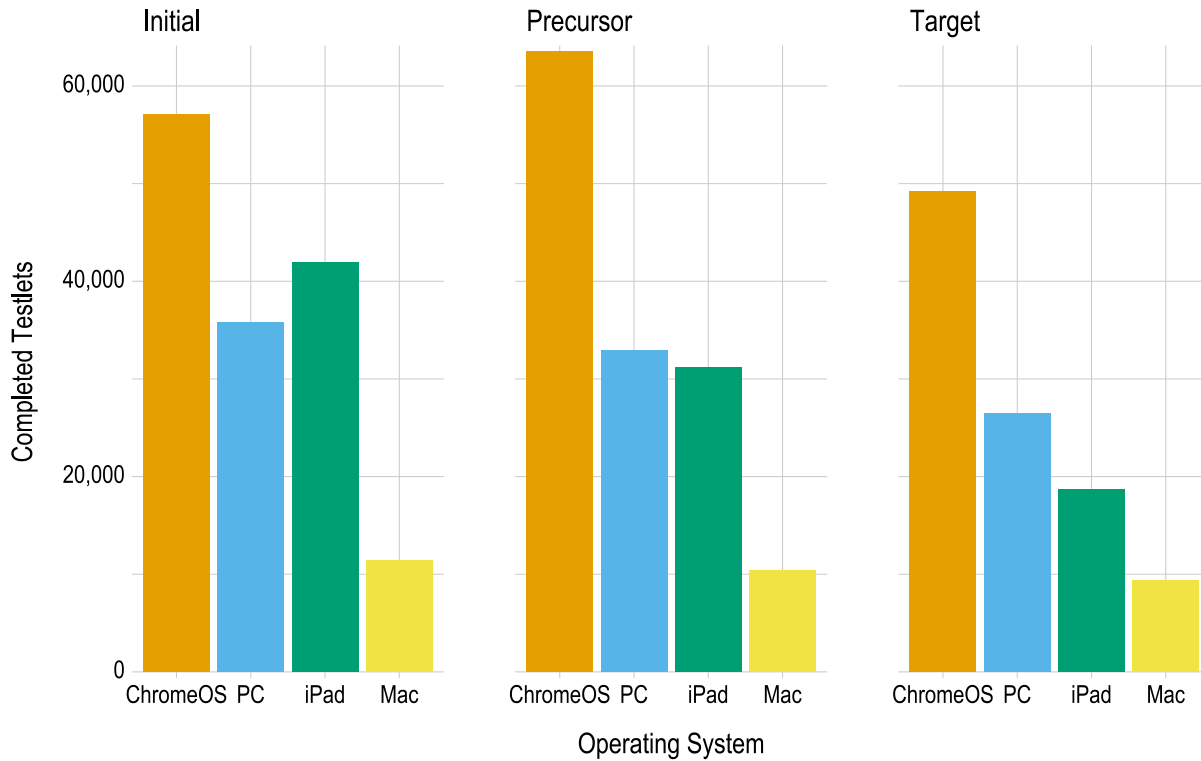
### **4.2.2. Device Usage**

Testlets may be administered on a variety of devices. Kite Student Portal captured the operating system used for each testlet completed. Although these data do not capture specific devices used to complete each testlet (e.g., SMART Board, switch system, etc.), they provide high-level information about how students access assessment content. For example, we can identify how often an iPad is used relative to a Chromebook or traditional PC. Figure 4.1 shows the number of testlets completed on each operating system by subject and linkage level for 2021–2022. Overall, 44% of testlets were completed on a Chromebook, 25% were completed on a PC, 24% were completed on an iPad, and 8% were completed on a Mac.



**Figure 4.1**

*Distribution of Devices Used for Completed Testlets*



### 4.2.3. Blueprint Coverage

Each student is assessed on all EEs included on the assessment blueprint.<sup>4</sup> Table 4.2 summarizes the number of EEs required for each grade or course.

**Table 4.2**

*Essential Elements Required for Blueprint Coverage*

Grade or Course	<i>n</i>
Elementary	9
Middle school	9
High school	9
Biology	10

Across all grades, 94% of students were assessed on all of the EEs and met blueprint requirements. Table 4.3 summarizes the total number of students and the percentage of students meeting blueprint

<sup>4</sup> For a description of the assessment blueprints see Chapter 3 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

requirements based on their complexity band. When comparing complexity band distributions, there was a slightly lower percentage of Foundational students not meeting requirements. However, all complexity band groups had over 91% of students meeting the coverage requirements.

**Table 4.3**

*Student Blueprint Coverage by Complexity Band*

Complexity Band	<i>n</i>	% meeting requirements
Foundational	6,986	91.2
Band 1	18,000	94.0
Band 2	13,573	94.8
Band 3	5,816	94.9

#### **4.2.4. Adaptive Delivery**

During the spring 2022 test administration, the science assessments were 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 4.4.

**Table 4.4**

*Correspondence of Complexity Bands and Linkage Levels*

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

Following the spring 2022 administration, analyses were conducted to determine the mean percentage of testlets that adapted from the first to second testlet administered for students within a grade or course and complexity band. The aggregated results can be seen in Table 4.5.

Due to small sample size, data regarding the adaptation of linkage levels was unavailable for Band 3 in grade 3. For the majority of students across all grades who were assigned to the Foundational Complexity Band by the First Contact survey, testlets did not adapt to a higher linkage level after the first assigned testlet (ranging from 57% to 71%). A similar pattern was seen for students assigned to Band 3, with the majority of students not adapting down to a lower linkage level after the first assigned testlet (ranging from 61% to 81%). In contrast, students assigned to Band 1 tend to adapt up to a higher linkage level after their first testlet (ranging from 54% to 74%). Consistent patterns were not as apparent for students who were assigned to Band 2. Results indicate that linkage levels of students assigned to higher complexity bands are more variable with respect to the direction in which students move between the first and second testlets. However, this finding of more variability in the higher complexity bands is consistent with prior years, which showed the same trend. Several factors may help explain these results, including more variability in student characteristics within this group and content-based differences across grades. For a description of previous findings, see Chapter 4 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017) and the subsequent technical manual updates (DLM Consortium, 2018a, 2018b, 2019, 2021b).

**Table 4.5**

*Adaptation of Linkage Levels Between First and Second Science Testlets (N = 44,375)*

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 (%)
Grade 3	39.9	60.1	68.1	31.9	22.5	46.7	30.8	*	*
Grade 4	43.0	57.0	73.3	26.7	25.2	46.3	28.5	63.1	36.9
Grade 5	40.3	59.7	73.7	26.3	27.5	43.2	29.4	64.8	35.2
Grade 6	32.7	67.3	70.8	29.2	32.9	40.0	27.1	61.0	39.0
Grade 7	34.2	65.8	73.1	26.9	34.8	38.8	26.4	61.6	38.4
Grade 8	38.8	61.2	71.1	28.9	38.3	41.7	20.0	67.0	33.0
Grade 9	33.8	66.2	60.7	39.3	45.7	38.7	15.5	80.6	19.4
Grade 10	29.0	71.0	61.3	38.7	40.1	40.9	19.0	80.7	19.3
Grade 11	32.9	67.1	57.0	43.0	40.9	40.7	18.5	78.8	21.2
Grade 12	29.3	70.7	54.0	46.0	37.7	40.2	22.1	74.2	25.8
Biology	32.9	67.1	54.1	45.9	20.5	43.1	36.4	62.8	37.2

\* These data were suppressed because  $n < 50$ .

*Note.* Foundational and Band 1 correspond to the 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.

After the second testlet is administered, testlets continue to adapt based on the same routing rules. Table 4.6 shows the total number and percentage of testlets that were assigned at each linkage level during the spring assessment window. Testlets were fairly evenly distributed across the three linkage levels, with the Initial and Precursor linkage levels being assigned slightly more often.

**Table 4.6**

*Distribution of Linkage Levels Assigned for Assessment*

Linkage level	<i>n</i>	%
Initial	146,235	37.7
Precursor	138,060	35.6
Target	103,795	26.7

#### **4.2.5. Administration Incidents**

DLM staff annually evaluates testlet assignment to ensure students are correctly assigned to testlets. Administration incidents that have the potential to affect scoring are reported to state education agencies in a supplemental Incident File. No incidents were observed during the 2021–2022 operational assessment windows. Assignment of testlets will continue to be monitored in subsequent years to track any potential incidents and report them to state education agencies.

#### **4.2.6. Accessibility Support Selections**

Accessibility supports provided in 2021–2022 were the same as those available in previous years. The DLM *Accessibility Manual* (DLM Consortium, 2021c) distinguishes 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 4.7 shows selection rates for the three categories of accessibility supports. Overall, 43,870 students (90%) had at least one support selected. The most commonly selected supports in 2021–2022 were human read aloud, spoken audio, and test administrator enters responses for student. For a complete description of the available accessibility supports, see Chapter 4 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017). Additionally, educators reported in the First Contact survey that 42% of students were able to access a computer independently, with or without assistive technology.

**Table 4.7**

*Accessibility Supports Selected for Students (N = 48,663)*

Support	<i>n</i>	%
Supports provided in Kite Student Portal		
Spoken audio	27,324	56.1
Magnification	7,186	14.8
Color contrast	4,507	9.3
Overlay color	2,141	4.4
Invert color choice	1,433	2.9
Supports requiring additional tools/materials		
Individualized manipulatives	20,474	42.1
Calculator	16,007	32.9
Single-switch system	2,042	4.2
Alternate form - visual impairment	1,201	2.5
Two-switch system	638	1.3
Uncontracted braille	59	0.1
Supports provided outside the system		
Human read aloud	38,757	79.6
Test administrator enters responses for student	26,590	54.6
Partner-assisted scanning	4,025	8.3
Language translation of text	766	1.6
Sign interpretation of text	680	1.4

### **4.3. Evidence From Monitoring Assessment Administration**

Monitoring of assessment administration was conducted using various materials and strategies. DLM project staff developed an assessment administration monitoring protocol for use by DLM staff, state education agency staff, and local education agency staff. Project staff also reviewed Service Desk contacts and hosted regular check-in calls to monitor common issues and concerns during the assessment window. This section provides an overview of all resources and supports as well as more detail regarding the assessment administration observation protocol and its use, check-in calls with states, and methods for monitoring testlet delivery.

#### **4.3.1. Test Administration Observations**

DLM project staff developed an assessment administration observation protocol to standardize data collection across observers and locations. This assessment administration protocol is available for use by state and local education agencies; however, participation in the test administration observations is not required. The majority of items in the protocol are based on direct recording of what is observed and require little inference or background knowledge. Information from the protocol is used to evaluate several assumptions in the validity argument, addressed in the Test Administration Observation Results section of this chapter.

One observation form is completed per testlet administered. Some items are differentiated for computer-delivered and educator-administered testlets. The four main sections include Preparation/Set Up, Administration, Accessibility, and Observer Evaluation. The Preparation/Set Up section includes documentation of the testing location, testing conditions, the testing device used for the testing session, and documentation of the test administrator’s preparation for the session. The Administration section is provided for the documentation of the student’s response mode, general test administrator behaviors during the session, subject-specific test administrator behaviors, any technical problems experienced with the Kite Suite, and documentation of student completion of the testlet. The Accessibility section focuses on the use of accessibility features, any difficulty the student encountered with the accessibility features, and any additional devices the student uses during the testing session. Finally, Observer Evaluation requires that the observer rate overall student engagement during the session and provide any additional relevant comments.

The protocol is available as an online survey (optimized for mobile devices and with branching logic) administered through Kite Survey Solutions, a survey platform within the Kite Suite.

Training resources are provided to state education agency staff to support fidelity of use of the assessment administration protocol and increase the reliability of data collected (see Table 4.8). State education agency staff have access to the *Test Administration Observation Training* video on the use of the *Test Administration Observation Protocol*. The links to this video, the *Guidance for Local Observers*, and the *Test Administrator Observation Protocol* are provided on the state side of the DLM website, and state education agencies are encouraged to use this information in their state monitoring efforts. State education agencies are able to use these training resources to encourage use of the protocol among local education agency staff. States are also cautioned that the protocol is only to be used to document observations for the purpose of describing the administration process. It is not to be used for evaluating or coaching test administrators or gauging student academic performance. This caution, as well as general instructions for completing and submitting the protocol, are provided in the form itself.

**Table 4.8**

*DLM Resources for Test Administration Monitoring Efforts*

Resource	Description
DLM Test Administration Observation Research Protocol (PDF)	Provides observers with a standardized way to describe the assessment administration.
Guide to Test Administration Observations: Guidance for Local Observers (PDF)	Provides observers with the purpose and use of the observation protocol as well as general instructions for use.
Test Administration Observation Training Video (Vimeo video)	Provides training on the use of the <i>Test Administration Observation Protocol</i> .

During 2021–2022, there were 48 assessment administration observations collected in six states. Table 4.9 shows the number of observations collected by state. Of the observations, 32 (67%) were of

computer-delivered assessments and 16 (33%) were of educator-administered testlets.

**Table 4.9**

*Educator Observations by State (N = 48)*

State	<i>n</i>	%
Arkansas	36	75.0
Missouri	2	4.2
West Virginia	10	20.8

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 4.10; behaviors were identified as supporting, neutral, or nonsupporting. For example, clarifying directions (62.5% 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 test administrator directly influenced the student’s answer choice. Overall, 74% of observed behaviors were classified as supporting, with 1% of observed behaviors reflecting nonsupporting actions.



**Table 4.10**

*Test Administrator Actions During Computer-Delivered Testlets (n = 32)*

Action	n	%
<b>Supporting</b>		
Clarified directions or expectations for the student	20	62.5
Read one or more screens aloud to the student	16	50.0
Navigated one or more screens for the student	12	37.5
Repeated question(s) before student responded	7	21.9
<b>Neutral</b>		
Used pointing or gestures to direct student attention or engagement	6	18.8
Used materials or manipulatives during the administration process	4	12.5
Used verbal prompts to direct the student’s attention or engagement (e.g., “look at this.”)	4	12.5
Asked the student to clarify or confirm one or more responses	2	6.2
Allowed student to take a break during the testlet	1	3.1
Entered one or more responses for the student	1	3.1
Repeated question(s) after student responded (gave a second trial at the same item)	0	0.0
<b>Nonsupporting</b>		
Physically guided the student to a response	1	3.1
Reduced the number of answer choices available to the student	0	0.0

*Note.* Respondents could select multiple responses to this question.

For DLM assessments, interaction with the system includes interaction with the assessment content as well as physical access to the testing device and platform. The fact that educators navigated one or more screens in 38% 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 the 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 4.11. Independent response selection was observed in 56% 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 3% 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 4.11**

*Student Actions During Computer-Delivered Testlets (n = 32)*

Action	n	%
Selected answers independently	18	56.2
Navigated screens independently	15	46.9
Selected answers after verbal prompts	10	31.2
Navigated screens after verbal prompts	8	25.0
Navigated screens after test administrator pointed or gestured	3	9.4
Skipped one or more items	3	9.4
Asked the test administrator a question	2	6.2
Independently revisited a question after answering it	2	6.2
Used materials outside of Kite Student Portal to indicate responses to testlet items	1	3.1
Revisited one or more questions after verbal prompt(s)	0	0.0

*Note.* Respondents could select multiple responses to this question.

Another assumption in the validity argument is that students are able to respond to tasks irrespective of sensory, mobility, health, communication, or behavioral constraints. This assumption was evaluated by having observers note whether there was difficulty with accessibility supports (including lack of appropriate available supports) during observations of educator-administered testlets. Of the 16 observations of educator-administered testlets, observers noted difficulty in 1 case (6%). For computer-delivered testlets, evidence to evaluate the assumption was collected by noting students who indicated responses to items using varied response modes such as gesturing (12%) and using manipulatives or materials outside of Kite (3%). Additional evidence for this assumption was gathered by observing whether students were able to complete testlets. Of the 48 test administration observations collected, students completed the testlet in 31 cases (65%).<sup>5</sup>

Finally, the test administration observations allow for an evaluation of the assumption 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 4.12 summarizes students' response modes for educator-administered testlets. The most frequently observed behavior was *gestured to indicate response to test administrator who selected answers*.

<sup>5</sup> In all instances where the testlet was not completed, no reason was provided by the observer.

**Table 4.12**

*Primary Response Mode for Educator-Administered Testlets (n = 16)*

Response mode	n	%
Gestured to indicate response to test administrator who selected answers	10	62.5
Verbally indicated response to test administrator who selected answers	6	37.5
No observable response mode	2	12.5
Eye gaze system indication to test administrator who selected answers	1	6.2

*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 PNP 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 1 of 32 (3%) observations of computer-delivered testlets, the test administrator entered responses on the student’s behalf. In all cases, the observers indicated that the entered response matched the student’s response.

### **4.3.2. Data Forensics Monitoring**

Two data forensics monitoring reports are available in Educator Portal. The first report includes information about testlets completed outside of normal business hours. The second report includes information about testlets that were completed within a short period of time.

The Testing Outside of Hours report allows state education agencies to specify days and hours within a day that testlets are expected to be completed. Each state can select its own days and hours for setting expectations. For example, a state could elect to flag any testlet completed outside of Monday through Friday from 6:00 a.m. to 5:00 p.m. local time. The Testing Outside of Hours report then identifies students who completed assessments outside of the defined expected hours. Overall, 2,812 (1%) science testlets were completed outside of the expected hours by 2,411 (5%) students.

The Testing Completed in a Short Period of Time report identifies students who completed a testlet within an unexpectedly short period of time. The threshold for inclusion in the report was testlet completion time of less than 30 seconds. The report is intended for state users to identify potentially aberrant response patterns; however there are many legitimate reasons a testlet may be submitted in a short time period. Overall, 10,262 (3%) testlets were completed in a short period of time by 4,904 (11%) students.

## **4.4. Evidence From Test Administrators**

This section first describes evidence collected from the spring 2022 test administrator survey. Data on user experience with the DLM System as well as student opportunity to learn is evaluated annually through a survey that test administrators are invited to complete after administration of the spring assessment. Test administrators receive one survey per rostered DLM student, which collects information about that student’s assessment experience. As in previous years, the survey was distributed to test administrators in

Kite Student Portal, where students completed assessments. The survey consisted of four blocks. Blocks 1 and 4 were administered in every survey. Block 1 included questions about the test administrator's perceptions of the assessments and the student's interaction with the content, and Block 4 included questions about the test administrator's background. Block 2 was spiraled, so test administrators received one randomly assigned section. In these sections, test administrators were asked about one of the following topics per survey: relationship to ELA instruction, relationship to mathematics instruction, or relationship to science instruction. Block 3 was added in 2021 and remained in the survey in 2022 to gather information about educational experiences during the COVID-19 pandemic.

#### **4.4.1. User Experience With the DLM System**

A total of 13,031 test administrators responded to the survey (67%) about 27,162 students' experiences. Test administrators are instructed to respond to the survey separately for each of their students. Participating test administrators responded to surveys for a median of two students. Test administrators reported having an average of 10 years of experience in science and 11 years of experience with students with significant cognitive disabilities.

The following sections summarize responses regarding both educator and student experience with the system.

##### **4.4.1.1. Educator Experience**

Test administrators 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 test administrators to respond on a 4-point scale: *strongly disagree*, *disagree*, *agree*, or *strongly agree*. Responses are summarized in Table 4.13.

Nearly all test administrators (96%) agreed or strongly agreed that they were confident administering DLM testlets. Most respondents (90%) agreed or strongly agreed that the required test administrator training prepared them for their responsibilities as test administrators. Most test administrators 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 (90%).

**Table 4.13**

*Test Administrator Responses Regarding Test Administration*

Statement	SD		D		A		SA		A+SA	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
I was confident in my ability to deliver DLM testlets.	121	1.4	248	3.0	3,458	41.4	4,520	54.2	7,978	95.6
Required test administrator training prepared me for the responsibilities of a test administrator.	204	2.5	632	7.6	4,187	50.3	3,301	39.7	7,488	90.0
I have access to curriculum aligned with the content measured by DLM assessments.	270	3.2	915	11.0	4,194	50.4	2,936	35.3	7,130	85.7
I used manuals and/or the DLM Educator Resource Page materials.	186	2.2	633	7.6	4,597	55.2	2,915	35.0	7,512	90.2

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

**4.4.1.2. Student Experience**

The spring 2022 test administrator survey included three items about how students responded to test items. Test administrators were asked to rate statements from *strongly disagree* to *strongly agree*. Results are presented in Table 4.14. The majority of test administrators agreed or strongly agreed that their students responded to items to the best of their knowledge, skills, and understandings; were able to respond regardless of disability, behavior, or health concerns; and had access to all necessary supports to participate.

**Table 4.14**

*Test Administrator Perceptions of Student Experience with Testlets*

Statement	SD		D		A		SA		A+SA	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Student responded to items to the best of his/her knowledge, skills, and understanding	981	4.0	1,810	7.3	12,930	52.2	9,050	36.5	21,980	88.7
Student was able to respond regardless of his/her disability, behavior, or health concerns	1,532	6.2	2,088	8.4	12,588	50.7	8,623	34.7	21,211	85.4
Student had access to all necessary supports to participate	863	3.5	1,192	4.8	12,898	52.2	9,769	39.5	22,667	91.7

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

Annual survey results show that a small percentage of test administrators disagree that their student was able to respond regardless of disability, behavior, or health concerns; had access to all necessary supports; and was able to effectively use supports. In spring 2020, DLM staff conducted educator focus groups with educators who disagreed with one or more of these survey items to learn about potential accessibility gaps in the DLM System (Kobrin et al., 2022). A total of 18 educators from 11 states participated in six focus groups. The findings revealed that many of the challenges educators described were documented in existing materials (e.g., wanting clarification about allowable practices that are described in the *Test Administration Manual*, such as substituting materials; desired use of not-allowed practices like hand-over-hand that are used during instruction). DLM staff are using the focus group findings to review existing materials and develop new resources that better communicate information about allowable practices to educators.

#### **4.4.2. Opportunity to Learn**

Table 4.15 reports the opportunity to learn results. Approximately 54% of responses (*n* = 13,365) 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 4.15**

*Educator Ratings of Portion of Testlets That Matched Instruction*

Subject	None		Some (< half)		Most (> half)		All		Not applicable	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Science	2,341	9.5	7,179	29.2	8,146	33.1	5,219	21.2	1,737	7.1

A subset of test administrators were asked to indicate the approximate number of hours they spent instructing students on each of the DLM science core ideas and in the science and engineering practices. Educators responded using a 6-point scale: 0 hours, 1–5 hours, 6–10 hours, 11–15 hours, 16–20 hours, or more than 20 hours. Table 4.16 and Table 4.17 indicate the amount of instructional time spent on DLM science core ideas and science and engineering practices, respectively. For all science core ideas and science and engineering practices, the most commonly selected response was 1–5 hours.

**Table 4.16**

*Instructional Time Spent on Science Core Ideas*

Core Idea	Median	Number of hours											
		0		1–5		6–10		11–15		16–20		>20	
		<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
<b>Physical Science</b>													
Matter and its interactions	1–5 hours	1,826	24.6	2,351	31.7	1,281	17.3	809	10.9	644	8.7	500	6.7
Motion and stability: Forces and interactions	1–5 hours	2,071	28.2	2,339	31.8	1,236	16.8	750	10.2	546	7.4	412	5.6
Energy	1–5 hours	1,862	25.5	2,314	31.7	1,266	17.3	784	10.7	633	8.7	446	6.1
<b>Life Science</b>													
From molecules to organisms: Structures and processes	1–5 hours	2,515	34.3	2,103	28.7	1,096	15.0	685	9.4	524	7.2	402	5.5
Ecosystems: Interactions, energy, and dynamics	1–5 hours	1,740	23.7	2,209	30.1	1,309	17.8	841	11.4	697	9.5	555	7.5
Heredity: Inheritance and variation of traits	1–5 hours	2,967	40.5	2,047	27.9	969	13.2	549	7.5	461	6.3	332	4.5
Biological evolution: Unity and diversity	1–5 hours	2,685	36.8	2,119	29.0	1,047	14.3	610	8.3	494	6.8	351	4.8
<b>Earth and Space Science</b>													
Earth’s place in the universe	1–5 hours	1,922	26.2	2,268	30.9	1,233	16.8	810	11.1	614	8.4	482	6.6
Earth’s systems	1–5 hours	1,936	26.4	2,283	31.1	1,195	16.3	798	10.9	619	8.4	505	6.9
Earth and human activity	1–5 hours	1,718	23.4	2,328	31.7	1,261	17.2	839	11.4	673	9.2	526	7.2



**Table 4.17**

*Instructional Time Spent on Science and Engineering Practices*

Science and engineering practice	Median	Number of hours											
		0		1–5		6–10		11–15		16–20		>20	
		<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Developing and using models	1–5 hours	1,964	26.7	2,566	34.9	1,200	16.3	664	9.0	536	7.3	430	5.8
Planning and carrying out investigations	1–5 hours	1,755	23.9	2,450	33.4	1,315	17.9	758	10.3	584	8.0	472	6.4
Analyzing and interpreting data	1–5 hours	1,482	20.3	2,373	32.5	1,361	18.6	811	11.1	689	9.4	595	8.1
Using mathematics and computational thinking	6–10 hours	1,396	19.1	2,229	30.5	1,278	17.5	815	11.2	721	9.9	866	11.9
Constructing explanations and designing solutions	1–5 hours	2,158	29.5	2,309	31.5	1,174	16.0	711	9.7	547	7.5	423	5.8
Engaging in argument from evidence	1–5 hours	2,619	35.7	2,210	30.2	1,045	14.3	593	8.1	481	6.6	380	5.2
Obtaining, evaluating, and communicating information	1–5 hours	1,706	23.2	2,269	30.9	1,228	16.7	789	10.7	694	9.5	657	8.9

Results from the test administrator survey were also correlated with total linkage levels mastered by grade band. The median of instructional time was calculated for each student across from educator 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 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 4.18 summarizes the Spearman rank-order correlations between instructional time and the total number linkage levels mastered, by grade band and course. Correlations ranged from 0.14 to 0.26. Based on guidelines from Cohen (1988), the observed correlations were small. However, the correlation for Biology is based on data from only 219 students who both participated in the Biology assessment and had this block of the test administrator survey completed. Thus, these results should be interpreted with caution.

**Table 4.18**

*Correlation Between Instruction Time in Science Linkage Levels Mastered*

Grade band	Correlation with instructional time
Elementary	.175
Middle school	.157
High school	.140
Biology	.256

Another dimension of opportunity to learn is student engagement with instruction. The First Contact survey contains two questions about student engagement during computer- and educator-directed instruction. Table 4.19 shows the percentage of students who demonstrated different levels of attention by instruction type. Overall, 87% of students demonstrated fleeting or sustained attention to computer-directed instruction and 86% of students demonstrated fleeting or sustained attention to educator-directed instruction.

**Table 4.19**

*Student Attention Levels During Instruction*

Type of instruction	Demonstrates little or no attention		Demonstrates fleeting attention		Generally sustains attention	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Computer-directed ( <i>n</i> = 42,089)	5,347	12.7	23,037	54.7	13,705	32.6
Educator-directed ( <i>n</i> = 45,686)	6,374	14.0	27,997	61.3	11,315	24.8

## **4.5. Conclusion**

Delivery of the DLM System was designed to align with instructional practice and be responsive to individual student needs. Assessment delivery options allow for necessary flexibility to reflect student needs while also including constraints to maximize comparability and support valid interpretation of results. The dynamic nature of DLM assessment administration is reflected in the initial input through the First Contact survey, as well as adaptive routing between testlets. Evidence collected from the DLM System, test administration monitoring, and test administrators indicates that students are able to successfully interact with the system to demonstrate their knowledge, skills, and understandings.

## 5. Modeling

To provide feedback about student performance, the Dynamic Learning Maps® (DLM®) Alternate Assessment System draws on a well-established research base in cognition and learning theory and relatively uncommon operational psychometric methods. The approach uses innovative, operational psychometric methods to provide feedback about student mastery of skills. This chapter describes the psychometric model that underlies the DLM System and describes the process used to estimate item and student parameters from student assessment data.

### 5.1. Psychometric Background

Learning maps, which are the networks of sequenced learning targets, are at the core of the DLM assessments in English language arts and mathematics. While development of a science learning map is planned for the future development work, the similarity across all subjects in scoring at the linkage level means the general background below is useful for understanding the current science scoring model, even though there is not currently an underlying map.

In general, a learning map is a collection of skills to be mastered that are linked together by connections between the skills. The connections between skills indicate what should be mastered prior to learning additional skills. Together, the skills and their prerequisite connections map out the progression of learning within a given subject. Stated in the vocabulary of traditional psychometric methods, a learning map defines a large set of discrete latent variables indicating students' learning status on key skills and concepts relevant to a large content domain, as well as a series of pathways indicating which topics (represented by latent variables) are prerequisites for learning other topics.

Because of the underlying map structure and the goal of providing more fine-grained information beyond a single raw or scale score value, student results are reported as a profile of skill mastery. This profile is created using diagnostic classification modeling, which draws on research in cognition and learning theory to provide information about student mastery of multiple skills measured by the assessment. Diagnostic classification models (DCMs) are confirmatory latent class models that characterize the relationship of observed responses to a set of categorical latent variables (e.g., Bradshaw, 2016; Rupp et al., 2010). DCMs are also known as cognitive diagnosis models (e.g., Leighton & Gierl, 2007) or multiple classification latent class models (Maris, 1999) and are mathematically equivalent to Bayesian networks (e.g., Almond et al., 2015; Mislevy & Gitomer, 1995; Pearl, 1988). This is the main difference from more traditional psychometric models, such as item response theory, which model a single, continuous latent variable. DCMs provide information about student mastery on multiple latent variables or skills of interest.

DCMs have primarily been used in educational measurement settings in which more detailed information about test-takers' skills is of interest, such as in assessing individual mathematics skills (e.g., Bradshaw et al., 2014), different levels of reading complexity (e.g., Templin & Bradshaw, 2014), and the temporal acquisition of science skills (e.g., Templin & Henson, 2008). To provide detailed profiles of student mastery of the skills, or attributes, measured by the assessment, DCMs require the specification of an item-by-attribute Q-matrix, indicating the attributes measured by each item. In general, for a given item,  $i$ , the Q-matrix vector would be represented as  $q_i = [q_{i1}, q_{i2}, \dots, q_{iA}]$ , where  $A$  is the total number of attributes. Similar to a factor pattern matrix in a confirmatory factor model, Q-matrix indicators are binary: either the item measures an attribute ( $q_{ia} = 1$ ) or it does not ( $q_{ia} = 0$ ).

For each item, there is a set of conditional item-response probabilities that corresponds to the student's possible mastery patterns. Although DCMs can be defined using any number of latent categories for each attribute, it is most common to use binary attributes, which provide more interpretable results to stakeholders (Bradshaw & Levy, 2019). When an item measures a single binary attribute, only two statuses are possible for any examinee: a master of the attribute or a nonmaster of the attribute.

In general, the modeling approach involves specifying the Q-matrix, determining the probability of being classified into each category of mastery (master or nonmaster), and relating those probabilities to students' response data to determine a posterior probability of being classified as a master or nonmaster for each attribute. For DLM assessments, the attributes for which probabilities of mastery are calculated are the Essential Element (EE) linkage levels.

## **5.2. Essential Elements and Linkage Levels**

Because the primary goal of the DLM assessments is to measure what students with the most significant cognitive disabilities know and can do, alternate grade-level expectations called EEs were created to provide students in the population access to the general education grade-level academic content. See Chapter 2 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017) for a complete description. Each EE has an associated set of linkage levels that are ordered by increasing complexity. There are three linkage levels for each science EE: Initial, Precursor, and Target.

## **5.3. Overview of the DLM Modeling Approach**

Many statistical models are available for estimating the probability of mastery for attributes in a DCM. The statistical model used to determine the probability of mastery for each linkage level for DLM assessments is the log-linear cognitive diagnosis model (LCDM). The LCDM is a DCM model that provides a general statistical framework for obtaining probabilities of class membership for each measured attribute (Henson et al., 2009). Student mastery statuses for each linkage level are obtained from a Bayesian estimation procedure, which contributes to an overall profile of mastery.

### **5.3.1. Model Specification**

Each linkage level was calibrated separately for each EE using separate LCDMs. Each linkage level within an EE is estimated separately because of the administration design in which overlapping data from students taking testlets at multiple levels within an EE is uncommon. Also, because items were developed to meet a precise cognitive specification, all master and nonmaster 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. As such, each class (i.e., masters or nonmasters) has a single probability of responding correctly to all items measuring the linkage level, as depicted in Table 5.1. Therefore, for each item measuring the same linkage level, the probability of providing a correct response is held constant for all students in each mastery class. Chapter 3 of this manual details item review procedures intended to support the fungibility assumption, and section 5.4.1 of this chapter describes empirical evidence to support this constraint.

**Table 5.1**

*Depiction of Fungible Item Parameters for Items Measuring a Single Linkage Level*

Item	Class 1 (Nonmasters)	Class 2 (Masters)
1	$\pi_1$	$\pi_2$
2	$\pi_1$	$\pi_2$
3	$\pi_1$	$\pi_2$
4	$\pi_1$	$\pi_2$
5	$\pi_1$	$\pi_2$

*Note.*  $\pi$  represents the probability of providing a correct response.

The DLM scoring model for the 2021–2022 administration was as follows. Each linkage level within each EE was considered the latent variable to be measured (the attribute). Using DCMs, a probability of mastery on a scale of 0 to 1 was calculated for each linkage level within each EE. Students were then classified into one of two classes for each linkage level of each EE: either master or nonmaster. As described in Chapter 6 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017), a posterior probability of at least .8 was required for mastery classification.

The general form of DCMs is shown in Equation 5.1. In Equation 5.1,  $\pi_{ic}$  is the conditional probability of a student in class  $c$  providing a correct response to item  $i$ , and  $x_{ij}$  is the observed response (i.e., 0 or 1) of student  $j$  to item  $i$ . Thus,  $\pi_{ic}^{x_{ij}}(1 - \pi_{ic})^{1-x_{ij}}$  represents the probability of a respondent in class  $c$  providing the observed response to item  $i$ . Finally,  $\nu_c$  represents the base rate probability that any given respondent belongs to class  $c$ .

$$P(X_j = x_j) = \sum_{c=1}^C \nu_c \prod_{i=1}^I \pi_{ic}^{x_{ij}} (1 - \pi_{ic})^{1-x_{ij}} \quad (5.1)$$

Different types of DCMs use different measurement models to define  $\pi_{ic}$  in Equation 5.1. For DLM assessments, item responses are modeled using the LCDM, as described by Henson et al. (2009). The LCDM defines the conditional probabilities using a generalized linear model with a logit link function. Specifically, using the LCDM,  $\pi_{ic}$  is defined as seen in Equation 5.2, where  $\alpha_c$  is a binary indicator of mastery status for a student in class  $c$  for that attribute.

$$\pi_{ic} = P(X_{ic} = 1 | \alpha_c) = \frac{\exp(\lambda_{i,0} + \lambda_{i,1}\alpha_c)}{1 + \exp(\lambda_{i,0} + \lambda_{i,1}\alpha_c)} \quad (5.2)$$

Equation 5.2 utilizes the LCDM notation described by Rupp et al. (2010), where the  $\lambda$  subscripts follow structure of “item, effect”, where the effect is defined as 0 = intercept, 1 = main effect. All items in a linkage level were assumed to measure that linkage level, meaning the Q-matrix for the linkage level was a column of ones. As such, each item measured one latent variable, resulting in two parameters per item: (a) an intercept ( $\lambda_{i,0}$ ) that corresponds to the probability of answering the item correctly for examinees who have not mastered the linkage level and (b) a main effect ( $\lambda_{i,1}$ ) that corresponds to the increase in the

probability of answering the item correctly for examinees who have mastered the linkage level. Because students who have mastered the linkage level should also have a higher probability of providing a correct response than students who have not,  $\lambda_{i,1}$  is constrained to be positive to ensure monotonicity (Henson et al., 2009). As per the assumption of item fungibility, a single set of probabilities was estimated for all items within a linkage level. Therefore, Equation 5.2 can be simplified to remove the item-level  $\lambda$  parameters. Equation 5.3 removes the item-level effects, showing that for each linkage level we now estimate only one intercept shared by all items ( $\lambda_0$ ) and one main effect shared by all items ( $\lambda_1$ ).

$$\pi_{ic} = P(X_{ic} = 1 | \alpha_c) = \frac{\exp(\lambda_0 + \lambda_1 \alpha_c)}{1 + \exp(\lambda_0 + \lambda_1 \alpha_c)} \quad (5.3)$$

Finally, because each linkage level is estimated separately as a single attribute LCDM, there are only two possible mastery classes (i.e., nonmasters and masters). Therefore, only a single structural parameter was needed ( $\nu$ ), which is the probability that a randomly selected student who is assessed on the linkage level is a master (i.e., the analogous map parameter). The base rate of the other class (i.e., nonmastery) is deterministically calculated as  $1 - \nu$ . In total, three parameters per linkage level are specified in the DLM scoring model: a fungible intercept, a fungible main effect, and the proportion of masters.

### 5.3.2. Model Calibration

A Bayesian approach was used to calibrate the DCMs. A Bayesian approach was preferred over a simpler maximum likelihood approach because the posterior distributions derived from Bayesian methods offer more robust methods for evaluating model fit (see section 5.4.1). We specifically selected an empirical Bayes procedure for several reasons. In any Bayesian approach, prior distributions must be specified for each parameter in the model. An Empirical Bayes procedure uses the data to estimate a prior distribution, whereas a standard Bayesian procedure would fix the prior distribution *a priori* (Carlin & Louis, 2001). The empirical priors offer several advantages. First, due to the number of models that are estimated (i.e., 102 linkage levels), an *a priori* specification of prior distributions would require the same priors for each model. By using empirical priors, we can select prior distributions specific to each linkage level, rather than using a single general prior that may be more or less appropriate for any given linkage level, thus increasing the information available in the estimation process (Nabi et al., 2022). Second, if *a priori* priors are used, there are many decisions that a practitioner must make when eliciting the fixed priors. Different decisions lead to different prior distributions, which would then affect the resulting posterior distributions (Falconer et al., 2022; Stefan et al., 2020). Using empirical priors removes the practitioner degrees of freedom that could result in different priors resulting from practitioner decisions. Finally, empirical prior distributions are often more informative than a general prior fixed *a priori*. More informative priors make the estimation more efficient by focusing the sampling more closely on the highest density area of the posterior distribution without biasing the final parameter estimates (Petroni et al., 2014).

Across all grades in science, there were 34 EEs, each with three linkage levels, resulting in a total of  $34 \times 3 = 102$  separate calibration models. Each separate calibration included all operational items for the EE and linkage level. Each model was estimated using a two-step Empirical Bayes procedure (Casella, 1985; Efron, 2014) using the software package rstan (Stan Development Team, 2022). The rstan package is an interface to the *Stan* probabilistic programming language (Carpenter et al., 2017). The first step of the process used for calibrating the DLM model consists of fitting a number of bootstrapped models with an

optimization algorithm to estimate the standard error of each parameter. The second step then used the standard errors from step 1 as prior distributions in a fully Bayesian estimation of the model using a Markov Chain Monte Carlo procedure. Each step is described in detail in the following sections.

### 5.3.2.1. Step 1: Estimation of Bootstrapped Models

In the first step, the data for each attribute were bootstrap resampled 100 times (Babu, 2011). For each bootstrap resample, the LCDM was fit using the low-memory Broyden-Fletcher-Goldfarb-Shanno optimization algorithm (Liu & Nocedal, 1989; Nocedal & Wright, 2006). The low-memory Broyden-Fletcher-Goldfarb-Shanno algorithm is a widely used maximum likelihood optimization algorithm that can efficiently estimate many types of models, including the LCDM. After estimating the LCDM on each of the bootstrapped samples, there are 100 estimates of each of the three model parameters. We denote these parameters as:

$$\begin{aligned}\lambda_0^* &= [\lambda_{0_1}, \lambda_{0_2}, \lambda_{0_3}, \dots, \lambda_{0_{100}}] \\ \lambda_1^* &= [\lambda_{1_1}, \lambda_{1_2}, \lambda_{1_3}, \dots, \lambda_{1_{100}}] \\ \nu^* &= [\nu_1, \nu_2, \nu_3, \dots, \nu_{100}]\end{aligned}$$

For each parameter, we calculated the mean value and standard deviation across the 100 bootstrap samples. These values were then used to define the prior distributions in the second step.

### 5.3.2.2. Step 2: Estimation of Final Bayesian Model

In the second step, the full data set was used to estimate the LCDM for each linkage level using Markov Chain Monte Carlo and the Hamiltonian Monte Carlo algorithm (Betancourt, 2018; Neal, 2011). The prior distribution for each parameter is defined using the values from the first step. The prior for the intercept ( $\lambda_0$ ) is defined as a normal distribution with a mean and standard deviation equal to the corresponding values from the first step. We define the mean and standard deviation of  $\lambda_0^*$  as  $\mu_{\lambda_0^*}$  and  $\sigma_{\lambda_0^*}$ , respectively. The prior for the intercept in the LCDM is then given as:

$$\lambda_0 \sim \mathcal{N}(\mu_{\lambda_0^*}, \sigma_{\lambda_0^*}) \quad (5.4)$$

Similarly, the main effect parameter ( $\lambda_1$ ) is also defined with a normal distribution. However, the prior distribution of the main effect is truncated at 0, forcing the main effect to be positive to ensure monotonicity in the LCDM.

$$\lambda_1 \sim \begin{cases} 0, & \text{if } \lambda_1 \leq 0 \\ \mathcal{N}(\mu_{\lambda_1^*}, \sigma_{\lambda_1^*}), & \text{otherwise} \end{cases} \quad (5.5)$$

Finally, because the base rate of linkage level class membership ( $\nu$ ) is a probability that must be between 0 and 1, a beta distribution is used for the prior. The beta distribution is governed by two shape parameters,  $\alpha$  and  $\beta$ . Given these two shape parameters, the mean of the beta distribution is given by:



$$\mu = \frac{\alpha}{\alpha + \beta} \quad (5.6)$$

and the variance as:

$$\sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \quad (5.7)$$

With some algebra, we calculate the values of the shape parameters given a mean ( $\mu_{\nu^*}$ ) and standard deviation ( $\sigma_{\nu^*}$ ):

$$\alpha_{\nu^*} = \left( \frac{1 - \mu_{\nu^*}}{\sigma_{\nu^*}^2} - \frac{1}{\mu_{\nu^*}} \right) \mu_{\nu^*}^2 \quad (5.8)$$

$$\beta_{\nu^*} = \alpha_{\nu^*} \left( \frac{1}{\mu_{\nu^*}} - 1 \right) \quad (5.9)$$

The prior distribution for  $\nu$  is then defined as:

$$\nu \sim \mathcal{B}(\alpha_{\nu^*}, \beta_{\nu^*}) \quad (5.10)$$

After the prior distributions have been defined, the LCDM was estimated. To ensure the posterior was adequately explored, four chains were estimated. For each chain, we specified 2,000 warm-up iterations and retained 1,000 post-warm-up iterations. This resulted in a posterior distribution of 4,000 draws (i.e., 1,000 from each of the four chains). After estimation, we ensured that the model had converged and adequately explored the posterior space by evaluating the  $\hat{R}$  and effective sample size metrics described by Vehtari et al. (2021). Using the cutoffs recommended by Vehtari et al. (2021), we ensured that all  $\hat{R}$  values were below 1.01 and that all effective sample sizes were greater than 400. After model evaluation, the mean of the posterior distribution for each of the three model parameters was taken. These parameter estimates were then used for scoring the DLM assessments.

### 5.3.3. Estimation of Student Mastery Probabilities

Once the LCDM parameters have been calibrated, student mastery probabilities are then obtained for each assessed linkage level. For DLM scoring, student mastery probabilities are *expected a posteriori*, or EAP, estimates. This is also the method most commonly used in scale score assessments (e.g., item response theory).<sup>6</sup> For each student  $j$  and linkage level  $l$ , EAP estimates of mastery probability,  $\hat{\alpha}_{jl}$ , are obtained using the following formula:

$$\hat{\alpha}_{jl} = \frac{\prod_{i=1}^{I_j} [\pi_{i1}^{X_{ji}} (1 - \pi_{i1})^{(1-X_{ji})}] \nu_1}{\sum_{c=0}^1 \prod_{i=1}^{I_j} [\pi_{ic}^{X_{ji}} (1 - \pi_{ic})^{(1-X_{ji})}] \nu_c} \quad (5.11)$$

<sup>6</sup> For a thorough discussion of the EAP estimates in scale score and diagnostic settings, see Chapter 10 of Rupp et al. (2010).

In Equation 5.11,  $X_{ji}$  is the dichotomous response of student  $j$  to item  $i$  and  $\pi_{ic}$  is the model-based probability of answering item  $i$  correctly, conditional on student  $j$  having mastery status  $c$  for the linkage level, as defined in Equation 5.3. The mastery status can take two values: masters ( $c = 1$ ) and nonmasters ( $c = 0$ ). Finally,  $\nu_c$  is the base rate probability of membership in each mastery status (see Equation 5.1). Thus, the numerator represents the likelihood of the student being in class  $c = 1$  (i.e., the master class), and the denominator is the total likelihood across both classes. The EAP estimate is then the proportion of the total likelihood that comes from the master class.

## 5.4. Model Evaluation

There are many ways to evaluate DCMs. Ravand and Baghaei (2020) suggest four main areas for evaluation: (1) fit, (2) classification consistency and accuracy, (3) item discrimination, and (4) congruence of attribute difficulty with substantive expectations. Fit can be further broken down into different types of fit (e.g., model fit and item fit).

Many of these aspects are described in other sections of this manual and published research. Item fit is described with other measures of item quality in Chapter 3 of this manual, and classification consistency is discussed with other measures of consistency and reliability in Chapter 8 of this manual. The congruence of difficulty and expectations is discussed in the work of Thompson and Nash (2019) and Thompson and Nash (2022). Finally, item discrimination is described later in this chapter in section 5.5.3, in the context of estimated model parameters.

In this section, we focus on two aspects that are critical to inferences of student mastery: model fit and classification accuracy. 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. Also called absolute model fit (e.g., Chen et al., 2013), this aspect involves an evaluation of the alignment between the three parameters estimated for each linkage level and the observed item responses. The second aspect is classification accuracy. This refers to how well the classifications represent the true underlying latent class. The accuracy of the assessment results (i.e., the classifications) is a prerequisite for any inferences that would be made from the results. Thus, the accuracy of the classifications is perhaps the most crucial aspect of model evaluation from a practical and operational standpoint. These aspects are discussed in the following sections.

### 5.4.1. Model Fit

Absolute model fit is evaluated through posterior predictive model checks, as described by Thompson (2019). Using parameter posterior distributions, we create a distribution for the expected number of students at each raw score point (i.e., the number of correct item responses). We then compare the observed number of students at each score point to the expected distribution using a  $\chi^2$ -like statistic. Finally, we can compare our  $\chi^2$ -like statistic to a distribution of what would be expected, given the expected distributions of students at each score point. This results in a posterior predictive  $p$ -value ( $ppp$ ), which represents how extreme our observed statistic is compared to the model-implied expectation. Very low values indicate poor model fit, whereas very high values may indicate overfitting. For details on the calculation of this statistic, see Thompson (2019).

Due to the large number of models being evaluated (i.e., 102 linkage levels), the *ppp* values were adjusted using the Holm correction, which is uniformly more powerful than the popular Bonferroni method (Holm, 1979). Linkage levels were flagged for misfit if the adjusted *ppp* value was less than .05. Table 5.2 shows the percentage of models with acceptable model fit by linkage level. Across all linkage levels, 56 (55%) of the estimated models showed acceptable model fit. Misfit was not evenly distributed across the linkage levels. The lower linkage levels were flagged at a higher rate than the higher linkage levels. This is likely due to the greater diversity in the student population at the lower linkage levels (e.g., required supports, expressive communication behaviors, etc.), which may affect item response behavior. To address the model misfit flags, we are prioritizing test development for linkage levels flagged for misfit so that testlets contributing to misfit can be retired.<sup>7</sup> We also plan to incorporate additional item quality statistics to the review of field test data to ensure that only items and testlets that conform to the model expectations are promoted to the operational assessment. Overall, however, the fungible LCDM models appear to largely reflect the observed data. Additionally, model fit is evaluated on an annual basis and continues to improve over time as a result of adjustments to the pool of available content (i.e., improved item writing practices, retirement of testlets contributing to misfit). Finally, it should be noted that a linkage level flagged for model misfit may still have high classification accuracy, indicating that student mastery classifications can be trusted, even in the presence of misfit.

**Table 5.2**

*Percentage of Models With Acceptable Model Fit (ppp > .05)*

Linkage Level	%
Initial	41.2
Precursor	32.4
Target	91.2

### 5.4.2. Classification Accuracy

The most practically important aspect of model fit for DCMs is classification accuracy. Classification accuracy is a measure of how accurate or uncertain classification decisions are for a given attribute in a DCM (the linkage level for DLM assessments). This measure of model fit is conceptualized by a summary of a  $2 \times 2$  contingency table of the true and model-estimated mastery statuses (Sinharay & Johnson, 2019).<sup>8</sup>

For an operational assessment, we do not know students' true mastery status. However, we can still estimate the classification accuracy for each linkage level, as shown by Wang et al. (2015) and Johnson and Sinharay (2018) with

$$\hat{P}_A = \frac{1}{N} \sum_{n=1}^N \tilde{\alpha} P(\alpha = 1 | \mathbf{x} = \mathbf{x}_n) + \frac{1}{N} \sum_{n=1}^N (1 - \tilde{\alpha}) P(\alpha = 0 | \mathbf{x} = \mathbf{x}_n). \quad (5.12)$$

<sup>7</sup> For a description of item development practices, see Chapter 3 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

<sup>8</sup> For a discussion of the closely related classification consistency, see Chapter 8 of this manual.

In Equation 5.12,  $N$  is the total number of students,  $\tilde{\alpha}$  is the model-estimated mastery status, and  $P(\alpha = 1 | \mathbf{X} = \mathbf{x}_n)$  is the model-estimated probability that the linkage level was mastered (or not mastered for  $\alpha = 0$ ). Johnson and Sinharay (2018) recommended interpretive guidelines for the classification accuracy,  $\hat{P}_A$ :  $\geq .99$  = Excellent,  $.95$ – $.98$  = Very Good,  $.89$ – $.94$  = Good,  $.83$ – $.88$  = Fair,  $.55$ – $.82$  = Poor, and  $< .55$  = Weak.

Across all estimated models, 61 linkage levels (60%) demonstrated at least fair classification accuracy. Table 5.3 shows the number and percentage of models within each linkage level that demonstrated each category of classification accuracy. Results are fairly consistent across linkage levels, with no one level showing systematically higher or lower accuracy. As was the case for model misfit, linkage levels flagged for low classification accuracy are prioritized for test development.

**Table 5.3**

*Estimated Classification Accuracy by Linkage Level*

Linkage Level	Weak	Poor	Fair	Good	Very Good	Excellent
	(%)	(%)	(%)	(%)	(%)	(%)
	0.00–.55	.55–.82	.83–.88	.89–.94	.95–.98	.99–1.00
Initial	0 (0.0)	3 (8.8)	2 (5.9)	22 (64.7)	7 (20.6)	0 (0.0)
Precursor	0 (0.0)	18 (52.9)	13 (38.2)	2 (5.9)	1 (2.9)	0 (0.0)
Target	0 (0.0)	20 (58.8)	10 (29.4)	1 (2.9)	2 (5.9)	1 (2.9)

When looking at absolute model fit and classification accuracy in combination, linkage levels flagged for absolute model misfit often have high classification accuracy. Of the 46 linkage levels that were flagged for absolute model misfit, 33 (72%) showed fair or better classification accuracy. Thus, even when misfit is present, we can be confident in the accuracy of the mastery classifications. In total, 87% of linkage levels ( $n = 89$ ) had acceptable absolute model fit and/or acceptable classification accuracy.

## 5.5. Calibrated Parameters

As stated in the previous section, the item parameters for diagnostic assessments are the conditional probability of nonmasters providing a correct response (i.e., the inverse logit of  $\lambda_0$ ) and the conditional probability of masters providing a correct response (i.e., the inverse logit of  $\lambda_0 + \lambda_1$ ). Because of the assumption of fungibility, parameters are calculated for each of the 102 linkage levels in science. Across all linkage levels, the conditional probability that masters provide a correct response is generally expected to be high, while it is expected to be low for nonmasters. In addition to the item parameters, the psychometric model also includes a structural parameter, which defines the base rate of class membership for each linkage level. A summary of the operational parameters used to score the 2021–2022 assessment is provided in the following sections.

### 5.5.1. 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.

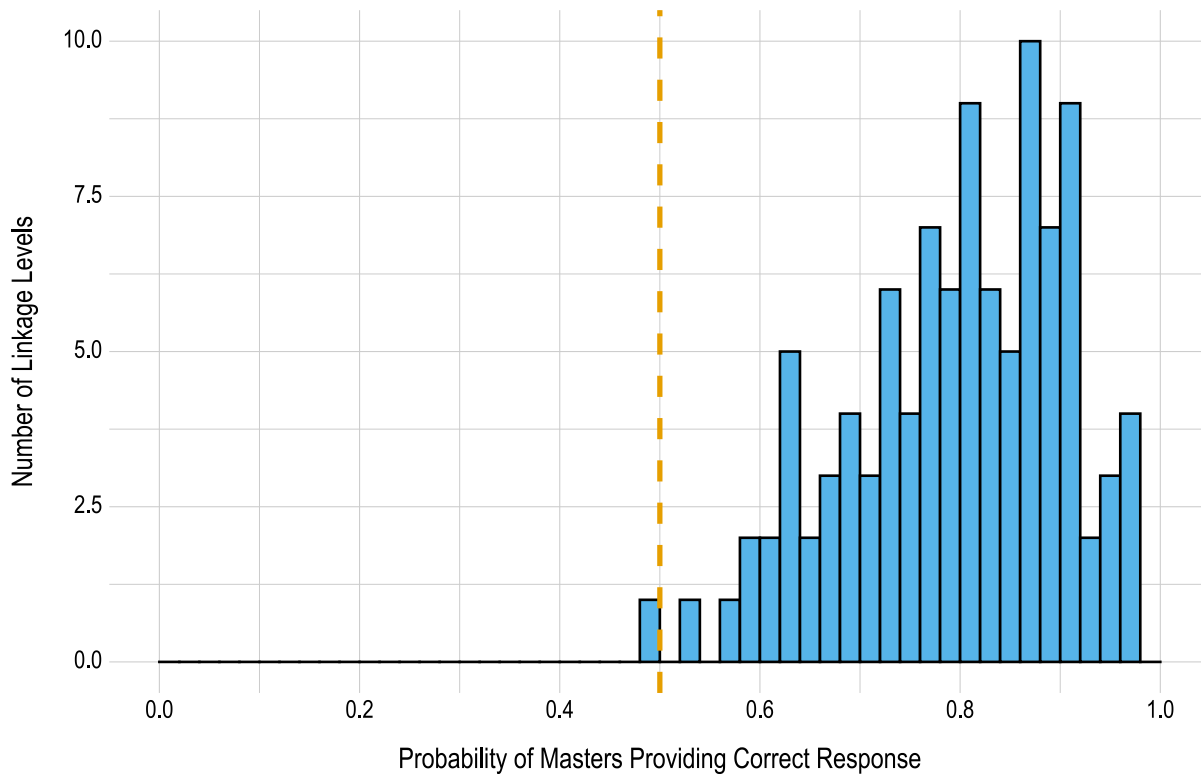
Instances where masters have a low probability of providing correct responses may indicate that the linkage level does not measure what it is intended to measure, or that students who have mastered the content select a response other than the key. These instances may result in students who have mastered the content providing incorrect responses and being incorrectly classified as nonmasters. This outcome has implications for the validity of inferences that can be made from results, including educators using results to inform instructional planning in the subsequent year.

Using the 2021–2022 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 .50 (i.e., equal likelihood of mastery or nonmastery), masters should have a greater than 50% chance of providing a correct response. The results in Figure 5.1 demonstrate that the vast majority of linkage levels ( $n = 101$ , 99%) performed as expected. Additionally, 95% of linkage levels ( $n = 97$ ) had a conditional probability of masters providing a correct response over .60. No linkage levels ( $n = 0$ , <1%) had a conditional probability of masters providing a correct response less than .40.

Thus, the vast majority of linkage levels performed consistently with expectations for masters of the linkage levels.

**Figure 5.1**

*Probability of Masters Providing a Correct Response to Items Measuring Each Linkage Level*



*Note.* Histogram bins are shown in increments of .02. Reference line indicates .50.

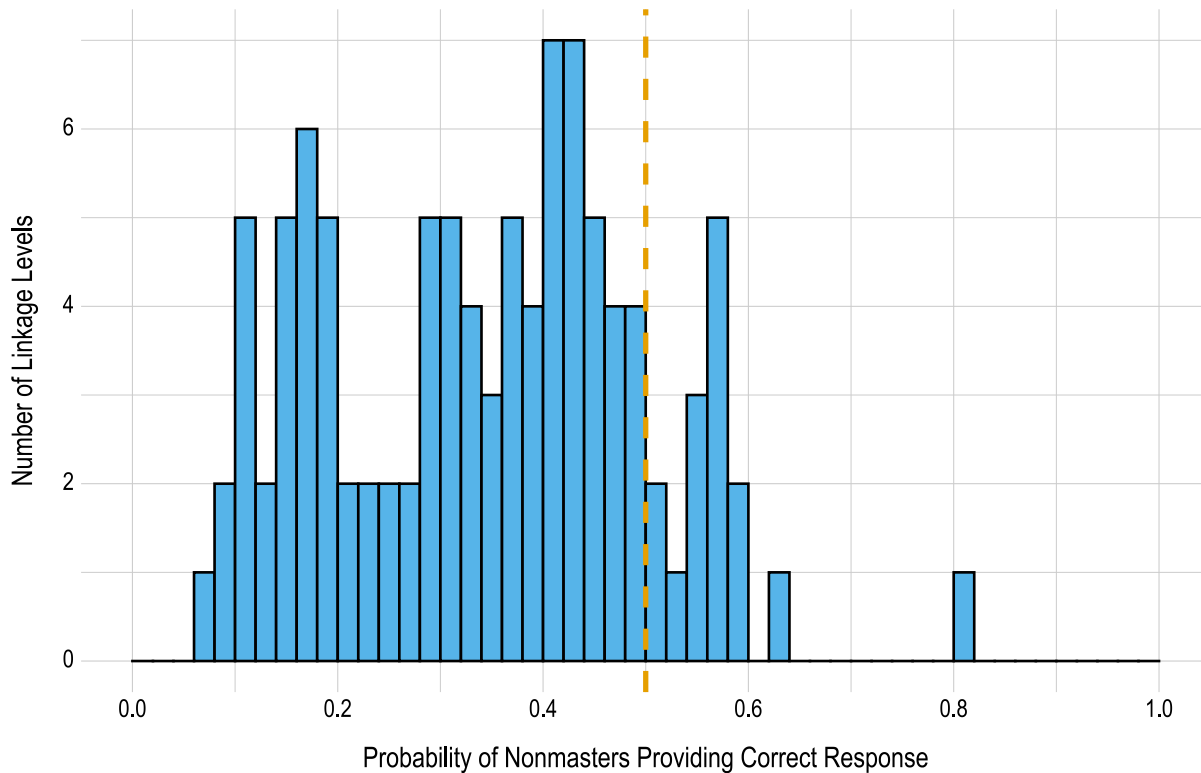
### **5.5.2. Probability of Nonmasters Providing Correct Response**

When items measuring each linkage level function as expected, nonmasters of the linkage level have a low probability of providing a correct response to items measuring the linkage level. Instances where nonmasters 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 educators using results to inform instructional planning in the subsequent year.

Figure 5.2 summarizes the probability of nonmasters providing correct responses to items measuring each of the 102 linkage levels. There is greater variation in the probability of nonmasters providing a correct response to items measuring each linkage level than was observed for masters, as shown in Figure 5.2. While the majority of linkage levels ( $n = 87$ , 85%) performed as expected, nonmasters sometimes had a greater than .50 chance of providing a correct response to items measuring the linkage level. Although most linkage levels ( $n = 60$ , 59%) have a conditional probability of nonmasters providing a correct response less than .40, 2 (2%) have a conditional probability for nonmasters providing a correct response greater than .60, indicating there are some linkage levels where nonmasters 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. All of the two linkage levels with a conditional probability for nonmasters providing a correct response greater than .60 were at the Target linkage level.

**Figure 5.2**

*Probability of Nonmasters Providing a Correct Response to Items Measuring Each Linkage Level*



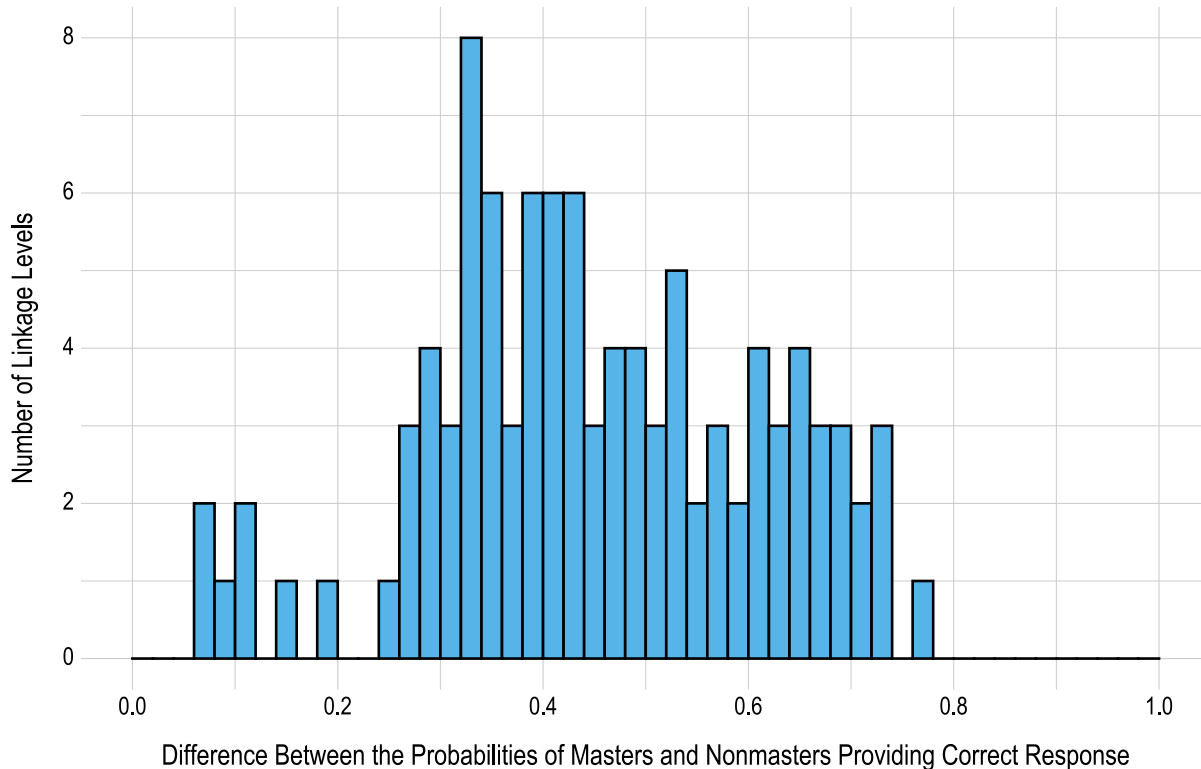
Note. Histogram bins are in increments of .02. Reference line indicates .50.

### 5.5.3. Item Discrimination

The discrimination of a linkage level represents how well the items are able to differentiate masters and nonmasters. For diagnostic models, this is assessed by comparing the conditional probabilities of masters and nonmasters providing a correct response. Linkage levels that are highly discriminating will have a large difference between the conditional probabilities, with a maximum value of 1.00 (i.e., masters have a 100% chance of providing a correct response and nonmasters a 0% chance). Figure 5.3 shows the distribution of linkage level discrimination values. Overall, 60% of linkage levels ( $n = 61$ ) have a discrimination greater than .40, indicating a large difference between the conditional probabilities (e.g., .75 to .35, .90 to .50, etc.). However, there were 3 linkage levels (3%) with a discrimination of less than .10, indicating that masters and nonmasters tend to perform similarly on items measuring these linkage levels. All of the three linkage levels with a discrimination of less than .10 were at the Target linkage level.

**Figure 5.3**

*Difference Between Masters' and Nonmasters' Probability of Providing a Correct Response to Items Measuring Each Linkage Level*



Note. Histogram bins are in increments of .02.

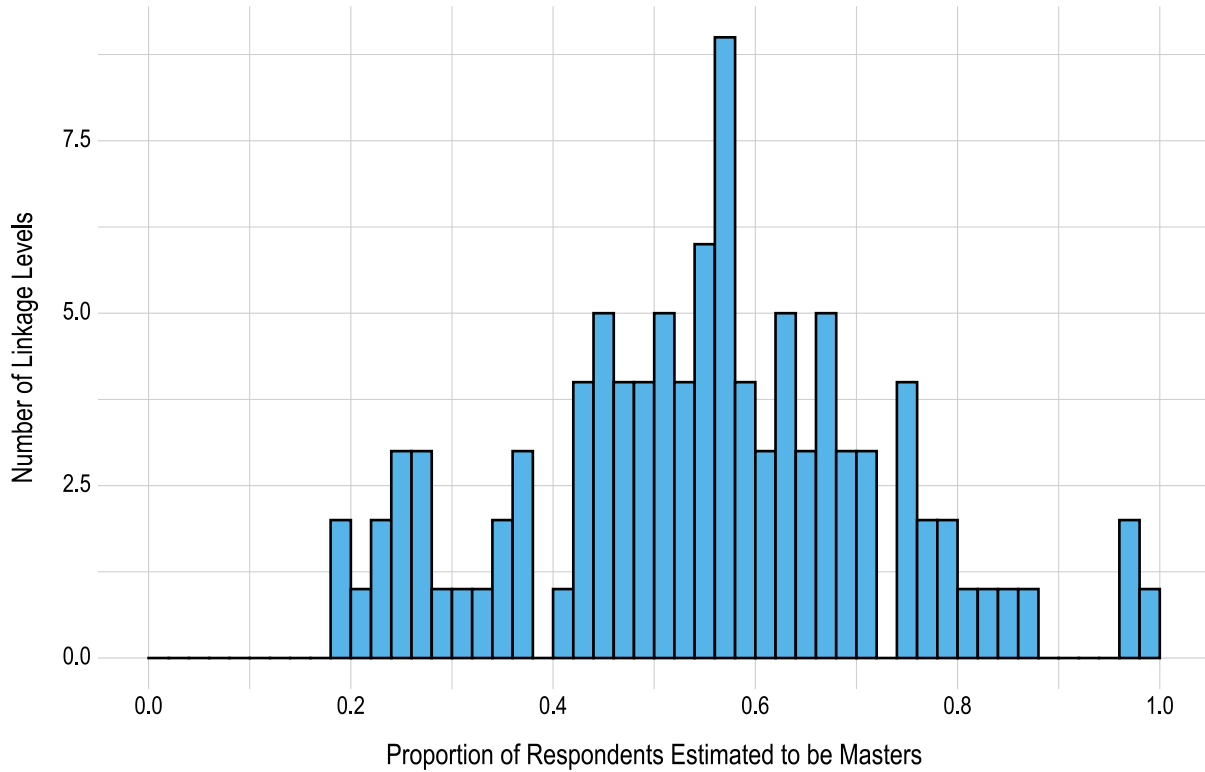
### **5.5.4. Base Rate Probability of Class Membership**

The base rate of class membership is the DCM structural parameter and represents the estimated proportion of students in each class for each EE and linkage level. A base rate close to .50 indicates that students assessed on a given linkage level are, *a priori*, equally likely to be a master or nonmaster. Conversely, a high base rate would indicate that students testing on a linkage level are, *a priori*, more likely to be masters. Figure 5.4 depicts the distribution of the base rate probabilities. Overall, 80% of linkage levels ( $n = 82$ ) had a base rate of mastery between .25 and .75. On the edges of the distribution, 6 linkage levels (6%) had a base rate of mastery less than .25, and 14 linkage levels (14%) had a base rate of mastery higher than .75. This indicates that students are more likely to 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 .02.

## 5.6. 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. A DCM is estimated 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 nonmasters, owing to the conceptual approach used to construct DLM testlets. An analysis of the estimated models indicates that the estimated models have acceptable levels of absolute model fit and classification accuracy. Additionally, the estimated parameters from each DCM are generally within the optimal ranges.

## 6. Standard Setting

The initial standard-setting process for the Dynamic Learning Maps® (DLM®) Alternate System in science derived cut points for then-tested grades 4, 5, 6, 8, and high school, to describe student achievement relative to four performance levels. For a description of the process, including the development of policy performance level descriptors, the three-day standard-setting meeting, evaluation of impact data and cut points, and development of grade specific performance level descriptors, see Chapter 6 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

After a new state that assessed students in science in all grades began administering DLM science assessments, additional cut points were derived for grades 3 and 7. For a description of the procedures and results of establishing grade 3 and 7 cut points, see Chapter 6 of the *2018–2019 Technical Manual Update—Science* (DLM Consortium, 2019).

## 7. Reporting and Results

Chapter 7 of the Dynamic Learning Maps® (DLM®) Alternate Assessment System *2015–2016 Technical Manual—Science* (DLM Consortium, 2017) 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 education agencies. Technical Manual updates provide a description of data files, score reports, and results for each corresponding academic year.

This chapter presents spring 2022 student participation data; the percentage of students achieving at each performance level; and subgroup performance by gender, race, ethnicity, and English learner status. This chapter also reports the distribution of students by the highest linkage level mastered during spring 2022. Finally, this chapter describes updates made to score reports during the 2021–2022 operational year. For a complete description of score reports and interpretive guides, see Chapter 7 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

### 7.1. Student Participation

During spring 2022, assessments were administered to 44,375 students in 20 states. Counts of students tested in each state are displayed in Table 7.1. The assessments were administered by 18,451 educators in 11,006 schools and 4,164 school districts. A total of 388,092 test sessions were administered during the spring assessment window. One test session is one testlet taken by one student. Only test sessions that were complete at the close of the spring assessment window counted toward the total sessions.

**Table 7.1**

*Student Participation by State (N = 44,375)*

State	Students (n)
Alaska	194
Arkansas	2,379
Delaware	334
District of Columbia	177
Illinois	4,612
Iowa	945
Kansas	911
Maryland	2,444
Missouri	2,298
New Hampshire	254
New Jersey	4,612
New Mexico	892
New York	7,779
North Dakota	233
Oklahoma	2,180
Pennsylvania	7,010
Rhode Island	411
Utah	3,558
West Virginia	677
Wisconsin	2,475

Table 7.2 summarizes the number of students assessed in each grade and course. More than 14,010 students participated in each of the elementary and the middle school grade bands.<sup>9</sup> In high school, over 15,300 students participated. The differences in high school grade-level participation can be traced to differing state-level policies about the grade(s) in which students are assessed.

<sup>9</sup> 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 7.2**

*Student Participation by Grade or Course (N = 44,375)*

Grade	Students ( <i>n</i> )
3	577
4	6,333
5	7,109
6	1,010
7	1,000
8	13,027
9	4,068
10	1,761
11	8,060
12	372
Biology	1,058

Table 7.3 summarizes the demographic characteristics of the students who participated in the spring 2022 administration. The distribution of students across the different subgroups was fairly consistent with prior years' distributions. The majority of participants were male (67%) and white (60%). About 7% of students were monitored or eligible for English learning services.

**Table 7.3**

*Demographic Characteristics of Participants (N = 44,375)*

Subgroup	<i>n</i>	%
<b>Gender</b>		
Male	29,762	67.1
Female	14,574	32.8
Nonbinary/undesigned	39	0.1
<b>Race</b>		
White	26,415	59.5
African American	9,210	20.8
Two or more races	5,143	11.6
Asian	2,170	4.9
American Indian	1,071	2.4
Native Hawaiian or Pacific Islander	283	0.6
Alaska Native	83	0.2
<b>Hispanic ethnicity</b>		
Non-Hispanic	34,895	78.6
Hispanic	9,480	21.4
<b>English learning (EL) participation</b>		
Not EL eligible or monitored	41,360	93.2
EL eligible or monitored	3,015	6.8

In addition to the spring assessment window, instructionally embedded science assessments are also made available for educators to optionally administer to students during the year. Results from the instructionally embedded 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 who completed at least one instructionally embedded assessment by state. A total of 4,139 students in 12 states took at least one instructionally embedded testlet during the 2021–2022 academic year.

**Table 7.4**

*Students Completing Instructionally Embedded Science Testlets by State (N = 4,139)*

State	<i>n</i>
Arkansas	660
Delaware	67
Iowa	319
Kansas	238
Maryland	1,191
Missouri	1,481
New Jersey	33
New Mexico	7
New York	3
North Dakota	46
Oklahoma	89
Utah	5

*Note.* Maryland required administration of instructionally embedded assessments during fall 2021.

Table 7.5 summarizes the number of instructionally embedded testlets taken in science. Across all states, students took 28,007 science testlets during the instructionally embedded window.

**Table 7.5**

*Number of Instructionally Embedded Science Testlets by Grade (N = 28,007)*

Grade	<i>n</i>
3	1,006
4	1,245
5	4,135
6	1,645
7	1,565
8	6,563
9	1,393
10	1,989
11	6,225
12	2,241
<i>Total</i>	<i>28,007</i>

## 7.2. Student Performance

Student performance on DLM assessments is interpreted using cut points,<sup>10</sup> which describe student achievement using four performance levels. A student’s performance level is determined based on the total number of linkage levels mastered across the assessed EEs.

For the spring 2022 administration, student performance was reported using the same four performance levels approved by the DLM Governance Board 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*. This performance level is considered to be meeting achievement expectations.
- 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 2022 administration for science. At the elementary level, the percentage of students who achieved at the At Target or Advanced levels (i.e., proficient) was approximately 12%; in middle school the percentage of students meeting or exceeding At Target expectations was approximately 22%; in high school the percentage was approximately 15%; in end-of-instruction biology the percentage was approximately 14%.

**Table 7.6**

*Percentage of Students by Grade and Performance Level*

Grade	Emerging (%)	Approaching (%)	At Target (%)	Advanced (%)	At Target + Advanced (%)
3 ( <i>n</i> = 577)	67.1	21.7	7.5	3.8	11.3
4 ( <i>n</i> = 6,333)	64.8	20.2	11.5	3.5	15.0
5 ( <i>n</i> = 7,109)	69.5	20.9	9.1	0.5	9.6
6 ( <i>n</i> = 1,010)	67.8	16.1	12.1	4.0	16.0
7 ( <i>n</i> = 1,000)	61.0	19.3	15.9	3.8	19.7
8 ( <i>n</i> = 13,027)	57.5	20.0	18.9	3.6	22.4
9 ( <i>n</i> = 4,068)	54.3	27.4	14.3	4.1	18.4
10 ( <i>n</i> = 1,761)	59.2	28.2	10.8	1.8	12.6
11 ( <i>n</i> = 8,060)	56.0	29.4	12.2	2.4	14.6
12 ( <i>n</i> = 372)	58.6	25.8	12.6	3.0	15.6
Biology ( <i>n</i> = 1,058)	68.0	18.1	10.9	3.0	13.9

<sup>10</sup> For a description of the standard setting process used to determine the cut points, see Chapter 6 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).



### **7.2.2. Subgroup Performance**

Data collection for DLM assessments includes demographic data on gender, race, ethnicity, and English learning status. Table 7.7 summarizes the disaggregated frequency distributions for science, collapsed across all assessed grade levels. Although state education agencies 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**

*Science Performance Level Distributions by Demographic Subgroup (N = 44,375)*

Subgroup	Emerging		Approaching		At Target		Advanced		At Target + Advanced	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
<b>Gender</b>										
Male	17,862	60.0	6,704	22.5	4,260	14.3	936	3.1	5,196	17.5
Female	9,036	62.0	3,407	23.4	1,810	12.4	321	2.2	2,131	14.6
Nonbinary/undesigned	22	56.4	10	25.6	6	15.4	1	2.6	7	17.9
<b>Race</b>										
White	15,778	59.7	6,137	23.2	3,750	14.2	750	2.8	4,500	17.0
African American	5,613	60.9	2,042	22.2	1,272	13.8	283	3.1	1,555	16.9
Two or more races	3,223	62.7	1,205	23.4	619	12.0	96	1.9	715	13.9
Asian	1,519	70.0	397	18.3	198	9.1	56	2.6	254	11.7
American Indian	552	51.5	268	25.0	187	17.5	64	6.0	251	23.4
Native Hawaiian or Pacific Islander	170	60.1	60	21.2	44	15.5	9	3.2	53	18.7
Alaska Native	65	78.3	12	14.5	6	7.2	0	0.0	6	7.2
<b>Hispanic ethnicity</b>										
Non-Hispanic	21,161	60.6	7,955	22.8	4,799	13.8	980	2.8	5,779	16.6
Hispanic	5,759	60.7	2,166	22.8	1,277	13.5	278	2.9	1,555	16.4
<b>English learning (EL) participation</b>										
Not EL eligible or monitored	25,142	60.8	9,424	22.8	5,654	13.7	1,140	2.8	6,794	16.4
EL eligible or monitored	1,778	59.0	697	23.1	422	14.0	118	3.9	540	17.9

## 7.3. Mastery Results

As described above, the student performance levels are determined by applying cut points to the total number of linkage levels mastered. In this section, we summarize student mastery of assessed EEs and linkage levels, including how students demonstrated mastery from among three scoring rules and the highest linkage level students tended to master.

### 7.3.1. Mastery Status Assignment

As described in Chapter 5 of this manual, student responses to assessment items are used to estimate the posterior probability that the student mastered each of the assessed linkage levels using diagnostic classification modeling. Students with a posterior probability of mastery greater than or equal to .80 are assigned a linkage level mastery status of 1, or mastered. Students with a posterior probability of mastery less than .80 are assigned a linkage level mastery status of 0, or not mastered. Maximum uncertainty in the mastery status occurs when the probability is .5 and maximum certainty when the probability approaches 0 or 1. After considering the risk of false positives and negatives and preliminary data analyses, and based on input from the DLM Technical Advisory Committee (TAC), the threshold used to determine mastery classifications was set at .80. In addition to the calculated 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 did not demonstrate mastery of the assessed linkage level, mastery was assigned at two linkage levels below the level that was assessed.

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 .80 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), rather than showing no evidence of mastery at all. Theoretical evidence for the use of the two-down rule based on DLM content structures is presented in Chapter 2 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

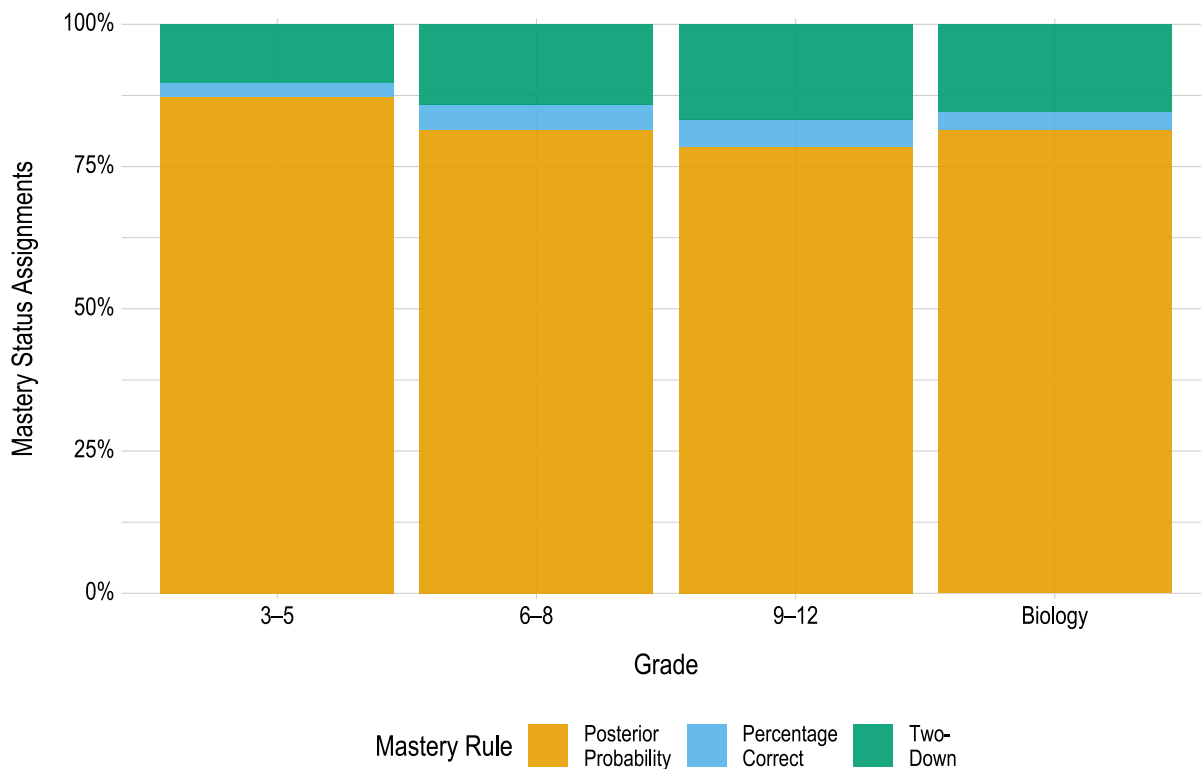
To evaluate the degree to which each mastery assignment rule contributed to students' linkage level mastery status during the 2021–2022 administration of DLM assessments, the percentage of mastery statuses obtained by each scoring rule was calculated, as shown in Figure 7.1. Posterior probability was 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. This means that EEs that were assessed by a student at the lowest two linkage levels (i.e., Initial and Precursor) are never categorized as having mastery assigned by the two-down rule. This is because the student would either master the assessed linkage level and have the EE counted under the posterior probability or 80% correct scoring rule, or all three scoring rules would agree on the score (i.e., no evidence of mastery), in which case preference is given to the posterior probability. Across grades, approximately 78%–87% of mastered linkage levels were derived from the posterior probability

obtained from the modeling procedure. Approximately 2%–5% of linkage levels were assigned mastery status by the percentage correct rule. The remaining approximately 10%–17% 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 .80 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 2021–2022 operational year, the rate of agreement between the two methods was 84%. However, in instances in which the two methods disagreed, the posterior probability method indicated a higher level of mastery (and therefore was implemented for scoring) in 75% 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 7.1**

*Linkage Level Mastery Assignment by Mastery Rule for Each Grade Band or Course*



### 7.3.2. Linkage Level Mastery

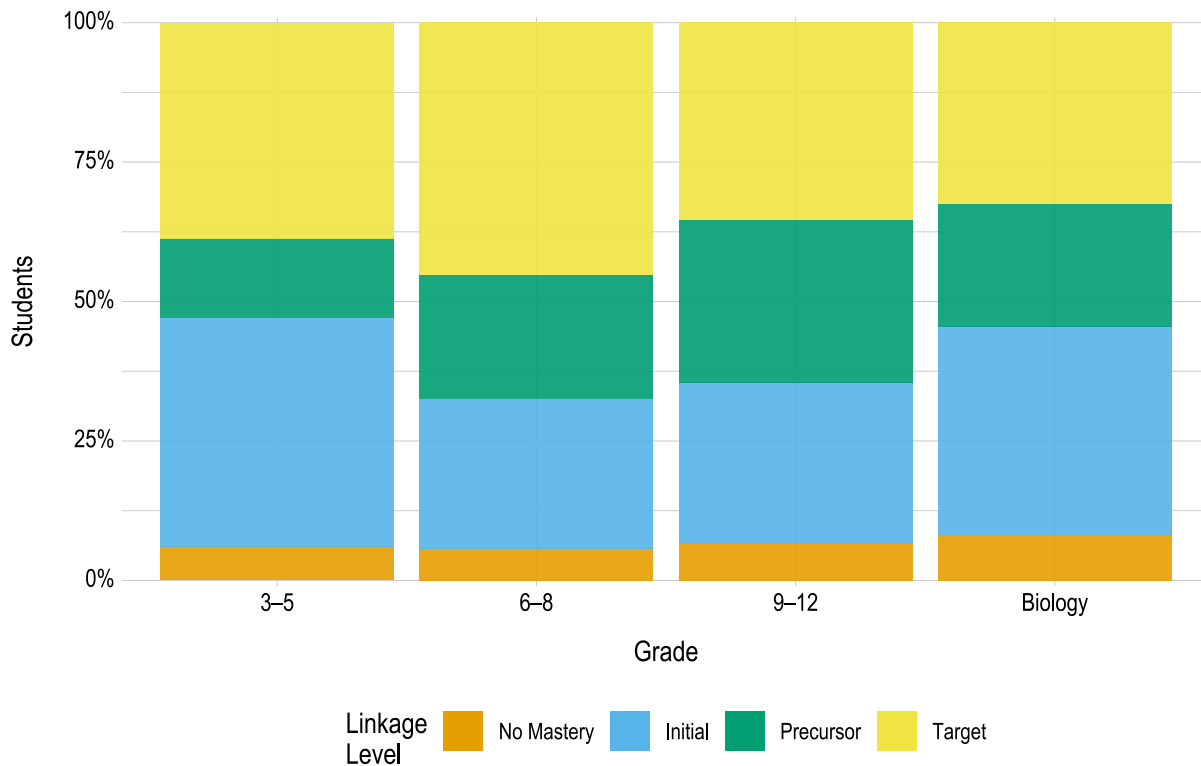
Scoring for DLM assessments determines the highest linkage level mastered for each EE. The linkage levels are (in order): Initial, Precursor, and Target. A student can be a master of zero, one, two, or all three

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 linkage level mastered.

Figure 7.2 displays the percentage of students who mastered each linkage level as the highest linkage level across all assessed EEs in science. For example, across all 3-5 science EEs, the Initial level was the highest level that students mastered 41% of the time. The percentage of students who mastered as high as the Target linkage level ranged from approximately 33% to 45%.

**Figure 7.2**

*Students' Highest Linkage Level Mastered Across Science Essential Elements by Grade*



## 7.4. Data Files

DLM assessment results were made available to DLM state education agencies following the spring 2022 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 science for all students who completed any testlets. In addition to the GRF, the states received 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 education agencies 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 education agencies are provided with an incident file describing students impacted; however, no incidents occurred during 2021–2022.

Consistent with prior delivery cycles, state education agencies were provided with a 2-week 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 education agencies submitted final GRFs via Educator Portal. The final GRF was used to generate score reports.

## **7.5. Score Reports**

Assessment results were provided to state education agencies to report to parents/guardians, educators, and local education agencies. Individual Student Score Reports summarized student performance on the assessment. 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 2022. One change to the Individual Student Score Reports is summarized below. For a complete description of score reports, including aggregated reports, see Chapter 7 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

### **7.5.1. Individual Student Score Reports**

Because of continued impacts of the COVID-19 pandemic on instruction and assessment, during 2021–2022, state education agencies were given the option to add a cautionary statement to Individual Student Score Reports, which indicated that the results may reflect the continued effects of the COVID-19 pandemic on student performance. Four states opted to include the cautionary statement on their individual score reports.


A sample Performance Profile and a sample Learning Profile reflecting the optional cautionary statement are provided in Figure 7.3 and Figure 7.4.

**Figure 7.3**

Example Page of the Performance Profile With Cautionary Statement for 2021–2022.

**REPORT DATE:** 08-03-2022  
**SUBJECT:** Science  
**GRADE:** 8

**Individual Student End-of-Year Report  
Performance Profile 2021-2022**



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


**DISTRICT ID:** DLM District  
**STATE:** DLM State  
**STATE ID:** DLM State ID

---

### Overall Results

Results from 2021–2022 may reflect the continued effects of the COVID-19 pandemic on student performance.

Middle school science allows students to show their achievement in 27 skills related to 9 Essential Elements. Student has mastered 22 of those 27 skills during Spring 2022. Overall, Student’s mastery of science fell into the third of four performance categories: **at target**.



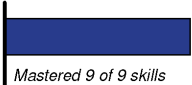
emerging      approaching the target      at target      advanced

<b>EMERGING:</b>	The student demonstrates <b>emerging</b> understanding of and ability to apply content knowledge and skills represented by the Essential Elements.
<b>APPROACHING THE TARGET:</b>	The student’s understanding of and ability to apply targeted content knowledge and skills represented by the Essential Elements is <b>approaching the target</b> .
<b>AT TARGET:</b>	The student’s understanding of and ability to apply content knowledge and skills represented by the Essential Elements is <b>at target</b> .
<b>ADVANCED:</b>	The student demonstrates <b>advanced</b> understanding of and ability to apply targeted content knowledge and skills represented by the Essential Elements.

### Domain

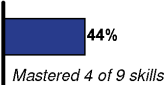
Bar graphs summarize the percent of skills mastered by domain. Not all students test on all skills due to availability of content at different levels per standard.

Physical Science



Mastered 9 of 9 skills

Life Science



Mastered 4 of 9 skills

Page 1 of 2

For more information, including resources, please visit <https://dynamiclearningmaps.org/states>.  
© The University of Kansas. All rights reserved. For educational purposes only. May not be used for commercial or other purposes without permission. "Dynamic Learning Maps" is a trademark of The University of Kansas.


**Figure 7.4**

*Example Page of the Learning Profile With Cautionary Statement for 2021–2022.*

**REPORT DATE:** 08-03-2022  
**SUBJECT:** Science  
**GRADE:** 8

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

**Individual Student End-of-Year Report  
Learning Profile 2021-2022**



**DISTRICT ID:** DLM District  
**STATE:** DLM State  
**STATE ID:** DLM State ID

Results from 2021–2022 may reflect the continued effects of the COVID-19 pandemic on student performance.

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 Spring 2022. Student was assessed on 9 out of 9 Essential Elements and 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.

Essential Element	Estimated Mastery Level		
	1	2	3 (Target)
SCI.EE.MS.PS1-2	Identify change	Gather data on properties before and after chemical changes	Interpret data on properties before and after chemical changes
SCI.EE.MS.PS2-2	Identify ways to change motion	Investigate and identify ways to change motion	Investigate and predict changes in motion
SCI.EE.MS.PS3-3	Identify objects or materials that minimize thermal energy transfer	Investigate objects/materials and predict changes in thermal energy transfer	Refine a device to minimize or maximize thermal energy transfer
SCI.EE.MS.LS1-3	Recognize major organs	Model how organs are connected	Make a claim about how organ structure and function support survival
SCI.EE.MS.LS1-5	Match organisms to habitats	Identify factors that influence growth of organisms	Interpret data to show environmental resources influence growth
SCI.EE.MS.LS2-2	Identify food that animals eat	Classify animals based on what they eat	Identify producers and consumers in a food chain

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

This report is intended to serve as one source of evidence in an instructional planning process. Results are based only on item responses from the full academic year. Because your child may demonstrate knowledge and skills differently across settings, the estimated mastery results shown here may not fully represent what your child knows and can do. For more information, including resources, please visit <https://dynamiclearningmaps.org/states>.  
 © The University of Kansas. All rights reserved. For educational purposes only. May not be used for commercial or other purposes without permission. "Dynamic Learning Maps" is a trademark of The University of Kansas.

Page 1 of 2

## 7.6. Quality-Control Procedures for Data Files and Score Reports

Changes to the quality-control procedures were made only to the extent of accommodating the revised score reports for 2021–2022 (i.e., checking to be sure the cautionary statement was correctly applied for states who opted to include it on score reports). For a complete description of quality-control procedures, see Chapter 7 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

## 7.7. Conclusion

Results for DLM assessments include students’ overall performance levels and linkage level mastery decisions for each assessed EE and linkage level. During spring 2022, assessments were administered to 44,375 students in 20 states. Between 10% and 22% of students achieved at the At Target or Advanced levels across all grades. Of the three scoring rules, linkage level mastery status was most frequently assigned by the diagnostic classification model.

Following the spring 2022 administration, three data files were delivered to state education agencies: GRF,



special circumstance code file, and exited students file. Lastly, state education agencies could opt to include cautionary text to score reports to aid in interpretation.

## 8. Reliability

The Dynamic Learning Maps® (DLM®) Alternate Assessment System reports student results as discrete mastery classifications derived from diagnostic classification models (DCMs) and aggregations of the individual classifications. As such, methods for estimating the reliability must reflect the innovative assessment design and unique reporting structure of the DLM System. The approach draws on previous research from classification-based assessments to describe the reliability of student results. This chapter discusses the methods used to estimate reliability, the factors that are likely to affect variability in reliability results, and an overall summary of reliability evidence.

### 8.1. Background Information on Reliability Methods

Reliability estimates quantify the degree of precision in a test score. Expressed another way, a reliability index specifies how likely scores are to vary from one test administration to another due to chance. Historically, reliability has been quantified using indices such as the Guttman-Cronbach alpha (Cronbach, 1951; Guttman, 1945), which provides an index of the proportion of variance in a test score that is due to variance in the trait. Values closer to 1.0 indicate variation in test scores comes from individual differences in the trait, whereas values closer to 0.0 indicate variation in test scores comes from random error.

Many traditional measures of reliability exist; their differences are due to assumptions each measure makes about the nature of the data from a test. For instance, the Spearman-Brown reliability formula assumes items are parallel, contributing equal amounts of information about the trait and having equal variance. The Guttman-Cronbach alpha assumes tau-equivalent items (i.e., items with equal information about the trait but not necessarily equal variances). As such, the alpha statistic is said to subsume the Spearman-Brown statistic, meaning that if the data meet the stricter definition of Spearman-Brown, then alpha will be equal to Spearman-Brown. As a result, inherent in any discussion of reliability is the fact that the metric of reliability is accurate to the extent that the assumptions of the test are met.

The DLM assessments are scored using diagnostic classification models (DCMs), which assume that students' knowledge, skills, and understandings are represented by discrete mastery statuses, rather than a continuous latent trait that characterizes more traditional classical test theory and item response theory models. As such, reliability-estimation methods based on item response theory estimates of ability are not applicable for the DLM assessments. Therefore, the reliability evidence may appear different from that reported when test scores are produced using traditional psychometric techniques such as classical test theory or item response theory. However, interpretation of indices for DLM assessments is consistent with traditional approaches. When a test is perfectly reliable (i.e., it has an index value of 1), any variation in test scores comes from individual differences in the trait within the sample in which the test was administered. When a test has zero reliability, then any variation in test scores comes solely from random error.

DCMs are models that produce classifications based on probability estimates for students. For the DLM System, the classification estimates are based on the set of attributes across the alternate achievement standards on which each student was assessed. The alternate achievement standards are themselves organized into larger content strands. In DLM terms, each content strand is called a topic, each of which is made up of standards called Essential Elements (EEs). Each EE is available at three linkage levels of complexity, which are the attributes in the DCM<sup>11</sup>: Initial, Precursor, and Target. Topics are organized into

<sup>11</sup> For more information on the specification of the DCMs, see Chapter 5 of this manual.

overarching domains for the subject. For the end-of-instruction Biology assessment, results are reported for topics, while the elementary, middle school, and high school assessment report results for the domain.

DLM testlets are written with items measuring the linkage level. Because of the DLM administration design, students do not take testlets outside of a single linkage level within an EE. Students take a single testlet measuring a given EE and linkage level; consequently, data obtained when students respond to testlets at adjacent linkage levels within an EE are sparse. Therefore, a linkage level DCM is used to score the assessment (i.e., estimate mastery proficiency; see Chapter 5 in this manual for more information).

The DCMs produce student-level posterior probabilities for each linkage level for which a student is assessed, with a threshold of 0.8 specified for demonstrating mastery. To guard against the model being overly influential, two additional scoring rules are applied. Students can also demonstrate mastery by providing correct responses to at least 80% of the items measuring the EE and linkage level.<sup>12</sup> Furthermore, because students are not assessed at more than one linkage level within an EE, students who do not meet mastery status for any assessed linkage level are assigned mastery status for the linkage level two levels below the lowest level on which they are assessed (unless the lowest level tested is either the Initial or Precursor, in which case students are considered nonmasters of all linkage levels within the EE). See Chapter 7 of this manual for a complete description of scoring rules for the DLM assessments.

The DLM score reports display linkage level mastery for each EE. Linkage level results are also aggregated for EEs within each topic in Biology and domain in science grade bands. Because of differences in the organization of the assessment blueprints, score reports in Biology summarize student results for EEs measuring each topic, while those in general science summarize student results for EEs measuring each domain. Reliability evidence for each type of science assessment is provided consistent with the level used to summarize performance in the student reports. Score reports also summarize overall performance with a performance level classification. The classification is determined by summing all linkage levels mastered and comparing the value with cut points determined during standard setting. For more information on cut points, see Chapter 6 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017). For more information on score reports, see Chapter 7 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

Consistent with the levels at which DLM results are reported, this chapter provides reliability evidence at six levels: (1) the classification accuracy of each linkage level within each EE (linkage level reliability), (2) the classification accuracy summarized for the three linkage levels (conditional evidence by linkage level), (3) the number of linkage levels mastered within each EE (EE reliability), (4) the number of linkage levels mastered within each topic in Biology and domain in science grade bands (topic or domain reliability), (5) the total number of linkage levels mastered (subject reliability), and (6) the classification to overall performance level (performance level reliability). As described in the next section, reliability evidence comes from simulated retests in which assessment data are generated for students with the estimated model parameters and student mastery probabilities.

The reliability methods and evidence presented in this chapter adhere to guidance given in the *Standards for Educational and Psychological Testing* (AERA et al., 2014). Where relevant, evidence provided in accordance with specific standards is noted.

---

<sup>12</sup> For a description of how often each scoring rule is utilized, see Chapter 7 of this manual.

## 8.2. Methods of Obtaining Reliability Evidence

Because the DLM psychometric model produces complex mastery results summarized at multiple levels of aggregation (linkage level, EE, topic or domain, subject, and performance level), rather than a traditional raw or scale score value, methods for evaluating reliability were based on simulated retests. For a simulation-based method of computing reliability, we generate simulated examinees with known characteristics, simulate assessment data using calibrated-model parameters, score the simulated retests using calibrated-model parameters, and compare estimated examinee characteristics with those characteristics known to be true in the simulation. For DLM assessments, the known characteristics of the simulated examinees are the set of linkage levels the examinee has mastered and not mastered.

Most methods for estimating the reliability of assessments scaled with DCMs are limited to attribute-level summaries of reliability (for a review see Sinharay & Johnson, 2019). Accordingly, these methods do not generalize to aggregated summaries of mastery classifications, such as those reported for DLM assessments. Simulated retests offer one method for estimating the reliability of DCM-based assessments at multiple levels of aggregation. At the attribute level (i.e., individual mastery classifications), simulated retests provide reliability estimates that are highly consistent with nonsimulation-based methods (Thompson, 2020). However, unlike the nonsimulation-based methods, simulated retests are able to support the evaluation of reliability for aggregations of individual mastery classifications (Thompson et al., 2019). In addition to supporting the evaluation of reliability evidence at multiple levels of reporting, simulated retests provide results consistent with classical reliability metrics in that perfect reliability is evidenced by consistency in classification, and zero reliability is evidenced by a lack of classification consistency.

The simulated retests used to estimate reliability for DLM versions of scores and classifications consider the unique design and administration of DLM assessments. Students typically take only 3–5 items per EE. Simulated retests are based on a replication of the administration process, including adaptive routing between testlets, and uses students' known mastery classifications from the operational assessment. Therefore, students may not receive the same testlets in the simulation as they did during their actual assessment (i.e., routing decisions may be different or different testlets may be assigned from the pool of available testlets). This means that the simulated retest offers a genuine approximation of actual retest assignment for any given student. Simulated retests replicate results of DLM assessments from actual examinees based on administration procedures specific to the DLM assessments. However, the use of simulation produces approximate estimates of reliability, which are contingent on the accuracy of the current scoring model. That is, reliability estimates are an upper bound on the true reliability. For the 2021–2022 administration, on advice of the DLM Technical Advisory Committee, the procedure for simulating the retest was updated to incorporate additional uncertainty into the simulation-assigned mastery status and testlets for each resampled student. The remaining sections of this chapter describe the current procedures and results with these updates, which provide a better estimate of the true reliability.

Simulated retests 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 reliability evidence reported here supports “interpretation for each intended score use,” as Standard 2.0 recommends (AERA et al., 2014, p. 42). The “appropriate evidence of reliability/precision” (AERA et al., 2014, p. 42) was assembled using a methodology that aligns to the design of the assessment and interpretations of results. The procedures

used to assemble reliability evidence align with all applicable standards.

### **8.2.1. Reliability Sampling Procedure**

The simulation design that was used to obtain the reliability estimates uses a resampling design to mirror DLM assessment data. In accordance with Standard 2.1, the sampling design uses the entire set of operational assessment data to generate simulated examinees (AERA et al., 2014, p. 42). Using 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 is as follows:

1. Draw with replacement the student record of one student from the operational assessment data (i.e., the spring assessment window). Use the student's originally scored linkage level mastery probabilities as the true values for the simulated student data. For linkage levels the drawn student was assessed on during the operational assessment, generate the mastery status from a Bernoulli distribution with a probability equal to the mastery probability. For linkage levels the student was not assessed on during the operational assessment, the probability is either fixed to 1 if the student has mastered a higher linkage level for the EE during the operational assessment, or is defined as the base rate of class membership for linkage levels higher than those assessed operationally (see Chapter 5 of this manual).
2. Simulate a new assessment based on administration rules. In practice, this means simulating one testlet at a time and applying routing rules (see Chapter 4 of this manual) to assign subsequent simulated testlets. Item responses are simulated for the assigned testlets from calibrated model parameters,<sup>13</sup> conditional on the linkage level mastery status determined in Step 1.
3. Score the simulated item responses using the operational DLM scoring procedure, estimating linkage level mastery or nonmastery for the simulated student. See Chapter 7 of this manual for more information.<sup>14</sup>
4. Calculate the aggregated summaries of linkage level mastery for the simulated retests (i.e., EE, topic or domain, subject, and performance level).
5. Compare the estimated linkage level mastery and aggregated summaries from the simulated retests to the values reported for the drawn student on the operational assessment.

Steps 1 through 5 are then repeated 100,000 times for each grade or course to create the full simulated retests data set. Figure 8.1 shows the steps of the simulation process as a flow chart.

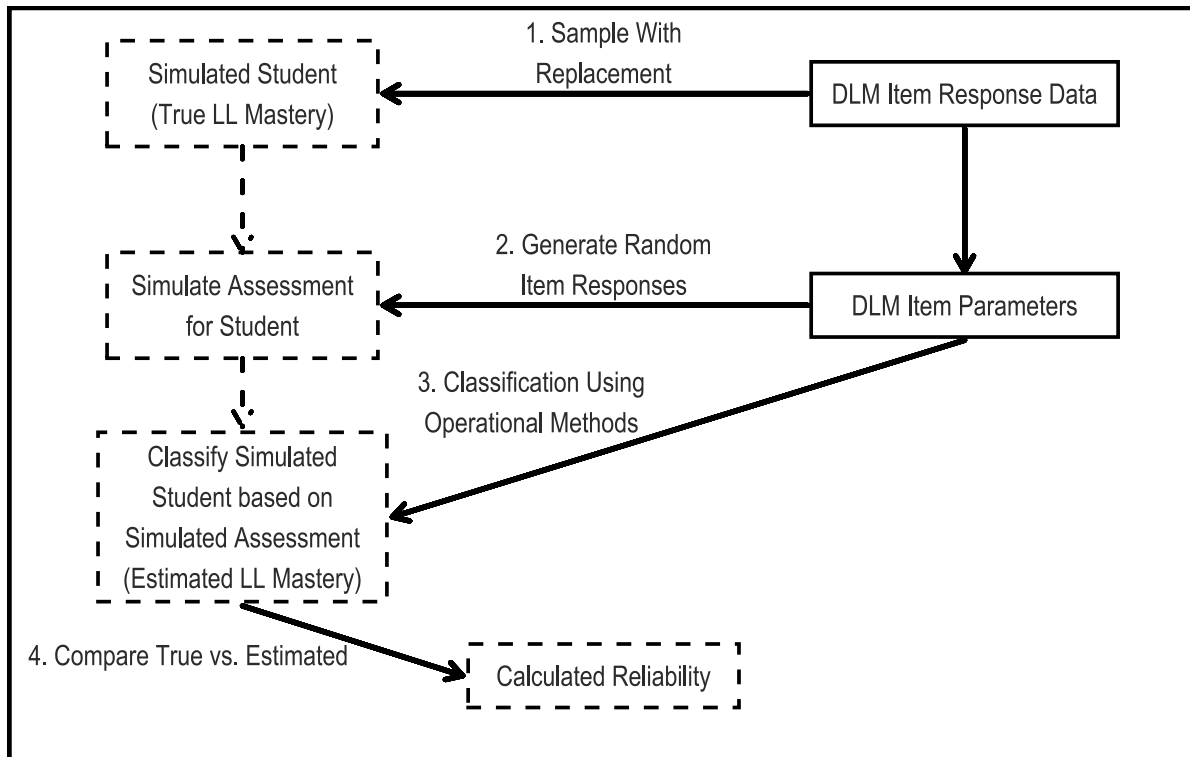
---

<sup>13</sup> Calibrated parameters were treated as true and fixed values for the simulated retests.

<sup>14</sup> All three scoring rules were implemented when scoring the simulated retests to be consistent with the operational scoring procedure.

**Figure 8.1**

*Simulation Process for Creating Reliability Evidence*



Note. LL = linkage level.

### 8.3. Reliability Evidence

This chapter provides reliability evidence for six levels of data: (1) linkage level reliability, (2) conditional reliability by linkage level, (3) EE reliability, (4) domain and topic reliability, (5) subject reliability, and (6) performance level reliability, which ensures that the simulation and resulting reliability evidence are aligned with Standard 2.2 (AERA et al., 2014, p. 42). Additionally, providing reliability evidence for each of the six levels ensures that these reliability estimation procedures meet Standard 2.5 (AERA et al., 2014, p. 43). With 34 EEs, each comprising 3 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. The website version of this report<sup>15</sup> provides a full report of reliability evidence for all 102 linkage levels and 34 EEs. The full set of evidence is provided in accordance with Standard 2.12 (AERA et al., 2014, p. 45).

Reliability evidence at each level is reported using various correlation coefficients. Correlation estimates mirror estimates of reliability from contemporary measures such as the Guttman-Cronbach alpha. For linkage level and conditional evidence by linkage level reliability, the tetrachoric correlation estimates the relationship between true and estimated linkage level mastery statuses. The tetrachoric correlation is a

<sup>15</sup> <https://2022-sci-techmanual.dynamiclearningmaps.org/8-reliability>

special case of the polychoric in which the variables are discrete. Both the polychoric and tetrachoric correlations provide more useful estimates of relationships between ordinal and discrete variables that would otherwise be attenuated using the standard correlation (i.e., the Pearson coefficient). For EE and performance level reliability, the polychoric correlation estimates the relationship between two ordinal variables: the true performance level or true number of linkage levels mastered and the corresponding estimated value. Finally, for subject and topic or domain reliability, the Pearson correlation estimates the relationship between the true and estimated numbers of linkage levels mastered.

Reliability evidence at each level is also reported using correct classification rates (raw and chance corrected), indicating the proportion of estimated classifications that match true classifications. The chance-corrected classification rate, kappa, represents the proportion of error reduced above chance. Kappa values above .6 indicate substantial-to-perfect agreement between estimated and true values (Landis & Koch, 1977). However, Cohen's kappa may be limited in this context. Numerous studies have shown that the kappa statistic tends to be too conservative when there are unbalanced categories (Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990; O'Leary et al., 2014; Pontius & Millones, 2011). This is often the case for DLM linkage levels, where the base rate of class membership can be less than .3 or greater than .7.<sup>16</sup> Thus, it is important to interpret the kappa statistic with caution and within the context of the other reporting metrics.

### **8.3.1. 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. This level of reliability reporting is the finest grain of reporting, and while it does not have a directly comparable classical test theory or item response theory analogue, its inclusion is important because it is the level at which mastery classifications are made for DLM assessments. All reported summary statistics of linkage level reliability are based on the resulting contingency tables: the comparison of true (operational assessment) and estimated (simulated retest) mastery statuses across all simulated examinees.

In addition to summary statistics from the simulated retests, we also calculated the classification consistency metric,  $\hat{P}_C$ , described by Johnson and Sinharay (2018). As the name implies, the classification consistency index is a measure of how consistent the student-level classifications are for each linkage level, and it is calculated from the estimated DCM parameters (see Chapter 5 of this manual for a description of the model parameters). The classification consistency metric is based on the estimated model parameters, and thus is only applicable to the linkage level, which is the unit of model estimation.<sup>17</sup> This metric is not based on simulated retests, and thus provides a measure of reliability independent from the simulation.

For each statistic, figures are given comparing the results of all 102 linkage levels. We report linkage level reliability evidence based on three summary statistics from the simulated retests and the nonsimulation-based classification consistency:

1. the tetrachoric correlation between estimated and true mastery status,
2. the classification agreement for the mastery status of each linkage level,

<sup>16</sup> See Chapter 5 of this manual for a summary of base rates of class membership.

<sup>17</sup> See Chapter 5 of this manual for a complete description of the model specification.



3. the classification consistency (Johnson & Sinharay, 2018), and
4. the classification agreement Cohen's kappa for the mastery status of each linkage level.

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 for individual linkage levels can be found in the website version of this report<sup>18</sup>). Results are given in both tabular and graphical forms. Table 8.1 and Figure 8.2 provide proportions and the number of linkage levels, respectively, that fall within prespecified ranges of values for the four linkage level reliability summary statistics (i.e., tetrachoric correlation, classification agreement rate, classification consistency, and Cohen's kappa).

The correlations and classification agreement rates show reliability evidence for the classification of mastery at the linkage level. Across all linkage levels, 14 (14%) had a tetrachoric correlation below .6, 0 (0%) had a percent classification agreement below .6, 3 (3%) had a classification consistency below .6, and 83 (81%) had a Cohen's kappa below .6. As previously described, Cohen's kappa may be limited in this context due to unbalanced class categories. Thus, the other three metrics provide a more useful evaluation of linkage level reliability.

Notably, Johnson and Sinharay (2018) recommend a cutoff of .7 for fair classification consistency. Overall, 79 (77%) linkage levels meet this cutoff, indicating that the linkage level classifications show a high degree of reliability.<sup>19</sup>

**Table 8.1**

*Reliability Summaries Across All Linkage Levels: Proportion of Linkage Levels Falling Within a Specified Index Range*

Reliability index	Index range								
	0.00– 0.59	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	.137	.088	.157	.157	.196	.127	.069	.069	.000
Percent classification agreement	.000	.000	.010	.059	.206	.245	.275	.137	.069
Classification consistency	.029	.059	.137	.088	.167	.245	.118	.098	.059
Cohen's kappa	.814	.098	.039	.049	.000	.000	.000	.000	.000

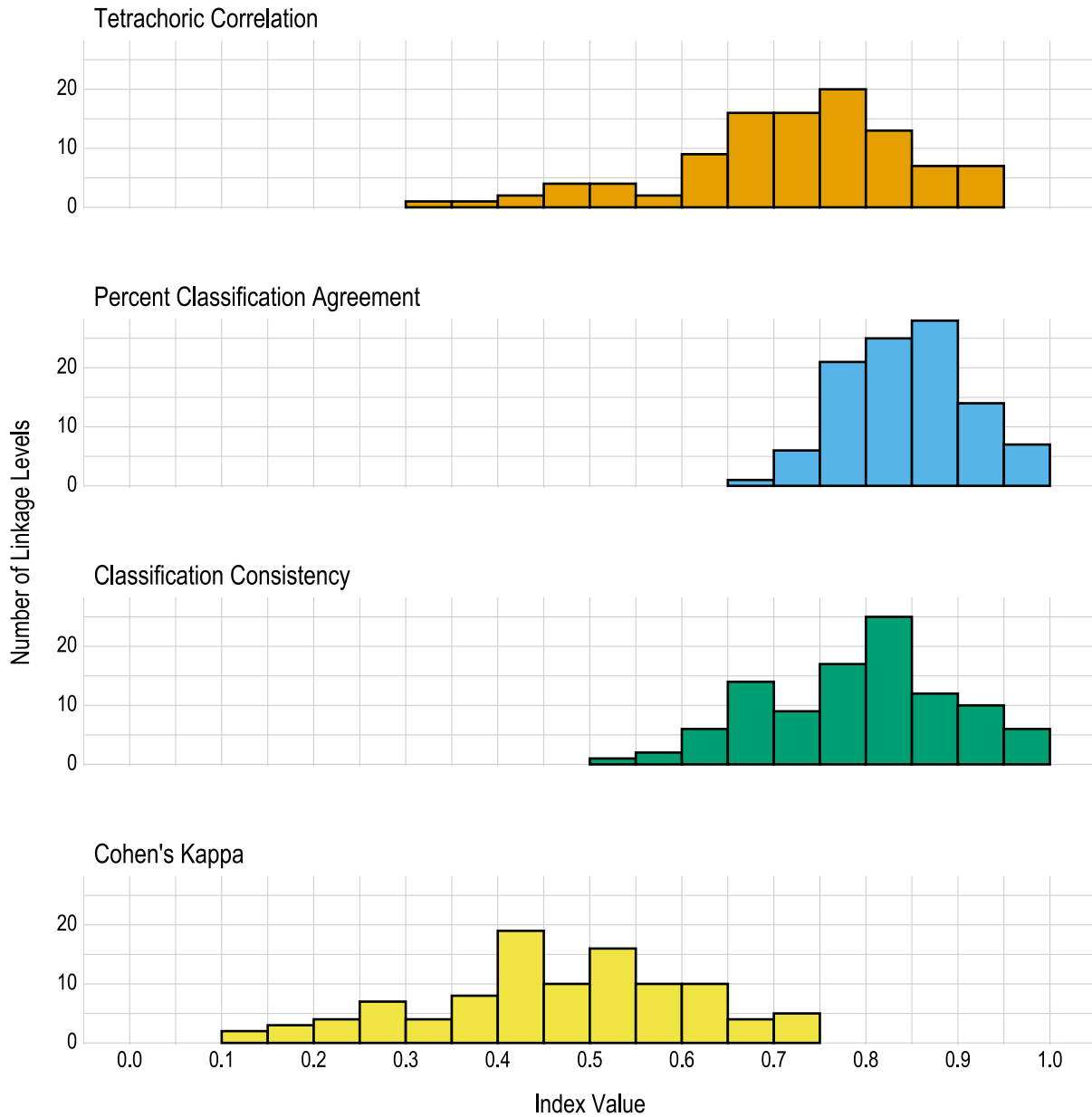
<sup>18</sup> <https://2022-sci-techmanual.dynamiclearningmaps.org/8-reliability>

<sup>19</sup> For a summary of the closely related classification accuracy, see Chapter 5 of this manual.



**Figure 8.2**

*Summaries of Linkage Level Reliability*



### **8.3.2. 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. Therefore, traditional measures of conditional reliability are not applicable. In particular, standard errors of measurement (inversely related to reliability) that are conditional on a continuous trait are based on the calculation of Fisher’s information, which involves taking

the second derivative-model likelihood function with respect to the latent trait. When classifications are the latent traits, however, the likelihood is not a smooth function regarding levels of the trait and therefore cannot be differentiated (Henson & Douglas, 2005; Templin & Bradshaw, 2013). In other words, because diagnostic classification modeling does not produce a total score or scale score, traditional methods of calculating conditional standard errors of measurement are not appropriate. However, because DLM assessments were designed to span the continuum of students' varying knowledge, skills, and understandings 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, as summarized by each of the three levels. Results are reported using the same four statistics used to summarize the overall linkage level reliability evidence (i.e., tetrachoric correlation, classification agreement rate, classification consistency, and Cohen's kappa).

Table 8.2 and Figure 8.3 provide the proportions and the number of linkage levels, respectively, that fall within prespecified ranges of values for each linkage level for the four reliability summary statistics (i.e., tetrachoric correlation, classification agreement rate, classification consistency, and Cohen's kappa). The correlations and classification agreement rates generally indicate that all three linkage levels provide reliable classifications of student mastery; results are fairly consistent across all linkage levels for each of the four statistics reported.

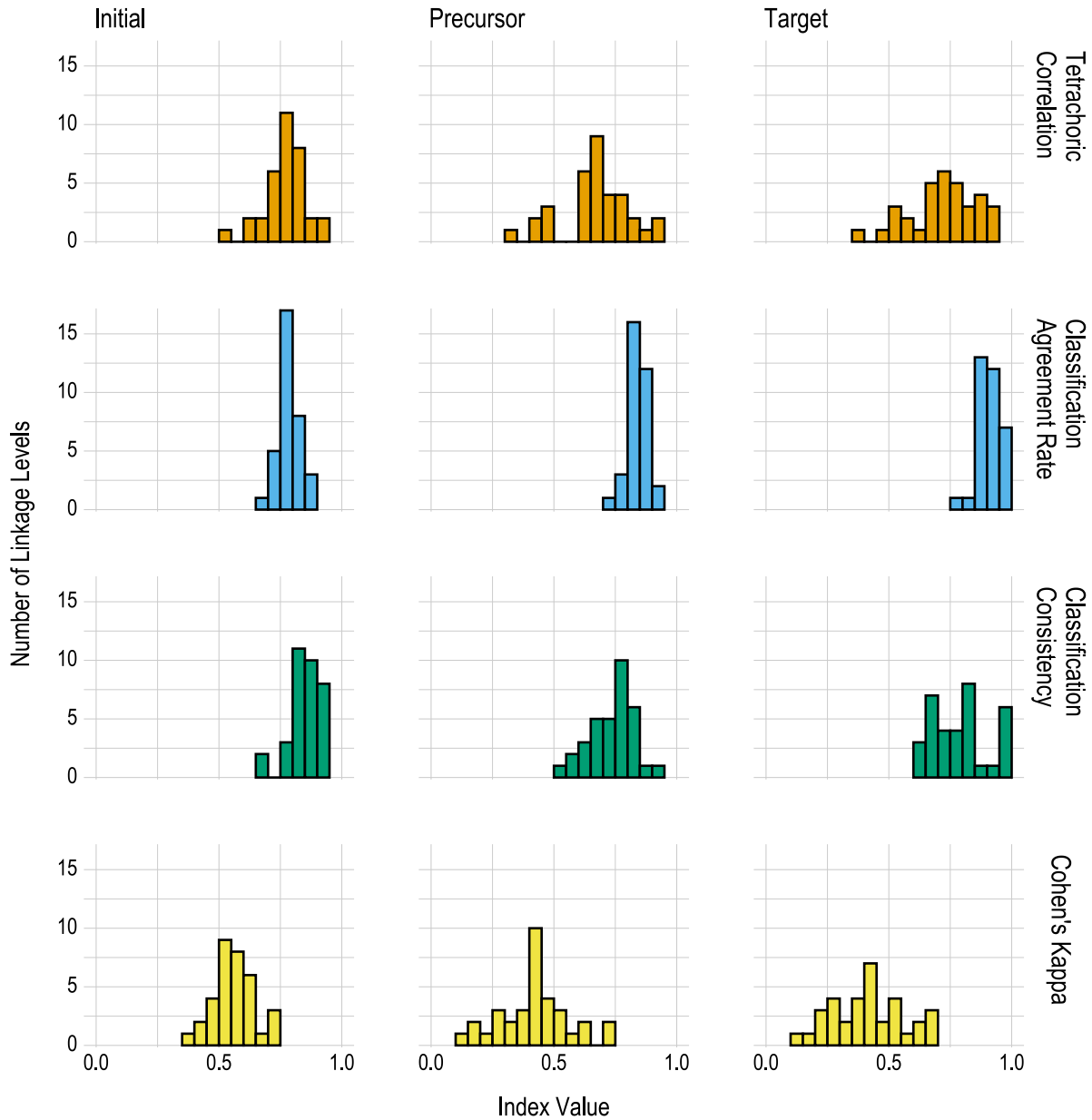
**Table 8.2**

*Reliability Summaries Across All Linkage Levels: Proportion of Linkage Levels Falling Within a Specified Index Range*

Reliability index	Index range								
	0.00– 0.59	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
<b>Initial</b>									
Tetrachoric correlation	.029	.059	.059	.176	.324	.235	.059	.059	.000
Percent classification agreement	.000	.000	.029	.147	.500	.235	.088	.000	.000
Classification consistency	.000	.000	.059	.000	.088	.324	.294	.235	.000
Cohen’s kappa	.706	.176	.029	.088	.000	.000	.000	.000	.000
<b>Precursor</b>									
Tetrachoric correlation	.176	.176	.265	.118	.118	.059	.029	.059	.000
Percent classification agreement	.000	.000	.000	.029	.088	.471	.353	.059	.000
Classification consistency	.088	.088	.147	.147	.294	.176	.029	.029	.000
Cohen’s kappa	.882	.059	.000	.059	.000	.000	.000	.000	.000
<b>Target</b>									
Tetrachoric correlation	.206	.029	.147	.176	.147	.088	.118	.088	.000
Percent classification agreement	.000	.000	.000	.000	.029	.029	.382	.353	.206
Classification consistency	.000	.088	.206	.118	.118	.235	.029	.029	.176
Cohen’s kappa	.853	.059	.088	.000	.000	.000	.000	.000	.000

**Figure 8.3**

*Conditional Reliability Evidence Summarized by Linkage Level*



### **8.3.3. Essential Element Reliability Evidence**

The first level of linkage level aggregation is the EE. EE-level results are reported as the highest linkage level mastered for each EE. Because EE-level results are an aggregation of the individual linkage level classifications, more traditional measures of the reliability (e.g., the classification consistency used for linkage levels) are not possible. Therefore, reliability results are only reported based on the simulated retests, which do offer a method for evaluating the reliability of aggregated classifications (Thompson et al.,

2019).

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 classification agreement rate for the number of linkage levels mastered within an EE, and
3. the classification agreement Cohen’s 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 for individual EEs can be found in the website version of this report<sup>20</sup>). Results are given in both tabular and graphical forms. Table 8.3 and Figure 8.4 provide the proportions and the number of EEs, respectively, falling within prespecified ranges of values for the three reliability summary statistics (i.e., classification agreement rate, kappa, correlation). Across all EEs, 8 (24%) had a polychoric correlation below .6, 3 (9%) had a percent classification agreement below .6, and 26 (76%) had a Cohen’s kappa below .6. In general, the reliability summaries show strong evidence for reliability for the number of linkage levels mastered within EEs.

**Table 8.3**

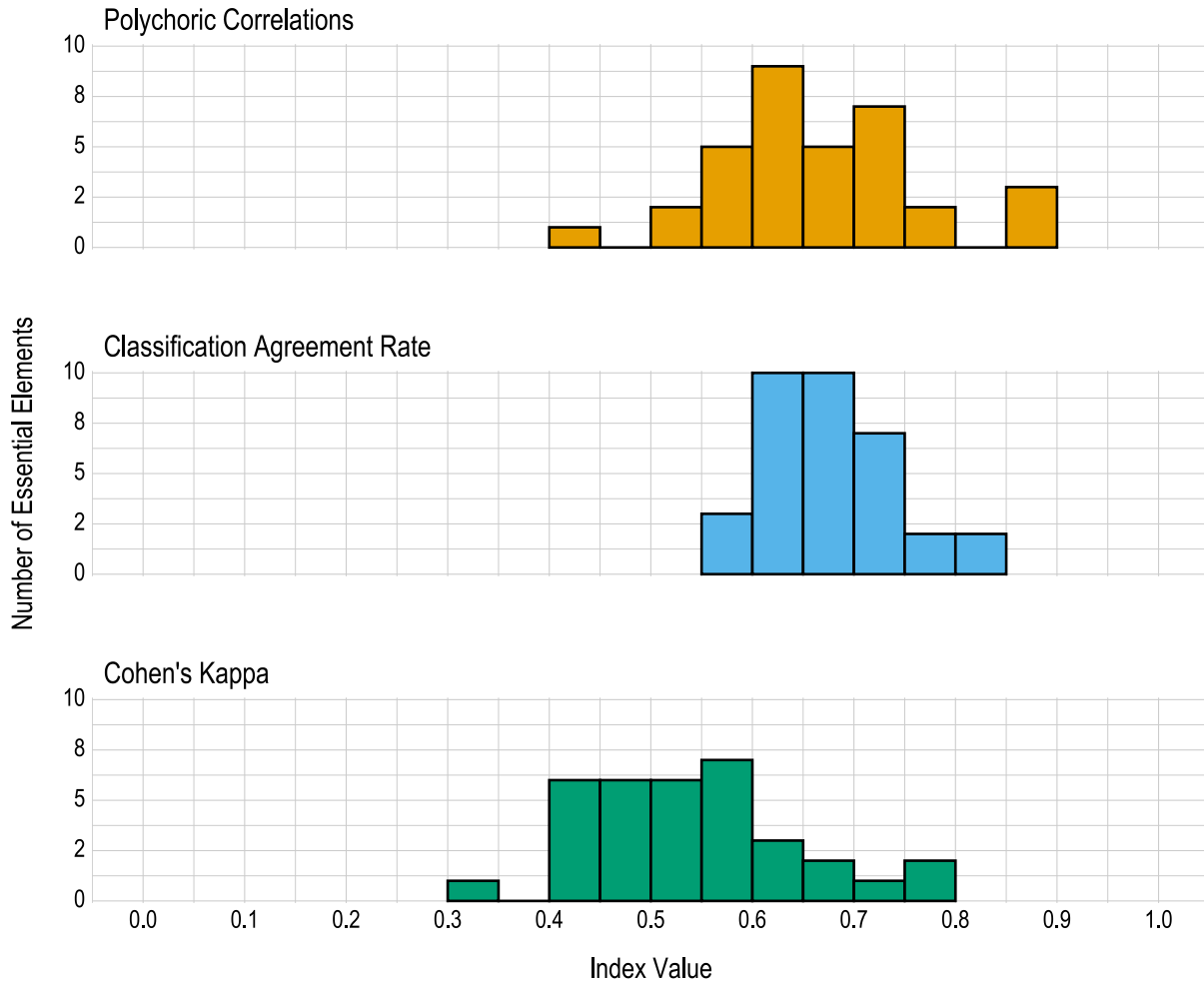
*Reliability Summaries Across All Essential Elements: Proportion of Essential Elements Falling Within a Specified Index Range*

Reliability index	Index range								
	0.00– 0.59	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	.235	.265	.147	.206	.059	.000	.088	.000	.000
Percent classification agreement	.088	.294	.294	.206	.059	.059	.000	.000	.000
Cohen’s kappa	.765	.088	.059	.029	.059	.000	.000	.000	.000

<sup>20</sup> <https://2022-sci-techmanual.dynamiclearningmaps.org/8-reliability>

**Figure 8.4**

*Number of Linkage Levels Mastered Within Essential Element Reliability Summaries*



### **8.3.4. Domain and Topic Reliability Evidence**

Science EEs are organized into topics and domains, which are akin to content strands for other assessments. Because Individual Student Score Reports summarize the number and percentage of linkage levels students mastered for each science domain or topic in Biology,<sup>21</sup> we also provide reliability evidence at these levels in their respective grade or course, in accordance with Standard 2.2, which indicates that reliability evidence should be provided consistent with the level(s) of scoring (AERA et al., 2014, p. 42).

Reliability at the domain or topic level provides consistency evidence for the number of linkage levels mastered across all EEs in each science domain for each grade or topic for Biology. Domain and topic reliability evidence compares the true and estimated number of linkage levels mastered across all tested

<sup>21</sup> See Chapter 7 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017) for more information.

levels for each of the three domains and Biology topics. Reliability is reported with three summary numbers:

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

Table 8.4 shows the three summary values for each domain in each grade and topic in Biology. Classification rate information is provided in accordance with Standard 2.16 (AERA et al., 2014, p. 46). The correlation and Cohen’s kappa summary statistics included in Table 8.4 also align with Standard 2.19 (AERA et al., 2014, p. 47). The correlations range from .172 to 1.000, the average classification agreement rates range from .922 to .999, and the average Cohen’s kappa values range from .830 to .999. This indicates that, overall, the domain and topic results provided on score reports are reliable.

**Table 8.4**  
*Summary of Domain and Topic Reliability Evidence*

Grade	Domain/Topic	Linkage levels mastered correlation	Average student classification agreement	Average student Cohen’s kappa
3	SCI.ESS	.611	.987	.977
3	SCI.LS	.480	.994	.992
3	SCI.PS	.844	.982	.964
4	SCI.ESS	.497	.965	.926
4	SCI.LS	.493	.993	.990
4	SCI.PS	.743	.948	.880
5	SCI.ESS	.529	.970	.939
5	SCI.LS	.498	.993	.990
5	SCI.PS	.763	.957	.903
6	SCI.ESS	.587	.951	.897
6	SCI.LS	.597	.982	.968
6	SCI.PS	.624	.972	.945
7	SCI.ESS	.604	.949	.894
7	SCI.LS	.598	.979	.962
7	SCI.PS	.653	.969	.939
8	SCI.ESS	.535	.922	.830
8	SCI.LS	.579	.972	.947
8	SCI.PS	.557	.946	.885

**Table 8.4**

*Summary of Domain and Topic Reliability Evidence (continued)*

Grade	Domain/Topic	Linkage levels mastered correlation	Average student classification agreement	Average student Cohen's kappa
9	SCI.ESS	.564	.961	.921
9	SCI.LS	.586	.973	.950
9	SCI.PS	.520	.942	.879
10	SCI.ESS	.606	.974	.951
10	SCI.LS	.605	.979	.962
10	SCI.PS	.590	.960	.921
11	SCI.ESS	.544	.960	.919
11	SCI.LS	.564	.971	.945
11	SCI.PS	.515	.944	.882
12	SCI.ESS	.550	.964	.928
12	SCI.LS	.569	.974	.951
12	SCI.PS	.510	.944	.882
Biology	SCI.LS1.A	.548	.948	.886
Biology	SCI.LS1.B	>.999	.999	.999
Biology	SCI.LS2.A	.372	.974	.953
Biology	SCI.LS3.B	.172	.999	.999
Biology	SCI.LS4.C	.529	.964	.927

### **8.3.5. Subject Reliability Evidence**

The next level of aggregation of linkage level mastery is for the subject overall. Subject reliability provides consistency evidence for the number of linkage levels mastered across all EEs for a given grade or course. Because students are assessed on multiple linkage levels across the assessed EEs in science, subject reliability evidence is similar to reliability evidence for testing programs that use summative assessments to describe overall performance in a subject. 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. Because subject-level reporting summarizes the total number of linkage levels a student mastered, the statistics reported for subject reliability are the same as those reported for domain and topic reliability. 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 classification agreement rate for which linkage levels were mastered, as averaged across all



- simulated students, and
- the classification agreement Cohen’s kappa for which linkage levels were mastered, as averaged across all simulated students.

Table 8.5 shows the three summary values for each grade and subject. The correlation between true and estimated number of linkage levels mastered ranges from .721 to .873. Students’ average classification agreement rates range from .866 to .956 and average Cohen’s kappa values range from .634 to .881. These values indicate that the total linkage levels mastered in each grade or course are reliably determined.

**Table 8.5**  
*Summary of Subject Reliability Evidence*

Grade	Linkage levels mastered correlation	Average student classification agreement	Average student Cohen’s kappa
3	.873	.956	.881
4	.783	.909	.731
5	.803	.918	.760
6	.789	.909	.745
7	.807	.904	.736
8	.747	.866	.634
9	.763	.892	.702
10	.779	.913	.766
11	.734	.892	.701
12	.754	.902	.730
Biology	.721	.896	.657

### 8.3.6. Performance Level Reliability Evidence

The final level of linkage level mastery aggregation is at the overall performance level. Results for DLM assessments are reported using four performance levels. The scoring procedure sums the linkage levels mastered across all EEs, and cut points are applied to distinguish between the four performance categories.<sup>22</sup>

Performance level reliability provides evidence for how reliably students are classified to the four performance levels for each subject and grade level. Because the 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 observed and estimated

<sup>22</sup> See Chapter 6 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017) and Chapter 6 of the *2018–2019 Technical Manual Update—Science* (DLM Consortium, 2019) for details on the standard setting procedure to determine the cut points.

performance levels from the simulated retests (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:

1. the polychoric correlation between the true and estimated performance levels within a grade and subject,
2. the classification agreement rate between the true and estimated performance levels within a grade and subject, and
3. the classification agreement Cohen’s kappa between the true and estimated performance levels within a grade and subject.

Table 8.6 presents this information across all grades in science. Polychoric correlations between true and estimated performance level range from .760 to .913. Classification agreement rates range from .638 to .790, and Cohen’s kappa values are between .594 and .817. These results indicate that the DLM scoring procedure of reporting performance levels based on total linkage levels mastered results in reliable classification of students to performance level categories.

**Table 8.6**

*Summary of Performance Level Reliability Evidence*

Grade	Polychoric correlation	Classification agreement rate	Cohen’s kappa
3	.913	.790	.817
4	.821	.714	.646
5	.844	.782	.663
6	.832	.715	.707
7	.841	.707	.725
8	.773	.655	.630
9	.787	.638	.616
10	.806	.705	.662
11	.760	.656	.594
12	.776	.666	.608
Biology	.761	.694	.606

## 8.4. Conclusion

In summary, reliability measures for the DLM assessment system address the standards set forth by AERA et al. (2014). The 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. The results indicate high levels of reliability for the individual linkage level mastery classifications, as well as for all levels of aggregation for which results are reported (i.e., EE, domain and topic, subject, and overall performance level). Because the reliability results depend on 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 administration process used in the operational assessment, which provides a rigorous evaluation of the variation in student results across simulated repeated assessment administrations.

## 9. Training and Professional Development

Chapter 10 of the Dynamic Learning Maps® (DLM®) Alternate Assessment System *2015–2016 Technical Manual—Science* (DLM Consortium, 2017) describes the training that was offered in 2015–2016 for state and local education agency staff, the required test administrator training, and the optional professional development provided. This chapter presents the participation rates and evaluation results from the 2021–2022 use of the optional instructional professional development. This chapter also describes the updates made to the professional development system during 2021–2022.

For a complete description of training and professional development for DLM assessments, including a description of training for state and local education agency staff, along with descriptions of facilitated and self-directed training, see Chapter 10 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017).

### 9.1. Updates to Required Test Administrator Training

Two main revisions were made to the DLM Required Test Administrator Training for 2021–2022. Slide content and narrations were revised to complement each of these revisions.

First, all modules were revised to comply with the Americans with Disabilities Act (ADA) guidelines. These revisions include alternative text for all images and tables, a unique title for each slide, and more succinct slide and narration content.

Second, all screenshots of the DLM website were removed from the modules in anticipation of an overhauled website launch in late July of 2021. The screenshots were no longer necessary, as all resources on the revised site are easier to find using filters or an enhanced search bar. The revised modules focus more on the name and usefulness of each resource rather than relying on directions for how to navigate the website.

Additionally, one minor revision was made to the slide and narration regarding the Security Agreement. The Security Agreement for 2021–2022 no longer has a choice to reject the agreement, only to a box to accept the agreement. Therefore, the slide and narration were revised to reflect this change.

The total time needed to complete the training, activities, and post-tests is still under three hours.

### 9.2. Instructional Professional Development

The DLM professional development system includes eight modules that address instruction in science and support educators in creating Individual Education Programs that are aligned with the DLM Essential Elements. While the modules were originally intended for educators who administer DLM assessments, demographic information suggests that preservice educators, related service providers, parents, and others also accessed and completed the modules.

The professional development system is built in WordPress, an open-source website content management system. The professional development modules and instructional support materials are available for anyone's use at <https://dlmpd.com> or through a direct link from the DLM website. These DLM professional development modules address instruction in science. The modules also address processes educators can apply to create Individual Education Programs that are aligned with the DLM Essential Elements and

supports they can provide to address the communication needs of the students they teach. Finally, the modules help them understand the components of the DLM assessment system more completely.

To support state and local education agencies in providing continuing education credits to educators who complete the modules, each module also includes a time-ordered agenda, learning objectives, and biographical information regarding the faculty who developed the training modules. There are a total of eight modules, which are described in section 9.2.1.2 of this chapter.

The eight modules are available in both self-directed and facilitated formats. The self-directed modules are available online, on-demand. The interactive modules include a combination of video-based content, embedded activities, and, for participants who would like to receive a certificate documenting their successful completion of the module, a five-item pre- and post-test. These certificates are sent directly to each participant's email when they score 80% or higher on the post-test.

Modules in the facilitated format were created for groups and by individuals who prefer to learn by reading the contents rather than interacting with video and other package contents.

In addition to the eight modules, the instructional professional development site provides instructional resources for educators. These include DLM Essential Element unpacking documents; vignettes that illustrate shared reading with students with the most complex needs across the grade levels; supports for augmentative and alternative communication for students who do not have a comprehensive, symbolic communication system; alternate pencils for educators to download and use with students who cannot use a standard pen, pencil, or computer keyboard; and links to Pinterest boards and other online supports.

During the 2021–2022 school year, teams at ATLAS worked in cooperation with the professional development team at the University of North Carolina (UNC) at Chapel Hill to develop a new science professional development module titled “Science and Engineering Practices #2: Developing and Using Models.” This module focuses on teaching the “developing and using models” component of one science and engineering practice through the use of a writing framework. The module also includes guidance for teaching the science and engineering and framework through differing levels of complexity using DLM Science Essential Elements and linkage levels. The writing and production process of the module included collaboration between teams to ensure alignment with existing English language arts (ELA) and mathematics learning modules for a more streamlined DLM learner experience. The module will be released for the 2022–2023 year, and evaluation data will be included in future technical manual updates.

### **9.2.1. Professional Development Participation and Evaluation**

There are two ways in which test administrators and educators may complete professional development modules: required test administrator training or optional professional development. As described in Chapter 10 of the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017), some states require professional development modules to be completed as part of the required test administrator training. States can require certain modules be completed by new test administrators, returning test administrators, or both. Test administrators completing professional development modules as part of the required test administrator training access the modules through the Moodle training site where the rest of the required test administration training is located. The second way in which professional development modules are

completed is through the DLM professional development website<sup>23</sup>. The modules on the professional development website may be accessed by anyone and can be completed at any time. Additionally, participants completing modules on the professional development website are administered a short evaluation survey following the module. With the exception of the evaluation survey, the content of the modules is identical between the required test administrator training and the professional development website.

### 9.2.1.1. Required Professional Development Participation

A total of nine states required at least one professional development module as part of their required test administrator training. The modules included in the required test administrator training are required of all relevant test administrators (i.e., new or returning, as specified by the state). For example, a test administrator who only administers science assessments may still be required to complete a module on instruction for ELA. Table 9.1 shows the number of modules required, by state, for new and returning test administrators, as well as the total number of modules completed. In total, 21,888 professional development modules were completed by 4,431 new and 3,033 returning test administrators as part of the required training.

**Table 9.1**

*Number of Professional Development Modules Completed as Part of the Required Test Administrator Training*

State	New test administrators		Returning test administrators		Total modules completed
	Required modules	Test Administrators	Required modules	Test Administrators	
Alaska	1	84	1	130	214
Delaware	2	359	—	—	718
Maryland	4	1,466	—	—	5,864
New Hampshire	1	152	—	—	152
New Mexico	3	824	—	—	2,472
Oklahoma	5	649	5	1,305	9,770
Rhode Island	1	110	1	190	300
West Virginia	2	203	1	364	770
Wisconsin	1	584	1	1,044	1,628

Table 9.2 shows which modules were required for new and returning test administrators across all states choosing to include professional development modules in the required training. For example, the *DLM Essential Elements* module was required for new test administrators in four states and was required for returning test administrators in two states.

<sup>23</sup> <https://dlmpd.com>

**Table 9.2**

*Professional Development Modules Selected for Inclusion in Required Test Administrator Training*

Module	States requiring for new test administrators	States requiring for returning test administrators	Total modules completed
DLM Claims and Conceptual Areas	1	—	1,466
DLM Essential Elements	4	2	3,993
Effective Instruction in Mathematics	3	2	3,987
Individual Education Programs Linked to the DLM Essential Elements	4	2	3,437
Principles of Instruction in English Language Arts	3	2	5,048
Universal Design for Learning	1	—	824
Who are Students With Significant Cognitive Disabilities?	4	1	3,133

### 9.2.1.2. Optional Professional Development Participation

In addition to the modules included in the required test administrator training, the DLM professional development website<sup>24</sup> contains modules specific to science. Table 9.3 shows the number of individuals who completed optional professional development science, as well as the total number of test administrators from each state who had a student assigned to them for the DLM assessment. In total, 450 science modules were completed in the self-directed format from August 1, 2021, to July 31, 2022. Since the first module was launched in the fall of 2017, a total of 3,502 modules have been completed on the professional development website.

<sup>24</sup> <https://dlmpd.com>

**Table 9.3**

*Number of Self-Directed Modules Completed in 2021–2022 by Educators in DLM States and Other Localities (N = 450)*

State	Participants	DLM test administrators	Total modules completed
Alaska	0	123	0
Arkansas	0	759	0
Delaware	0	167	0
District of Columbia	5	93	17
Illinois	1	2,005	1
Iowa	1	543	1
Kansas	1	544	1
Maryland	4	1,046	16
Missouri	3	1,114	5
New Hampshire	0	193	0
New Jersey	3	1,859	4
New Mexico	1	463	7
New York	4	3,978	14
North Dakota	0	165	0
Oklahoma	3	989	10
Pennsylvania	4	3,253	25
Rhode Island	1	174	1
Utah	7	714	14
West Virginia	1	342	3
Wisconsin	38	1,026	117
Non-DLM state and other locations	55	—	214

*Note.* Participant counts may include individuals who are not educators or test administrators (e.g., pre-service educators).

To evaluate educator perceptions of the utility and applicability of the modules, DLM staff asked educators to respond to a series of evaluation questions on completion of each self-directed module. Three questions asked about importance of content, whether new concepts were presented, and the utility of the module. Educators responded using a 4-point scale ranging from *strongly disagree* to *strongly agree*. A fourth question asked whether educators planned to use what they learned, with the same response options. During the 2021–2022 year, educators completed the evaluation questions 69% of the time. The responses were generally positive, as illustrated in Table 9.4. Across all science modules, 62% of respondents either agreed or strongly agreed with each statement.

To evaluate the consistency in the ratings for each module, we calculated Cronbach’s (1951) alpha from the four items for each module using all ratings from fall 2017 through the 2021–2022 year. Across all modules, alpha ranged from .92 to .97 with an average value of .96, suggesting high internal consistency in



responses.

**Table 9.4**

*Response Rates and Rate of Agree or Strongly Agree on 2021–2022 Self-Directed Module Evaluation Questions*

Module	Total modules completed ( <i>n</i> )	Response rate	The module addressed content that is important for professionals working with SWSCDs. (%)	The module presented me with new ideas to improve my work with SWSCDs. (%)	Completing this module was worth my time and effort. (%)	I intend to apply what I learned in the module to my professional practice. (%)
DLM Science Standards Framework Part 1	34	85.3	79.4	79.4	79.4	79.4
DLM Science Standards Framework Part 2	26	80.8	76.9	73.1	73.1	73.1
Instructional Strategies for Teaching DLM Science Part 1	127	75.6	68.5	67.7	67.7	67.7
Instructional Strategies for Teaching DLM Science Part 2	153	58.8	49.7	49.7	44.4	49.0
Instructional Strategies for Teaching DLM Science Part 3	67	74.6	74.6	73.1	74.6	74.6
Science and Engineering Practices #6: Constructing Explanations	14	57.1	57.1	50.0	57.1	50.0

**Table 9.4**

*Response Rates and Rate of Agree or Strongly Agree on 2021–2022 Self-Directed Module Evaluation Questions (continued)*

Module	Total modules completed ( <i>n</i> )	Response rate	The module addressed content that is important for professionals working with SWSCDs. (%)	The module presented me with new ideas to improve my work with SWSCDs. (%)	Completing this module was worth my time and effort. (%)	I intend to apply what I learned in the module to my professional practice. (%)
Science and Engineering Practices Part 1	20	45.0	45.0	45.0	45.0	45.0
Science and Engineering Practices Part 2	9	77.8	77.8	77.8	77.8	77.8
<i>Total</i>	450	68.9	63.1	62.2	60.9	62.2

*Note.* SWSCDs = students with significant cognitive disabilities.

### **9.3. Conclusion**

During 2021–2022, the required test administrator training had minor revisions to increase accessibility and focus on the most relevant content for those administering the DLM assessments. Instructional professional development modules continued to be available, and educators provided consistently positive feedback regarding the importance and relevance of the modules. Finally, a new module was developed, which will be available to educators in future years.

## 10. Validity Evidence

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 assessment provides students with the most significant cognitive disabilities the opportunity to demonstrate what they know and can do.

The 2021–2022 was the seventh operational administration of the DLM science assessments. This technical manual update provides updated evidence from the 2021–2022 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 10.1. Evidence summarized in this manual builds on the original evidence included in the *2015–2016 Technical Manual—Science* (DLM Consortium, 2017) and in the subsequent technical manual updates (DLM Consortium, 2018a, 2018b, 2019, 2020, 2021b). Together, the documents summarize the validity evidence collected to date.

**Table 10.1**

*Review of Technical Manual Update Contents*

Chapter	Contents
1	Provides an overview of information updated for the 2021–2022 year
2	Not updated for 2021–2022
3, 4	Provides evidence collected during 2021–2022 of assessment development and administration, including field-test information, item analyses, and test administrator survey results
5	Describes the statistical model used to produce results based on student responses, along with a summary of item parameters
6	Not updated for 2021–2022
7, 8	Describes results and analyses from the seventh 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 evidence collected during 2021–2022 on participation in professional development modules, including participant evaluations
10	Summarizes validity evidence collected during the 2021–2022 academic year

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.

## 10.1. Validity Evidence Summary

The accumulated evidence available by the end of the 2021–2022 year provides additional support for the validity argument. Three scoring interpretation and use claims are summarized in Table 10.2. 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 (Standards, AERA et al., 2014)*. While many sources of evidence contribute to multiple propositions, Table 10.2 lists the primary associations. For example, Proposition 4 is indirectly supported by content-related evidence described for Propositions 1 through 3. Table 10.3 shows the titles and sections for the chapters cited in Table 10.2.

**Table 10.2**

*DLM Alternate Assessment System Claims and Sources of Updated Evidence for 2021–2022*

Claim	Sources of evidence*				
	Test content	Response processes	Internal structure	Relations with other variables	Consequences of testing
1. Mastery results represent what students know and can do.	3.1, 3.2, 3.3, 3.4, 4.1, 4.3, 4.4, 4.5, 7.1, 7.2	4.1, 4.2	3.3, 3.4, 3.6, 5.1, 8.1		3.5, 7.1, 7.2
2. Results indicate summative performance relative to alternate achievement standards.	7.1, 7.2		8.1		3.5, 7.1, 7.2
3. Results can be used for instructional decision-making.					3.5

\* See Table 10.3 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 10.3**

*Evidence Sources Cited in Table 10.2*

Evidence no.	Chapter	Section
3.1	3	Testlet and Item Writing
3.2	3	External Reviews
3.3	3	Operational Assessment Items for 2021–2022
3.4	3	Field Testing
3.5	3	Educator Perception of Assessment Content
3.6	3	Evaluation of Item-Level Bias
4.1	4	User Experience With the DLM System
4.2	4	Accessibility Support Selections
4.3	4	Test Administration Observations
4.4	4	Student Experience
4.5	4	Opportunity to Learn
5.1	5	All
7.1	7	Student Performance
7.2	7	Score Reports
8.1	8	All

## 10.2. Continuous Improvement

As noted previously in this manual, 2021–2022 was the seventh year the DLM Science Alternate Assessment System was operational. While the 2021–2022 assessments were carried out in a manner that supports the validity of inferences made from results for the intended purposes, the DLM Governance Board is committed to continual improvement of assessments, educator 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 2022–2023. This section describes notable improvements from the sixth year to the seventh year of operational administration, as well as examples of improvements to be made during the 2022–2023 year.

### **10.2.1. Improvements to the Assessment System**

Overall, there were no significant changes to the item-writing procedures, item flagging outcomes, assessment administration procedures, or the process for scoring assessments from previous years to 2021–2022.

Based on an ongoing effort to improve the evaluation of the psychometric model used for scoring the assessments, the methods for calibrating the psychometric model and estimating reliability were updated in 2021–2022. The model calibration now implements a Bayesian rather than a maximum likelihood estimator, which allows for more accurate evaluations of model fit. The simulated retest method for estimating reliability now more closely approximates the conditions of test administration to better estimate the consistency in reported results. Both of these changes were implemented on the advice of the DLM Technical Advisory Committee (TAC).

The validity evidence collected in 2021–2022 expands upon the data compiled in the first six operational years for four of the critical sources of evidence as described in the *Standards* (AERA et al., 2014): evidence based on test content, response process, internal structure, and consequences of testing. Specifically, analysis of the opportunity to learn data contributed to the evidence collected based on test content. Test administrator 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. Test administrator survey responses also provided evidence based on consequences of testing. We summarize studies planned for 2022–2023 to provide additional validity evidence in the following section.

### **10.2.2. Future Research**

The continuous improvement process also leads to future directions for research to inform and improve the DLM System in 2022–2023 and beyond. The section describes some areas for further investigation.

DLM staff members are planning several studies for spring 2023 to collect data from educators in the states administering DLM assessments. The test administrator survey will collect information on educator ratings of student mastery as additional evidence to evaluate the extent that mastery ratings are consistent with other measures of student knowledge, skills, and understandings. In addition, the test administrator survey will continue to provide a source of data from which to investigate changes over time in the long-term effects of the assessment system for students and educators. DLM staff are also examining new ways to collect information on students' opportunity to learn and evaluate the extent to which educators provide aligned instruction. DLM staff will continue to collaborate with the DLM Governance Board on additional data collection as needed.

In addition to data collected from students and educators, there is an ongoing research agenda to improve the evaluation of item- and person-level model fit. A modeling subcommittee of DLM TAC members guides this research agenda.

Advice from the DLM TAC and DLM Governance Board will guide all future studies, using processes established over the life of the DLM System.



## 11. References

- Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian networks in educational assessment*. Springer. <https://doi.org/10.1007/978-1-4939-2125-6>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.
- Babu, G. J. (2011). Resampling methods for model fitting and model selection. *Journal of Biopharmaceutical Statistics*, 21(6), 1177–1186. <https://doi.org/10.1080/10543406.2011.607749>
- Betancourt, M. (2018, July 15). *A conceptual introduction to Hamiltonian Monte Carlo*. arXiv. <http://arxiv.org/abs/1701.02434>
- Bradshaw, L. (2016). Diagnostic classification models. In A. A. Rupp & J. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (1st ed., pp. 297–327). John Wiley & Sons. <https://doi.org/10.1002/9781118956588.ch13>
- Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice*, 33(1), 2–14. <https://doi.org/10.1111/emip.12020>
- Bradshaw, L., & Levy, R. (2019). Interpreting probabilistic classifications from diagnostic psychometric models. *Educational Measurement: Issues and Practice*, 38(2), 79–88. <https://doi.org/10.1111/emip.12247>
- Camilli, G., & Shepard, L. A. (1994). *Method for Identifying Biased Test Items* (4th). SAGE Publications, Inc.
- Carlin, B. P., & Louis, T. A. (2001). Empirical Bayes: Past, present and future. In A. E. Raftery, M. A. Tanner, & M. T. Wells (Eds.), *Statistics in the 21st century*. Chapman and Hall/CRC. <https://doi.org/10.1201/9781420035391>
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 1–32. <https://doi.org/10.18637/jss.v076.i01>
- Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, 39(2), 83–87. <https://doi.org/10.2307/2682801>
- Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, 50(2), 123–140. <https://doi.org/10.1111/j.1745-3984.2012.00185.x>
- Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43, 551–558. [https://doi.org/10.1016/0895-4356\(90\)90159-M](https://doi.org/10.1016/0895-4356(90)90159-M)
- Clark, A. K., Karvonen, M., Swinburne Romine, R., & Kingston, N. (2018, April 12–16). *Teacher use of score reports for instructional decision-making: Preliminary findings*. National Council on Measurement in Education Annual Meeting, New York, NY. [https://dynamiclearningmaps.org/sites/default/files/documents/presentations/NCME\\_2018\\_Score\\_Report\\_Use\\_Findings.pdf](https://dynamiclearningmaps.org/sites/default/files/documents/presentations/NCME_2018_Score_Report_Use_Findings.pdf)
- Clark, A. K., Kobrin, J., & Hirt, A. (2022). *Educator perspectives on instructionally embedded assessment* (Research Synopsis No. 22-01). University of Kansas, Accessible Teaching, Learning, and Assessment Systems. [https://dynamiclearningmaps.org/sites/default/files/documents/publication/IE\\_Focus\\_Groups\\_project\\_brief.pdf](https://dynamiclearningmaps.org/sites/default/files/documents/publication/IE_Focus_Groups_project_brief.pdf)
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159. <https://doi.org/10.1037//0033-2909.112.1.155>

- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1177/0146621612445470>
- DLM Consortium. (2021a). *Test Administration Manual 2021–2022*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems.
- Dynamic Learning Maps Consortium. (2017). *2015–2016 Technical Manual—Science*. University of Kansas, Center for Educational Testing and Evaluation.
- Dynamic Learning Maps Consortium. (2018a). *2016–2017 Technical Manual Update—Science*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems.
- Dynamic Learning Maps Consortium. (2018b). *2017–2018 Technical Manual Update—Science*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems.
- Dynamic Learning Maps Consortium. (2019). *2018–2019 Technical Manual Update—Science*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems.
- Dynamic Learning Maps Consortium. (2020). *2019–2020 Technical Manual Update—Science*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems.
- Dynamic Learning Maps Consortium. (2021b). *2020–2021 Technical Manual Update—Science*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems.
- Dynamic Learning Maps Consortium. (2021c). *Accessibility Manual 2021–2022*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems.
- Dynamic Learning Maps Consortium. (2021d). *Educator Portal User Guide*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems.
- Dynamic Learning Maps Consortium. (2022a). *2021–2022 Technical Manual—Instructionally Embedded Model*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems.
- Dynamic Learning Maps Consortium. (2022b). *2021–2022 Technical Manual—Year-End Model*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems.
- Efron, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statistical Science*, 29(2), 285–301. <https://doi.org/10.1214/13-STS455>
- Falconer, J. R., Frank, E., Polaschek, D. L. L., & Joshi, C. (2022). Methods for eliciting informative prior distributions: A critical review. *Decision Analysis*. <https://doi.org/10.1287/deca.2022.0451>
- Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43, 543–549. [https://doi.org/10.1016/0895-4356\(90\)90158-L](https://doi.org/10.1016/0895-4356(90)90158-L)
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10(4), 255–282. <https://doi.org/10.1007/BF02288892>
- Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, 29(4), 262–277. <https://doi.org/10.1177/0146621604272623>
- Henson, R., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74(2), 191–210. <https://doi.org/10.1007/s11336-008-9089-5>
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70. <http://www.jstor.org/stable/4615733>
- Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power raters using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, 14(4), 329–349. [https://doi.org/10.1207/S15324818AME1404\\_2](https://doi.org/10.1207/S15324818AME1404_2)

- Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement, 55*(4), 635–664. <https://doi.org/10.1111/jedm.12196>
- Karvonen, M., Bechard, S., & Wells-Moreaux, S. (2015, April 16–20). *Accessibility considerations for students with significant cognitive disabilities who take computer-based alternate assessments* [Paper presentation]. American Educational Research Association Annual Meeting, Chicago, IL.
- Karvonen, M., Wakeman, S. Y., Browder, D. M., Rogers, M. A. S., & Flowers, C. (2011). *Academic curriculum for students with significant cognitive disabilities: Special education teacher perspectives a decade after IDEA 1997* (Research Report). National Alternate Assessment Center. <https://files.eric.ed.gov/fulltext/ED521407.pdf>
- Kobrin, J., Clark, A. K., & Kavitsky, E. (2022). *Exploring educator perspectives on potential accessibility gaps in the Dynamic Learning Maps alternate assessment* (Research Synopsis No. 22-02). University of Kansas, Accessible Teaching, Learning, and Assessment Systems. [https://dynamiclearningmaps.org/sites/default/files/documents/publication/Accessibility\\_Focus\\_Groups\\_project\\_brief.pdf](https://dynamiclearningmaps.org/sites/default/files/documents/publication/Accessibility_Focus_Groups_project_brief.pdf)
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. <https://doi.org/10.2307/2529310>
- Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511611186>
- Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming, 45*(1), 503–528. <https://doi.org/10.1007/BF01589116>
- Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*(2), 187–212. <https://doi.org/10.1007/BF02294535>
- Mislevy, R. J., & Gitomer, D. H. (1995). The role of probability-based inference in an intelligent tutoring system. *User Modeling and User-Adapted Interaction, 5*(3–4), 253–282. <https://doi.org/10.1007/BF01126112>
- Nabi, S., Nassif, H., Hong, J., Mamani, H., & Imbens, G. (2022). Bayesian meta-prior learning using empirical Bayes. *Management Science, 68*(3), 1737–1755. <https://doi.org/10.1287/mnsc.2021.4136>
- National Research Council. (2012). *A Framework for K-12 science education: Practice, crosscutting concepts, and core ideas*. The National Academies Press.
- Neal, R. (2011, May 10). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (Vol. 20116022). Chapman and Hall/CRC. <https://doi.org/10.1201/b10905-6>
- NGSS Lead States. (2013). *Next Generation Science Standards: For States, by States*. The National Academies Press.
- Nitsch, C. (2013). *Dynamic Learning Maps: The Arc parent focus groups*. The Arc. <https://dynamiclearningmaps.org/sites/default/files/documents/publication/TheArcParentFocusGroups.pdf>
- Nocedal, J., & Wright, S. J. (2006). *Numerical optimization*. Springer. <https://doi.org/10.1007/978-0-387-40065-5>
- O’Leary, S., Lund, M., Ytre-Hauge, T. J., Holm, S. R., Naess, K., Dalland, L. N., & McPhail, S. M. (2014). Pitfalls in the use of kappa when interpreting agreement between multiple raters in reliability studies. *Physiotherapy, 100*, 27–35. <https://doi.org/10.1016/j.physio.2013.08.002>

- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann.  
<https://doi.org/10.1016/C2009-0-27609-4>
- Petrone, S., Rousseau, J., & Scricciolo, C. (2014). Bayes and empirical Bayes: Do they merge? *Biometrika*, *101*(2), 285–302. <https://doi.org/10.1093/biomet/ast067>
- Pontius, R. G., Jr., & Millones, M. (2011). Death to kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, *32*, 4407–4429. <https://doi.org/10.1080/01431161.2011.552923>
- Ravand, H., & Baghaei, P. (2020). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing*, *20*(1), 24–56.  
<https://doi.org/10.1080/15305058.2019.1588278>
- Rupp, A. A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.
- Sinharay, S., & Johnson, M. S. (2019). Measures of agreement: Reliability, classification accuracy, and classification consistency. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of Diagnostic Classification Models* (pp. 359–377). Springer International Publishing.  
[https://doi.org/10.1007/978-3-030-05584-4\\_17](https://doi.org/10.1007/978-3-030-05584-4_17)
- Stan Development Team. (2022). RStan: The R interface to Stan [R package version 2.21.5].  
<https://mc-stan.org/>
- Stefan, A. M., Evans, N. J., & Wagenmakers, E.-J. (2020). Practical challenges and methodological flexibility in prior elicitation. *Psychological Methods*, *27*(2), 177–197.  
<https://doi.org/10.1037/met0000354>
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370.  
<https://www.jstor.org/stable/1434855>
- Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*(2), 251–275. <https://doi.org/10.1007/s00357-013-9129-4>
- Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*(2), 317–339.  
<https://doi.org/10.1007/s11336-013-9362-0>
- Templin, J., & Henson, R. (2008, March 24–28). *Understanding the impact of skill acquisition: Relating diagnostic assessments to measurable outcomes* [Paper presentation]. American Educational Research Association Annual Meeting, New York, NY.
- Thompson, W. J. (2019). *Bayesian psychometrics for diagnostic assessments: A proof of concept* (Research Report No. 19-01). University of Kansas, Accessible Teaching, Learning, and Assessment Systems. <https://doi.org/10.35542/osf.io/jzqs8>
- Thompson, W. J. (2020). *Reliability for the Dynamic Learning Maps assessments: A comparison of methods* (Technical Report No. 20-03). University of Kansas; Accessible Teaching, Learning, and Assessment Systems. [https://dynamiclearningmaps.org/sites/default/files/documents/publication/Reliability\\_Comparison.pdf](https://dynamiclearningmaps.org/sites/default/files/documents/publication/Reliability_Comparison.pdf)
- Thompson, W. J., Clark, A. K., & Nash, B. (2019). Measuring the reliability of diagnostic mastery classifications at multiple levels of reporting. *Applied Measurement in Education*, *32*(4), 298–309.  
<https://doi.org/10.1080/08957347.2019.1660345>

- Thompson, W. J., & Nash, B. (2019, April 4–8). Empirical methods for evaluating maps: Illustrations and results. In M. Karvonen (Chair), *Beyond learning progressions: Maps as assessment architecture* [Symposium]. National Council on Measurement in Education Annual Meeting, Toronto, Canada. [https://dynamiclearningmaps.org/sites/default/files/documents/presentations/Thompson\\_Nash\\_Empirical\\_evaluation\\_of\\_learning\\_maps.pdf](https://dynamiclearningmaps.org/sites/default/files/documents/presentations/Thompson_Nash_Empirical_evaluation_of_learning_maps.pdf)
- Thompson, W. J., & Nash, B. (2022). A diagnostic framework for the empirical evaluation of learning maps. *Frontiers in Education, 6*, 714736. <https://doi.org/10.3389/feduc.2021.714736>
- Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved  $\hat{R}$  for assessing convergence of MCMC. *Bayesian Analysis, 16*(2), 667–718. <https://doi.org/10.1214/20-BA1221>
- Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement, 52*(4), 457–476. <https://doi.org/10.1111/jedm.12096>
- Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF* (Working Paper). University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science. Prince George, Canada.

## **A. Supplemental Information About Assessment Design and Development**

### **A.1. 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.

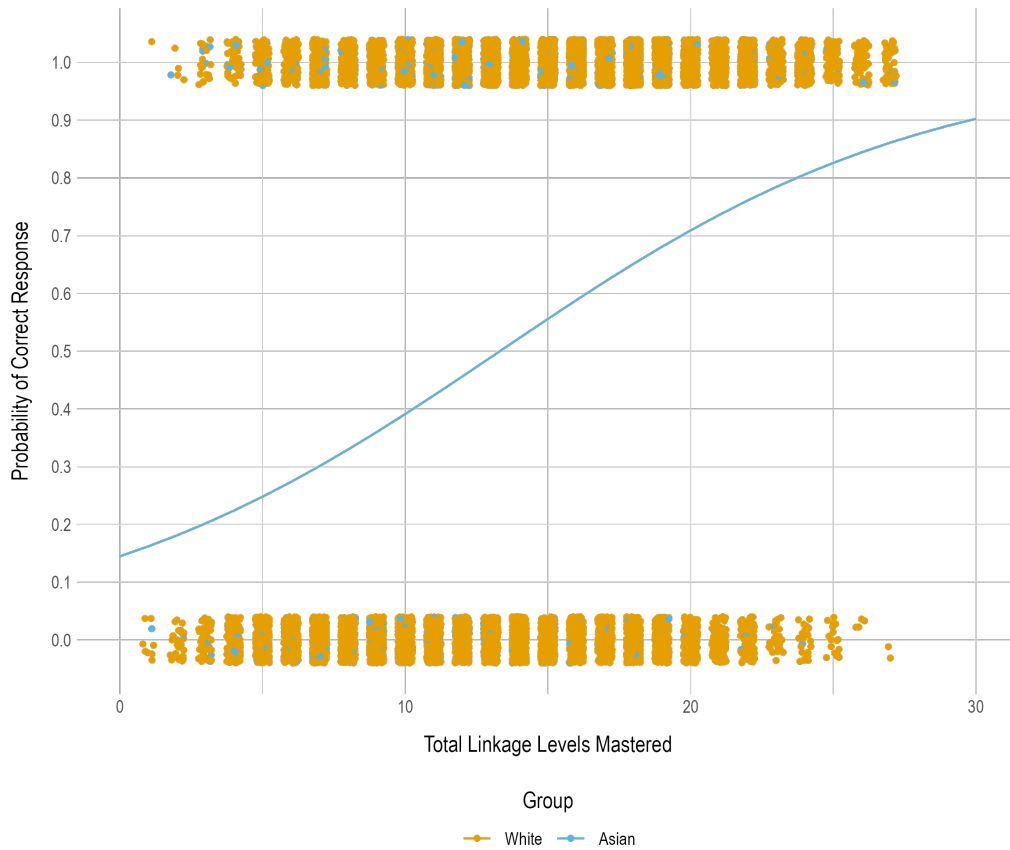
#### ***A.1.1. Uniform Model***

These plots show items that had a non-negligible effect-size change when comparing equation 3.2 to equation 3.1. In this model, the probability of a correct response was modeled as a function of ability and gender.



**Item 51571**

$\chi^2 = 15.02, p = 0.0001$ ; Nagelkerke's  $R^2 = 0.90$ , Zumbo & Thomas: *large*, Jodoin & Gierl: *large*



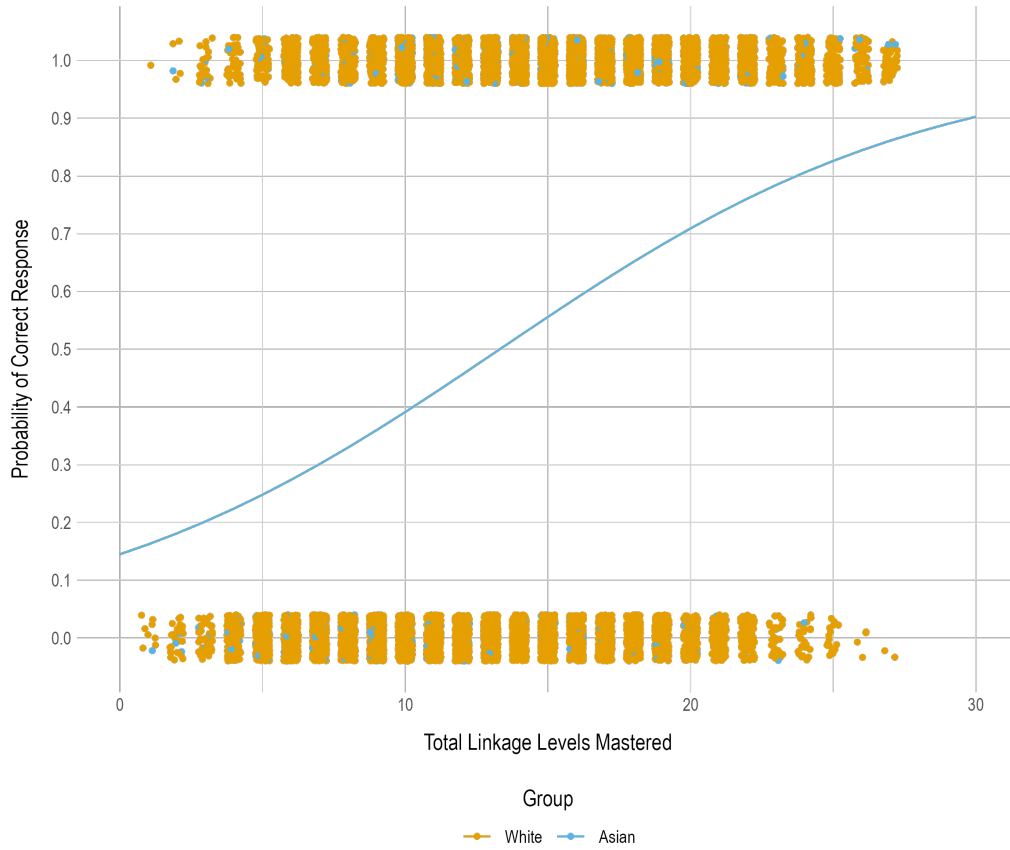
$n = 14,143$

**A.1.2. Combined Model**

These plots show items that had a non-negligible effect-size change when comparing equation 3.3 to equation 3.1. In this model, the probability of a correct response was modeled as a function of ability, gender, and their interaction.

**Item 51571**

$\chi^2 = 15.02, p = 0.0005$ ; Nagelkerke's  $R^2 = 0.90$ , Zumbo & Thomas: *large*, Jodoin & Gierl: *large*



$n = 14,143$