DYNAMIC®
LEARNING MAPS

2016–2017 Technical Manual

Science

January 2018

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

Acknowledgments

The publication of this technical manual represents the culmination of a body of work in the service of creating 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 manual possible. We are especially grateful for the contributions of Annie Davidson and to the members of the Dynamic Learning Maps® (DLM®) Technical Advisory Committee who graciously provided their expertise and feedback. Members of the Technical Advisory Committee include:

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

DLM project staff who made significant writing contributions to this technical manual are listed below with gratitude.

Brooke Nash, Ph.D., *Psychometrician Senior*
Amy Clark, Ph.D., *Psychometrician Senior*
William Jacob Thompson, *Psychometrician Assistant*
Meagan Karvonen, Ph.D., *Director*

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 are listed below with gratitude.

Lori Anderson
Sue Bechard
Brianna Beitling
Lisa Harkrader
Neal Kingston
Allison Lawrence
Lee Ann Mills
Michael Muenks
Haley Olson
Qianqian Pan
Michelle Shipman
Lindsay Rhuter
Russell Swinburne Romine
Susan K. Thomas
Xiaolin Wang
Lisa Weeks

Revision history

Date	Revision
4/18/2019	Corrected number of educators, schools, and school districts in VII.1. Student Participation

Table of Contents

I. INTRODUCTION 7
I.1. Background 7
I.2. Technical Manual Overview 7

II. ESSENTIAL ELEMENT DEVELOPMENT 10
II.1. Purpose of Essential Elements for Science 10

III. ITEM AND TEST DEVELOPMENT 12
III.1. Items and Testlets 12
III.1.A. Items 12
III.1.B. Item Writing 13
III.2. External Reviews 14
III.2.A. Review Recruitment, Assignments, and Training 14
III.2.B. Results of Reviews 15
III.2.C. Content-Team Decisions 15
III.3. Operational Assessment Items for 2016–2017 16
III.4. Field Testing 18
III.4.A. Description of Field Tests 19
III.4.B. Field-Test Results 19

IV. TEST ADMINISTRATION 24
IV.1. Overview of Key Administration Features 24
IV.1.A. Test Windows 24
IV.2. Implementation Evidence 24
IV.2.A. Adaptive Delivery 24
IV.2.B. Administration Incidents 28
IV.2.C. User Experience with Dynamic Learning Maps System 30
IV.3. Conclusion 36

V. MODELING 37
V.1. Overview of the Psychometric Model 37
V.2. Calibrated Parameters 38
V.2.A. Probability of Master Providing Correct Response 38
V.2.B. Probability of Non-Master Providing Correct Response 39
V.3. Mastery Assignment 42
V.4. Model Fit 43
V.4.A. Description of Methods 43
V.4.B. Background on Model Fit Calculation 44
V.4.C. Procedure for Evaluating Model Fit 49

V.4.D. Results 49
V.4.E. Operational Evaluation of Model Fit 53
V.4.F. Summary of Model Fit Analyses 56
V.5. Conclusion 57

VI. STANDARD SETTING 59

VII. ASSESSMENT RESULTS 60
VII.1. Student Participation 60
VII.2. Student Performance 62
VII.2.A. Overall Performance 63
VII.2.B. Subgroup Performance 64
VII.2.C. Linkage-Level Mastery 66
VII.3. Data Files 67
VII.4. Score Reports 67
VII.4.A. Individual Student Score Reports 67
VII.5. Quality Control Procedures for Data Files and Score Reports 70
VII.5.A. Manual Quality Control Checks 70

VIII. RELIABILITY 71
VIII.1. Background Information on Reliability Methods 71
VIII.2. Methods of Obtaining Reliability Evidence 71
VIII.2.A. Reliability Sampling Procedure 72
VIII.2.B. Reliability Evidence 73
VIII.3. Conclusion 84

IX. VALIDITY STUDIES 86
IX.1. Evidence Based on Test Content 86
IX.1.A. Opportunity to Learn 86
IX.2. Evidence Based on Response Processes 89
IX.2.A. Evaluation of Test Administration 89
IX.2.B. Test-Administration Observations 90
IX.3. Evidence Based on Internal Structure 94
IX.3.A. Evaluation of Item-Level Bias 95
IX.3.B. Internal Structure Across Linkage Levels 98
IX.4. Evidence Based on Consequences of Testing 99
IX.4.A. Teacher Perception of Assessment Contents 99
IX.4.B. Score-Report Interpretation Tutorial 100
IX.5. Conclusion 101

X. TRAINING AND INSTRUCTIONAL ACTIVITIES 103

XI. CONCLUSION AND DISCUSSION 104

XI.1. Validity Evidence Summary 105
XI.2. Continuous Improvement 107
XI.2.A. Operational Assessment 107
XI.2.B. Future Research 108

REFERENCES 110

Table of Tables

Table 1. Professional Roles of External Reviewers 14

Table 2. Population Density for Schools of External Reviewers 15

Table 3. Distribution of 2016–2017 Operational Science Testlets by Grade Band (*N* = 109) 16

Table 4. Distribution of 2016–2017 Science Field-Test Testlets by Grade Band (*N* = 82) 19

Table 5. Science Content Team Response to Item Flags for Each Grade Band (*N* = 16) 23

Table 6. Correspondence of Complexity Bands and Linkage Levels 25

Table 7. Adaptation of Linkage Levels Between First and Second Science Testlets by Grade Band (*N* = 19,686) 27

Table 8. Number of Students Affected by Each 2017 Incident 28

Table 9. Incident Summary for 2016–2017 Operational Testing, Science 29

Table 10. Teacher Response Regarding Test Administration 31

Table 11. Ease of Using KITE Client 32

Table 12. Ease of Using Educator Portal 33

Table 13. Overall Experience With KITE Client and Educator Portal 33

Table 14. Personal Needs and Preferences Profile Supports Selected for Students (*N* = 20,242) .. 35

Table 15. Teacher Reports of Student Accessibility Experience 36

Table 16. Univariate Contingency Table 46

Table 17. Bivariate Contingency Table 47

Table 18. Example Model-Level Univariate Fit 48

Table 19. Specification of *k*-Fold Estimation Procedure 49

Table 20. Average Number of Linkage Levels That Performed Better Over Five Validation Sets 50

Table 21. Average Number of Flagged Linkage Levels Using Limited Information Indices Over Five Validation Sets 52

Table 22. Student Participation by State (*N* = 19,686) 60

Table 23. Student Participation by Grade (*N* = 19,686) 61

Table 24. Demographic Characteristics of Participants 62

Table 25. Percentage of Students by Grade and Performance Level 64

Table 26. Students at Each Performance Level by Demographic Subgroup (*N* = 19,686) 64

Table 27. Percentage of Students Demonstrating Highest Linkage Level Mastered Across Essential Elements by Grade 66

Table 28. Summary of Performance-Level Reliability Evidence 75

Table 29. Summary of Content-Area Reliability Evidence by Grade 76

Table 30. Summary of Science-Domain Reliability Evidence by Grade 77

Table 31. Reliability Summaries Across All Essential Elements (EEs): Proportion of EEs Falling Within a Specified Index Range 79

Table 32. Example of True and Estimated Mastery Status From Reliability Simulation 80

Table 33. Reliability Summaries Across All Linkage Levels: Proportion of Linkage Levels Falling Within a Specified Index Range 82

Table 34. Teacher Ratings of Portion of Testlets That Matched Instruction 86

Table 35. Instruction Time Spent on Science Domains, in Hours 87

Table 36. Instruction Time Spent on Science and Engineering Practices, in Hours 88

Table 37. Correlation Between Instruction Time in Science Domain and Linkage Levels Mastered in That Domain 89

Table 38. Teacher Perceptions of Student Experience With Testlets 90

Table 39. Distribution of Teacher Observations by State (*N* = 32) 91

Table 40. Test Administrator Actions During Computer-Delivered Testlets (*n* = 22) 92

Table 41. Student Actions During Computer-Delivered Testlets (*n* = 22) 93

Table 42. Primary Response Mode for Teacher-Administered Testlets (*N* = 10) 94

Table 43. Items Flagged for Evidence of Uniform Differential Item Functioning by Grade Band 97

Table 44. Items Flagged for Evidence of Differential Item Functioning for the Combined Model by Grade Band 97

Table 45. Teacher Perceptions of Assessment Content 100

Table 46. Teacher Confidence in Ability to Interpret and Use Dynamic Learning Maps Score Reports Prior to Tutorial (*N* = 92) 101

Table 47. Review of Technical Manual Update Contents 104

Table 48. Dynamic Learning Maps Science Alternate Assessment System Propositions and Sources of Updated Evidence for 2016–2017 106

Table 49. Evidence Sources Cited in Previous Table 107

Table of Figures

Figure 1. Shown are *p* values for 2016–2017 operational science items. 17

Figure 2. Standardized difference *z* scores for 2016–2017 operational science items. 18

Figure 3. Shown are *p* values for 2016–2017 science items field-tested during spring window 20

Figure 4. Standardized difference *z* scores for 2016–2017 science items field-tested during spring window. 21

Figure 5. Probability of masters providing a correct response to items measuring each linkage level. 39

Figure 6. Probability of non-masters providing a correct response to items measuring each linkage level. 41

Figure 7. Linkage-level mastery assignment by mastery rule for each grade band. 43

Figure 8. Comparison of base-rate mastery probability in the fungible and nonfungible models. 54

Figure 9. Total linkage levels mastered comparison. 55

Figure 10. Change in percentage of students achieving at each performance level. 56

Figure 11. Page 1 of the performance profile for 2016–2017. 69

Figure 12. Simulation process for creating reliability evidence. 73

Figure 13. Number of linkage levels mastered within Essential Element reliability summaries. 79

Figure 14. Linkage-level reliability summaries. 83

Figure 15. Conditional reliability evidence summarized by linkage level. 84

I. INTRODUCTION

During the 2016–2017 academic year, the Dynamic Learning Maps® (DLM®) Alternate Assessment System offered assessments of student achievement in mathematics, English Language Arts (ELA), and science for students with the most significant cognitive disabilities in grades 3–8 and high school. Due to differences in the development timeline for science, separate technical manuals were prepared for ELA and mathematics (see Dynamic Learning Maps [DLM] Consortium, 2016b and DLM Consortium, 2016c).

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

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

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

I.1. Background

In 2016–2017, DLM science assessments were administered to students in nine states: Alaska, Illinois, Iowa, Kansas, Maryland, Missouri, Oklahoma, West Virginia, and Wisconsin.

In 2016–2017, the Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS) at the University of Kansas continued to partner with the Center for Literacy and Disability Studies at the University of North Carolina at Chapel Hill and the Center for Research Methods and Data Analysis at the University of Kansas. The project was also supported by a Technical Advisory Committee (TAC).

I.2. Technical Manual Overview

This manual provides evidence to support the DLM Consortium's assertion of technical quality and the validity of assessment claims.

Chapter I provides an overview of the assessment and administration for the 2016–2017 academic year and a summary of contents of the remaining chapters. While subsequent chapters describe the essential components of the assessment system separately, several key topics are addressed throughout this manual, including accessibility and validity.

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

Chapter III outlines procedural evidence related to test content. The chapter includes summaries of external reviews for content, bias, and accessibility. The final portion of the chapter describes the operational and field-test content available for 2016–2017.

Chapter IV provides an overview of the fundamental design elements that characterize test administration and how each element supports the DLM theory of action. The chapter provides updated information about administration incidents and evidence for spring routing in the system, as well as teacher-survey results collected during 2016–2017 regarding educator experience, administration of instructionally embedded assessments, and system accessibility.

Chapter V provides a summary of the psychometric model that underlies the DLM project and describes the process used to estimate item and student parameters from student test data. The chapter includes a summary of calibrated parameters, mastery assignment for students, and evidence of model fit. For a complete description of the modeling method, see Chapter V in the *2015–2016 Technical Manual – Science* (DLM Consortium, 2017b).

Chapter VI was not updated for 2016–2017. See Chapter VI in the *2015–2016 Technical Manual – Science* (DLM Consortium, 2017b) for a description of the methods, preparations, procedures, impact data, and results of the standard-setting meeting.

Chapter VII reports the 2016–2017 operational results, including student participation data. The chapter details the percentage of students at each performance level; subgroup performance by gender, race, ethnicity, and English learner status; and the percentage of students who showed mastery at each linkage level. Finally, the chapter provides descriptions of all types of score reports, data files, and quality control methods.

Chapter VIII focuses on reliability evidence, including a summary of the methods used to evaluate assessment reliability and results by performance level, content area, domain, Essential Element, linkage level, and conditional linkage level. For a complete description of the reliability background and methods, see Chapter VIII in the *2015–2016 Technical Manual – Science* (DLM Consortium, 2017b).

Chapter IX describes additional validation evidence not covered in previous chapters. The chapter provides study results for four of the five critical sources of evidence: test content, internal structure, response process, and consequences of testing. For evidence of relation to other variables, see Chapter IX in the *2015–2016 Technical Manual – Science* (DLM Consortium, 2017b).

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

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

II. 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 assessments 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 (2013; NGSS), see Chapter II of the *2015–2016 Technical Manual – Science* (Dynamic Learning Maps [DLM] Consortium, 2017b).

The 2015–2016 alternate assessments for science were based on the version of the EEs for science developed from the *Framework* and the NGSS during Phase 1 of the two-phase project. This approach addressed member states' need for the immediate creation of an end-of-year assessment in elementary, middle, and high school grade bands, as well as an end-of-course assessment in high school biology.

While additional work on the EEs is expected to occur after the development of a research-based learning map model that will also inform future decisions about the content and test design, the purpose of the *2015–2016 Technical Manual – Science* (DLM Consortium, 2017b) and this update is to provide evidence only for Phase 1 of the science project. Phase 2 of the science project will require its own separate accumulation and evaluation of evidence to support validity claims aligned to its theory of action.

II.1. Purpose of Essential Elements for Science

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

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

III. ITEM AND TEST DEVELOPMENT

Chapter III of the *2015–2016 Technical Manual – Science* (Dynamic Learning Maps® [DLM®] Consortium, 2017b) describes general item- and test-development procedures. This chapter provides an overview of updates to item and test development for the 2016–2017 academic year. The first portion of the chapter provides a summary of item and testlet information, followed by the 2016–2017 external reviews of items and testlets for content, bias, and accessibility. The next portion of the chapter describes the operational assessments for 2015–2016, followed by a section describing field tests administered in 2016–2017.

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

III.1. Items and Testlets

This section describes information pertaining to items and testlets administered as part of the DLM assessment system. For a complete summary of item- and testlet-development procedures, see Chapter III of the *2015–2016 Technical Manual – Science* (DLM Consortium, 2017b).

III.1.A. Items

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

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

- Arrange single-word answer options in alphabetical order.
- Arrange answer options that are phrases or sentences by logic (e.g., order as appears in a passage, stanza, or paragraph; order from key, chart, or table; chronological order; atomic number from periodic table; etc.), or, if no logical alternative is available, by length from shortest to longest.

- If following the arrangement guidelines results in consistently having the first option as the key (or the second or the third) for all items in a testlet, the order may be rearranged to avoid creating a pattern.

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

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

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

III.1.B. ITEM WRITING

Development of DLM items and testlets for science began in the winter of 2015. Additional items and testlets were developed during that summer. In 2015, item writing occurred during two events in which content and special education specialists worked on-site in Kansas City, Missouri, or Lawrence, Kansas, respectively, to develop DLM assessments. While each testlet developed in 2015 consisted of three items, item development continued in 2016 with the specific goal of developing five-item testlets. To this end, new testlets were created or items were added to the existing, nonoperational three-item testlets. A description of the development process and item writer characteristics is provided in this section. A description of the item writers from 2015 is provided in Chapter III of the *2015–2016 Technical Manual – Science* (DLM Consortium, 2017b).

The science test development team was responsible for writing the five-item testlets in 2016. The team consisted of three staff members: one graduate research assistant and two full-time staff members with combined experience teaching science, students with disabilities in elementary, middle, and high school grade levels. All three team members had at least master's-level degrees in education and significant experience writing, editing, or reviewing DLM items.

Using the pool of testlets developed during the July 2015 science item-writing workshop but not yet operational, the test-development team wrote two additional items for as many of the testlets as possible. In other words, additional items were written when the context of the testlet lent itself to additional ways of measuring the construct. When this was not an option, the team wrote new testlets. Testlets developed during 2016 were externally reviewed and subsequently field-tested in the spring of 2017. The results from the external review process are provided in the next section.

III.2. External Reviews

The purpose of external review is to evaluate items and testlets developed for the DLM Alternate Assessment System. Using specific criteria established for DLM assessments, reviewers recommended that the content be accepted, revised, or rejected. Feedback from external reviewers was used to make final decisions about assessment items before they were field-tested.

Overall, the process and review criteria for external review in 2016–2017 remained the same as those used in 2015–2016. Minor changes were made, including using fewer reviewers who completed more assignments.

Across all three content areas in the DLM assessment system (i.e., English language arts, mathematics, and science), the external review criteria and process appear to be a useful and effective review of content by outside panelists. Over the 5 years that the process has been implemented for the DLM system, modifications have been made to improve any noted difficulties, resulting in fewer field-test items being flagged for review each year across content areas.

III.2.A. Review Recruitment, Assignments, and Training

In 2016–2017, a volunteer survey was used to recruit external review panelists. Volunteers for the external review process completed the Qualtrics survey to capture demographic information as well as information about their education and experience. These data were then used to identify panel types (content, bias and sensitivity, and accessibility) for which the volunteers were eligible. A total of 37 people completed the required training, 12 of whom were placed on external review panels.

Of the 12 reviewers placed on panels, five completed reviews. Each reviewer was assigned to one of the three panel types. There were five reviewers: two on accessibility panels, two on content panels, and one on a bias and sensitivity panel. In addition, three power reviewers and two hourly reviewers examined all three panel types as needed for each content area.

The professional roles reported by the 2016–2017 reviewers are shown in Table 1. Reviewers who reported other roles included a specialized teacher and a supervisor of special education.

Table 1. Professional Roles of External Reviewers

Role	***n***	**%**
Classroom teacher	2	40.0
Instructional coach	1	20.0
Other	2	40.0

Reviewers had diverse experience teaching students with the most significant cognitive disabilities. Science reviewers had a median of 23 years of experience, with a minimum of 15 and a maximum of 27 years of experience.

All science reviewers were female and non-Hispanic/Latino. Most reviewers self-identified as Caucasian, although one reviewer reported that she was African American. Population density of schools in which reviewers taught or held a position is reported in Table 2. 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.

Table 2. Population Density for Schools of External Reviewers

Population density	***n***	**%**
Rural	1	20.0
Suburban	2	40.0
Urban	2	40.0

Review assignments were given throughout the year. Reviewers were notified by email each time they were assigned collections of testlets. Each review assignment required 1.5 to 2 hours to complete. In most cases, reviewers were given between 10 days and 2 weeks to complete an assignment.

III.2.B. Results of Reviews

Most of the content externally reviewed during the 2016–2017 academic year was included in the spring testing window. On a limited basis, reviewers examined content for the upcoming 2017–2018 school year. For science, 100% of items and testlets were rated as *accept* or *review*. No content was recommended for rejection. A summary of the content-team decisions and outcomes is provided here.

III.2.C. Content-Team Decisions

Because multiple reviewers examined each item and testlet, external review ratings were compiled across panel types, following the same process used in the previous 2 years. DLM content teams reviewed and summarized the recommendations provided by the external reviewers for each item and testlet. Based on the combined information, there were five decision options: (a) no pattern of similar concerns—accept as is, (b) pattern of minor concerns—will be addressed, (c) major revision needed, (d) reject, and (e) more information needed.

DLM content teams documented the decision category applied by external reviewers to each item and testlet. Following this process, content teams made a final decision to accept, revise, or reject each item and testlet. The science content team retained 100% of items and testlets sent out for external review. Of the items and testlets that were revised, most required only minor changes (e.g., minor rewording but concept remained unchanged), as opposed to major changes (e.g., stem or answer option replaced). The science team made 43 minor revisions to items and 31 minor revisions to testlets.

III.3. Operational Assessment Items for 2016–2017

Operational assessments were administered during the spring testing window. A total of 169,603 operational test sessions were administered; one test session is one testlet taken by one student. Only test sessions that were completed or in progress at the close of the testing window were included in the total number of test sessions.

Table 3 summarizes the total number of operational testlets for 2016–2017. A total of 109 operational testlets were available across the three grade bands; no science states in 2016–2017 participated in the end-of-instruction biology assessment. This total also included 27 braille testlets and one combination of EE and linkage level for which more than one testlet was available during an operational window due to having both a BVI and general version of the testlet available.

Table 3. Distribution of 2016–2017 Operational Science Testlets by Grade Band (*N* = 109)

Grade band	***n***
Elementary	36
Middle school	36
High school	37

Similar to 2015–2016, *p* values were calculated for all operational items to summarize information about item difficulty.

Figure 1 shows the *p* values for each operational item in science. To prevent items with a small sample size from skewing results, the student sample-size cutoff for inclusion in the *p* values plots was 20. The *p* values for most science items were between .5 and .8.

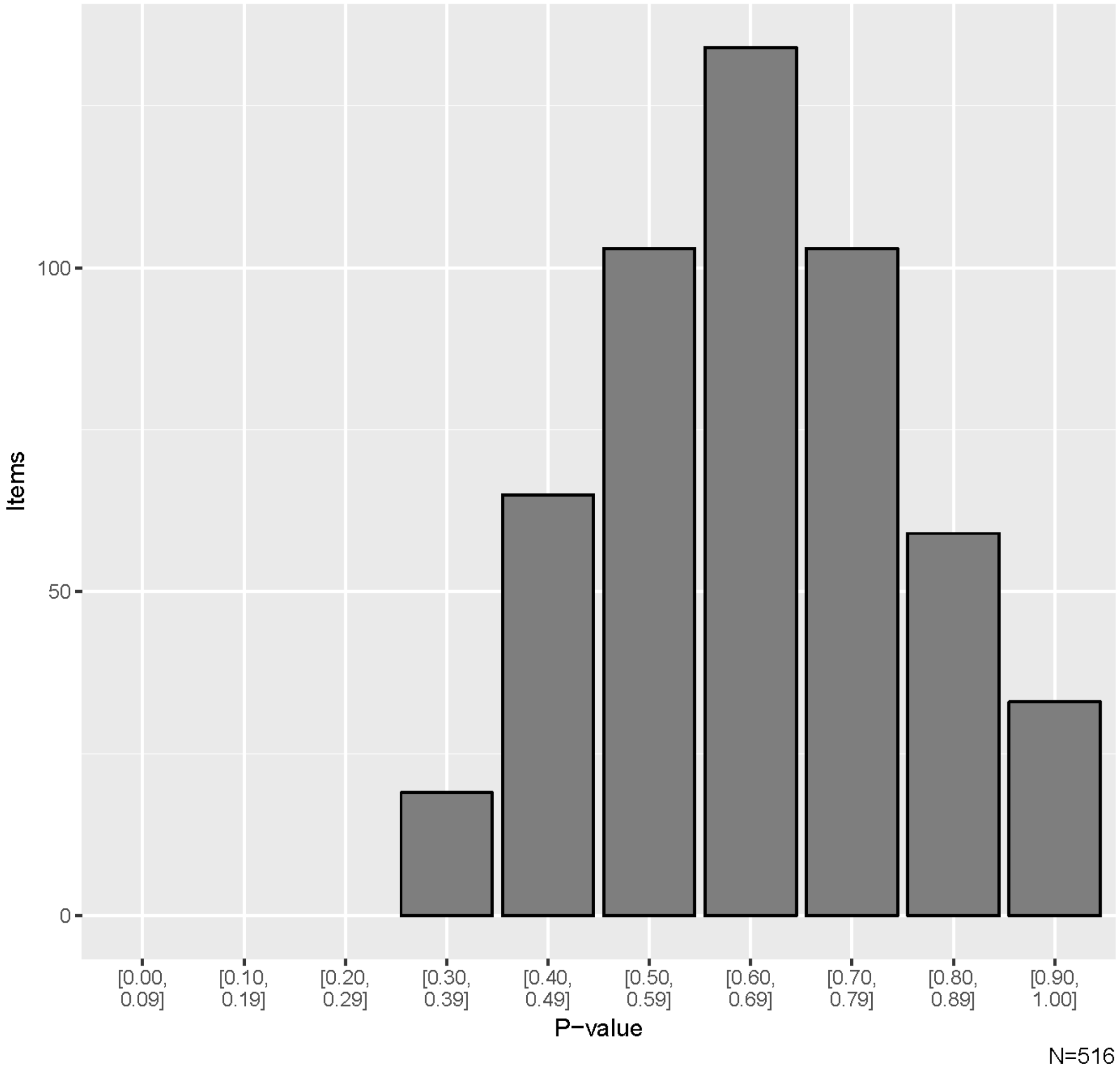


Figure 1. Shown are *p* values for 2016–2017 operational science items.
Note. Items with a sample size of less than 20 were omitted.

Standardized difference values were also calculated for all operational items with a student sample size of at least 20 to compare the *p* value for the item to the *p* values of all other items measuring the same EE and linkage-level combination. The standardized difference values provide one source of evidence of internal consistency. Figure 2 summarizes the standardized difference values for operational items. Most items fell within two standard deviations of the mean for the EE and linkage level. As additional data are collected and decisions are made regarding item-pool replenishment, item standardized difference values will be considered along with item-misfit analyses to determine which items and testlets are recommended for retirement.

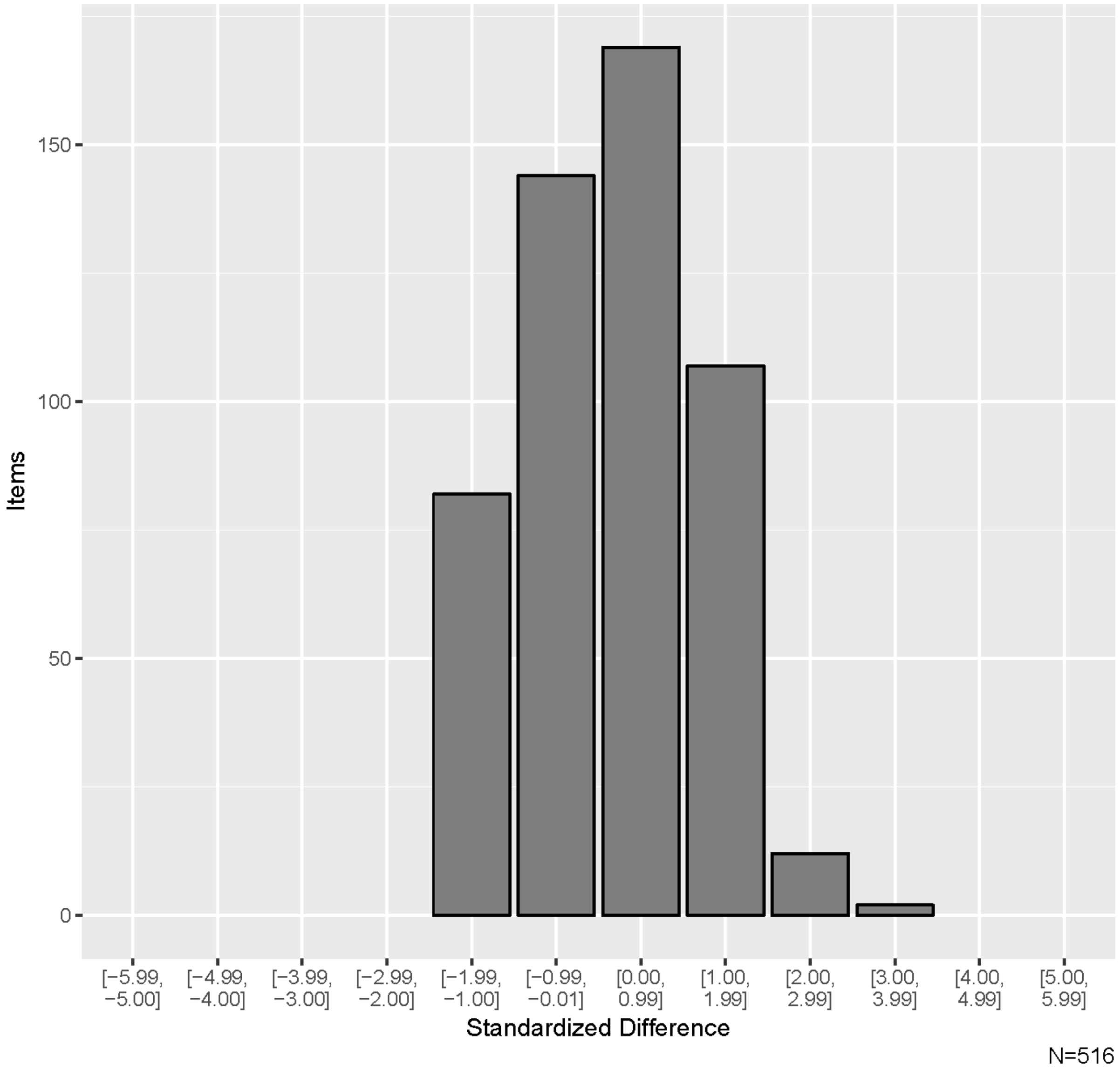


Figure 2. Standardized difference *z* scores for 2016–2017 operational science items.
Note. Items with a sample size of less than 20 were omitted.

III.4. Field Testing

During the 2016–2017 academic year, DLM field tests were administered to evaluate item quality for EEs assessed at each grade band for science, using the five-item testlets described earlier in this chapter. Field testing is conducted to deepen operational pools so that multiple testlets are available in spring windows. By deepening the operational pools, testlets can also be evaluated for retirement when other testlets perform better. A complete summary of prior pilot and field-test events can be found in the *Summary of the Dynamic Learning Maps Science Alternate Assessment Development Process* (Nash & Bechard, 2016).

III.4.A. DESCRIPTION OF FIELD TESTS

Collection of field-test data during the spring window in science was first implemented in the 2016–2017 academic year. During the spring administration, all students received one field-test testlet in science upon completion of all operational testlets.

The spring field-test administration was designed to both evaluate the new five-item testlets and collect data for each participating student at more than one linkage level for an EE to support future modeling development. (See Chapter V of this manual for more information.) As such, the field-test testlets were assigned at one linkage level below the last linkage level at which the student was assessed. Because of the process of assigning the testlet one linkage level lower than the last testlet, no Target-level testlets were field-tested during the spring window.

Testlets were made available for spring field testing in 2016–2017 for each section of the assessment. Table 4 summarizes the total number of field-test testlets by grade band for 2016–2017. A total of 82 field-test testlets were available.

Table 4. Distribution of 2016–2017 Science Field-Test Testlets by Grade Band (*N* = 82)

Grade band	***n***
Elementary	27
Middle school	28
High school	27

Participation in spring field testing was not required in any state, but teachers were encouraged to administer all available testlets to their students. The participation rate for science field testing in 2016–2017 was 72.1% (N= 14,200). The high participation rate allowed all testlets to meet sample-size requirements (i.e., responses from at least 20 students) and thus undergo statistical and content review before moving to the operational pool.

III.4.B. FIELD-TEST RESULTS

Data collected during each field test are compiled, and statistical flags are implemented ahead of content-team review. Flagging criteria serve as a source of evidence for content teams in evaluating item quality; however, final judgments are content based, taking into account the testlet as a whole.

III.4.B.i. Item Flagging

Criteria used for item flagging during previous field-test events were retained for 2016–2017. Items were flagged for review if they met any of the following statistical criteria:

- The item was too challenging, as indicated by a percentage correct (*p* value) below 35%. This value was selected as the threshold for flagging because most DLM items consist of three response options, so a value of less than 35% may indicate chance selection of the 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.

Reviewed items had a sample size of at least 20 cases. Figure 3 summarizes the *p* values for items field-tested during the 2016–2017 spring window. Most items fell above the 35% threshold for flagging. Test-development teams for each content area reviewed items below the threshold.

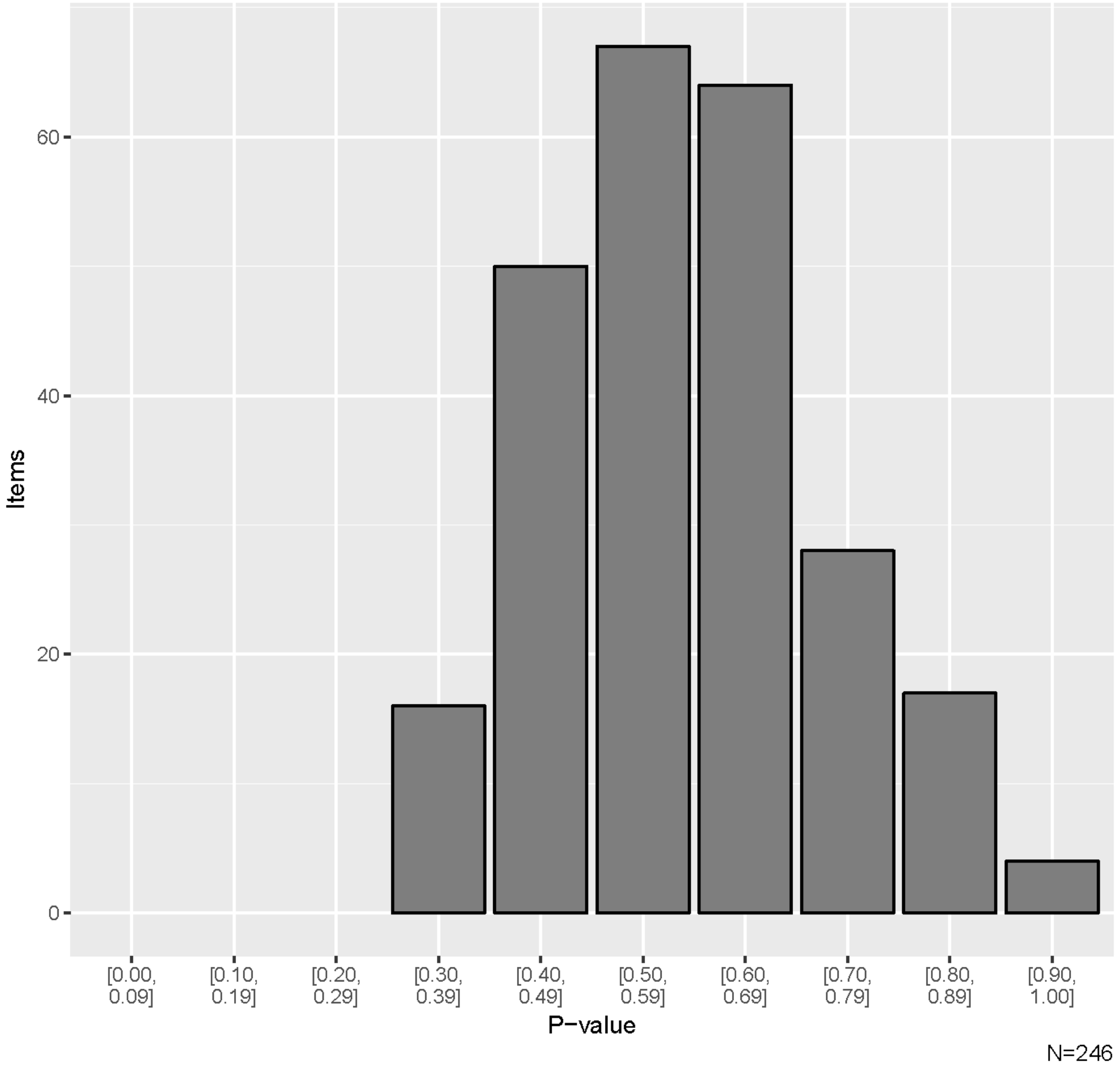


Figure 3. Shown are *p* values for 2016–2017 science items field-tested during spring window. *Note*. Items with a sample size of less than 20 were omitted.

Figure 4 summarizes the standardized difference values for items field-tested during the 2016–2017 spring window. Most items fell within two standard deviations of the mean for the EE and linkage level. Test-development teams for each content area reviewed items below the threshold.

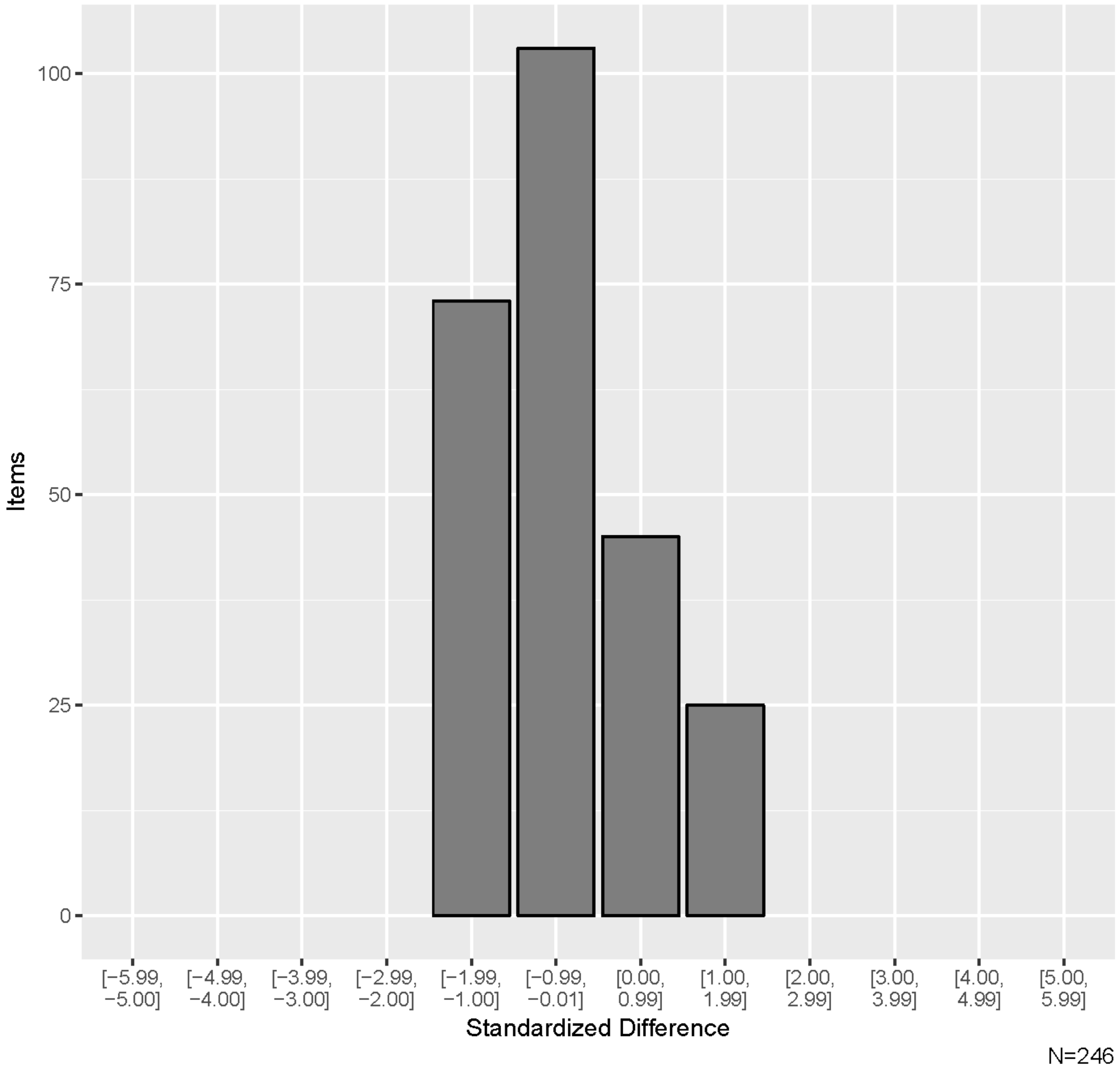


Figure 4. Standardized difference *z* scores for 2016–2017 science items field-tested during spring window.
Note. Items with a sample size of less than 20 were omitted.

III.4.B.ii. Item Data Review Decisions

Using the same procedures used in prior field-test windows, the test-development team made four types of item-level decisions as they reviewed 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. Reject 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 was aligned to the node. 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 incorrectly keyed (i.e., no correct answer or incorrect answer option was labeled as the correct option), items that were misaligned to the node, distractors that could be argued to be partially correct, or unnecessary complexity in the language of the stem.

After reviewing flagged items, reviewers looked at all items rated at 3 or 4 within the testlet to help determine whether the testlet would be retained or rejected. Here, the test-development team could elect to keep the testlet (with or without revision) or reject it. If an edit was to be made, it was assumed the testlet needed retesting. The entire testlet was rejected if the test-development team determined the flagged items could not be adequately revised.

III.4.B.iii. Results of Item Analysis and Content-Team Review

A total of 16 items were flagged due to their *p* values and/or standardized difference values. The test-development team 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.

Table 5 provides the test-development team's counts for acceptance, revision, and rejection for all field-test flagged items. No items were rejected or required revisions as a result of the review.

Table 5. Science Content Team Response to Item Flags for Each Grade Band (*N* = 16)

	Flagged	Accept		Revise		Reject	
Grade band	**item count**	***n***	**%**	***n***	**%**	***n***	**%**
Elementary	3	3	100.0	0	0.0	0	0.0
Middle school	5	5	100.0	0	0.0	0	0.0
High school	8	8	100.0	0	0.0	0	0.0

Decisions to recommend testlets for retirement occur on an annual basis following the completion of the operational testing year. When multiple testlets are available for an EE and linkage-level combination, test-development teams may recommend the retirement of testlets that perform poorly compared to others measuring the same EE and linkage level. The retirement process will begin after the 2016–2017 academic year and will be reported in the *2017–2018 Technical Manual Update – Science*.

IV. TEST ADMINISTRATION

Chapter IV of the *2015–2016 Technical Manual – Science* (Dynamic Learning Maps® [DLM®], 2017b) describes general test administration and monitoring procedures. This chapter describes procedures and data collected in 2016–2017, including a summary of adaptive routing, administration errors, Personal Needs and Preferences (PNP) profile selections, and teacher-survey responses regarding user experience and accessibility.

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

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

IV.1. Overview of Key Administration Features

This section describes updates to the key, overarching features of Dynamic Learning Maps (DLM) test administration for 2016–2017. For a complete description of key administration features, including information on assessment delivery, KITE® Client, and linkage-level selection, see Chapter IV of the *2015–2016 Technical Manual – Science* (DLM Consortium, 2017b). Additional information about administration can be found in the *Test Administration Manual 2016–2017* (DLM Consortium, 2016a) and the *Educator Portal User Guide* (Dynamic Learning Maps Consortium, 2017a).

IV.1.A. Test Windows

During the consortium-wide spring testing window, which occurred between March 15 and June 9, 2017, students were assessed on each Essential Element (EE) on the blueprint. Each state set its own testing window within the larger consortium spring window.

IV.2. Implementation Evidence

This section describes evidence collected for 2016–2017 during the operational implementation of the DLM Alternate Assessment System. The categories of evidence include data relating to the adaptive delivery of testlets in the spring window, administration incidents user experience, and accessibility.

IV.2.A. Adaptive Delivery

During the spring 2017 test administration, the science assessment was adaptive between testlets, following the same routing rules applied in 2015–2016. 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, skill, and ability 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. Table 6 shows the correspondence between the First Contact complexity bands and first assigned linkage levels.

Table 6. Correspondence of Complexity Bands and Linkage Levels

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

For a complete description of adaptive delivery procedures, see Chapter IV of the *2015–16 Technical Manual – Science* (DLM Consortium, 2017b).

Following the spring 2017 administration, analyses were conducted to determine the mean percentage of testlets that adapted up a linkage level, stayed at the same linkage level, or adapted down a linkage level from the first to second testlet administered for students within a grade and complexity band. The aggregated results can be seen in Table 7.

In comparison to 2015–2016, results were similar for students assigned to the Foundational complexity band. However, some differences were seen for students assigned to the other three bands. In particular, students who were assigned to Bands 1 and 2 adapted up a linkage level more frequently between their first and second testlets in comparison to 2015–2016. For students assigned to Band 2, the percentage of testlets that did not adapt was similar to 2015–2016 results; however, the percentage that adapted down a level decreased. Finally, while in both years the majority of students assigned to Band 3 did not adapt up or down linkage levels between first and second testlets, there was a slight increase in this trend in 2016–2017.

Overall, patterns seen for students assigned to the Foundational and Band 3 complexity bands are expected given the limited directions in which they can adapt. The shift in more testlets adapting up a linkage level for Band 1 and Band 2 students may be explained by several factors, including more opportunity for students to learn science content and interact with the assessment and more variability in student characteristics within this group.

Table 7. Adaptation of Linkage Levels Between First and Second Science Testlets by Grade Band (*N* = 19,686)

	Foundational[*]		Band 1[*]		Band 2			Band 3[*]	
Grade band	Adapted up (%)	Did not adapt (%)	Adapted up (%)	Did not adapt (%)	Adapted up (%)	Did not adapt (%)	Adapted down (%)	Did not adapt (%)	Adapted down (%)
3–5	35.9	64.1	73.9	26.1	40.9	40.0	19.1	69.6	30.4
6–8	26.7	73.3	57.6	42.4	51.5	30.9	17.6	67.9	32.1
9–12	27.1	72.9	53.7	46.3	37.4	38.0	24.6	80.1	19.9

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

IV.2.B. Administration Incidents

Monitoring of testlet assignment during the 2016–2017 operational assessment window uncovered two incidents that potentially affected a small number of students' experience with the science assessment. Table 8 provides a summary of the number of students affected by each incident, as delivered to states in the Incident File (see Chapter VII of this manual for more information). Following delivery of the Incident File on the predetermined timeline, a script was created to identify students who were actually affected by each incident, narrowing from the list of those potentially affected.[1] These values are also reported in Table 8. This script will be modified such that the 2018 Incident File and beyond reports only students actually affected by the incident rather than students who may have been affected.

The most frequent incident was potential incorrect scoring and misrouting caused by a system database load issue. While there was no evidence that the database did not record student responses as intended, this incident was reported out of an abundance of caution in the unlikely event student responses were stored as skips by the system during the load issue period. For Incident Code 5 (i.e., misrouting due to testlet re-administration after student transfer), the impacted state was provided information about the affected student and given the option to revert the student's assessment back to the end of the last correctly completed testlet (i.e., the point at which routing failed) and have the student complete the remaining testlets as intended. Additional details about the two incidents are provided in Table 9. Overall, the administration incidents affected less than 0.001% of students testing in science.

Table 8. Number of Students Affected by Each 2017 Incident

Incident code	Incident description	Potential effect as reported in Incident File		Actual effect	
		n	%	*n*	%
4	Potential incorrect scoring and misrouting due to KITE database load issue.	16	< 0.001	0*	0.00*
5	Misrouting due to testlet re-administration after student transfer.	1	< 0.001	1	< 0.001

Note. Incident codes 1–3 did not affect the science assessment.
*Estimated actual effect. There is no evidence that the database did not record student responses.

[1] Due to the type of incidents that affected the science assessment, the number of students reported as potentially affected was the number that were actually affected.

Table 9. Incident Summary for 2016–2017 Operational Testing, Science

Incident no.	Issue	Type	Summary
4	Potential incorrect scoring and misrouting due to KITE database load issue	Technology: Capacity	An integration server used by the test-delivery application experienced server load issues April 4, 2017 8:50 a.m.–2:05 p.m. CST and April 5, 2017, 8:45–10:55 a.m. CST. As a matter of regular practice, if the database times out before a student's response is submitted, the system starts the student at the beginning of the testlet the next time the testlet is opened. There is no evidence that the database did not record all student responses during the two impacted periods. Because items may be intentionally skipped as a matter of practice or student choice, and because there is no evidence that the database did not record student responses, it is assumed that all responses were recorded by the database as intended. However, out of an abundance of caution, testlets with one or more missing responses submitted during the two time periods were identified and provided to states for review. States were given the option to revert students to the end of the previously submitted testlet and resume testing, or to let students proceed forward as usual.
5	Misrouting due to testlet re-administration after student transfer	Technology: Administration	The student transferred to a different school, district, and/or teacher, and the system reassigned a previously taken testlet. During the second administration, the student provided different responses, resulting in a different percent correct for routing purposes.

As in 2015–2016, the Incident File was delivered to state partners with the General Research File (GRF; see Chapter VII of this manual for more information), providing a list of all students potentially affected by each issue. States could use the Incident File and their own accountability policies and practices to determine possible invalidation of student records. All issues were corrected for subsequent administration. Testlet assignment will continue to be monitored in subsequent years to track any incidents and report them to state partners.

IV.2.C. USER EXPERIENCE WITH DYNAMIC LEARNING MAPS SYSTEM

User experience with the system was evaluated through a spring 2017 survey disseminated to teachers who had administered a DLM assessment during the spring window. In 2017, the survey was distributed to teachers via KITE Client, the platform students use to complete assessments. Each student was assigned a teacher survey for their teacher to complete. The survey included three sections. The first and third sections were fixed, while the second section was spiraled, with teachers responding to subsets pertaining to accessibility, Educator Portal and KITE Client feedback, the relationship of assessment content to instruction, and teacher experience with the system.

A total of 6,619 teachers from states participating in the science assessment responded to the survey (response rate of 84.4%) for 14,991 students. This reflects a substantial increase in the rate of responding teachers compared to those observed during previous delivery of surveys in Qualtrics (e.g., 2016 response rate was 11.5%). Because of the difference in response rates over years and changes to the structure and content of the survey, the spring 2017 administration is treated as baseline data collection. Comparisons of data collected from 2016 are not included in this manual.

Participating teachers responded to surveys for between one and 20 students. Teachers most frequently reported having 0 to 5 years of experience in science and in teaching students with the most significant cognitive disabilities. The median number of years of experience in each of these areas was 6 to 10. Approximately 52% indicated they had experience administering the DLM assessment in all three operational years.

The sections that follow summarize user experience with the system and accessibility. Additional survey results are summarized in Chapter IX of this manual. For responses to the 2015–2016 teacher survey, see Chapter IV and Chapter IX in the *2015–2016 Technical Manual – Science* (DLM Consortium, 2017b).

IV.2.C.i. Educator Experience

Respondents were asked to reflect on their experiences with the assessments and their comfort level and knowledge in administering them. Most questions required respondents to use a 4-point scale: *strongly disagree*, *disagree*, *agree*, or *strongly agree*. Responses are summarized in Table 10.

Teachers responded that they were confident administering DLM testlets (96.0% agreed or strongly agreed). Respondents believed that the required test administrator training prepared them for their responsibilities as test administrators (87.2% agreed or strongly agreed). Moreover, most teachers reported that manuals and the Educator Resources page on the DLM webpage helped them understand how to use the system (88.1%); that they knew how to use accessibility supports, allowable supports, and options for flexibility (92.6%); and that the Testlet Information Pages helped them deliver the testlets (87.6%).

Table 10. Teacher Response Regarding Test Administration

	SD		D		A		SA		A+SA	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Confidence in ability to deliver DLM testlets.	26	1.2	61	2.8	1,007	45.8	1,104	50.2	2,111	96.0
Test administrator training prepared respondent for responsibilities of test administrator.	73	3.3	207	9.5	1,183	54.2	720	33.0	1,903	87.2
Manuals and DLM Educator Resource Page materials helped respondent understand how to use assessment system.	44	2.0	215	9.8	1,273	58.3	653	29.9	1,926	88.1
Respondent knew how to use accessibility features, allowable supports, and options for flexibility.	32	1.5	129	5.9	1,287	58.8	739	33.8	2,026	92.6
Testlet Information Pages helped respondent prepare to deliver the testlets.	57	2.6	215	9.8	1,255	57.4	660	30.2	1,915	87.6

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

IV.2.C.ii. KITE System

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

KITE Client is used for the administration of DLM testlets. Teachers were asked to rate their experiences with KITE Client and to evaluate the ease of each step using a 5-point scale: *very hard*, *somewhat hard*, *neither hard nor easy*, *somewhat easy*, or *very easy*. Table 11 summarizes teacher responses.

Respondents found it either somewhat easy or very easy to enter the site (74.4%), navigate within a testlet (77.5%), submit a completed testlet (83.3%), record a response (83.2%) and administer testlets on various devices (69.7%). Open-ended survey-response feedback indicated that testlets were easy to administer and that technology had improved compared to previous years.

Table 11. Ease of Using KITE Client

	VH		SH		N		SE		VE		SE+VE	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Enter the site	57	2.5	177	7.8	349	15.3	680	29.8	1,018	44.6	1,698	74.4
Navigate within a testlet	45	2.0	150	6.6	318	13.9	689	30.2	1,079	47.3	1,768	77.5
Record a response	24	1.1	72	3.2	283	12.5	620	27.3	1,272	56.0	1,892	83.3
Submit a completed testlet	31	1.4	79	3.5	270	11.9	582	25.7	1,306	57.6	1,888	83.2
Administer testlets on various devices	55	2.4	135	6.0	496	21.9	656	28.9	924	40.8	1,580	69.7

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

Educator Portal is the software used to store and manage student data and to enter PNP and First Contact information. Teachers were asked to assess the ease of navigating and using Educator Portal for its intended purposes, using the same scale used regarding KITE Client; these data are summarized in Table 12. Overall, respondents' feedback improved from the previous year but was still somewhat mixed. The majority of teachers found it somewhat easy or very easy to navigate the site (56.6%), enter PNP and First Contact information (65.7%), manage student data (56.6%), manage their own accounts (59.9%), and manage tests (54.9%).

Table 12. Ease of Using Educator Portal

	VH		SH		N		SE		VE		SE+VE	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Navigate the site	135	5.9	420	18.4	434	19.0	729	32.0	562	24.6	1,291	56.6
Enter PNP/Access profile and First Contact information	67	2.9	273	12.0	441	19.4	870	38.2	628	27.6	1,498	65.7
Manage student data	97	4.3	367	16.2	520	22.9	793	35.0	491	21.6	1,284	56.6
Manage my account	79	3.5	307	13.5	527	23.2	844	37.1	518	22.8	1,362	59.9
Manage tests	130	5.7	418	18.4	477	21.0	731	32.1	518	22.8	1,249	54.9

Note. VH = very hard; SH = somewhat hard; N = neither hard nor easy; SE = somewhat easy; VE = very easy; SE+VE = somewhat easy and very easy; PNP = Personal Needs and Preferences Profile.

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

Finally, respondents were asked to rate their overall experience with KITE Client and Educator Portal on a 4-point scale: *poor*, *fair*, *good*, and *excellent*. Results are summarized in Table 13. The majority of respondents reported a positive experience with KITE Client. Nearly 75.9% of respondents said their experience was good or excellent, while 65.0% reported their overall experience with Educator Portal was good or excellent.

Overall feedback from teachers indicated that KITE Client was easy to navigate and user-friendly. Additionally, teachers provided useful feedback for improvements to Educator Portal that will be considered for subsequent technology development to improve user experience for 2017–2018 and beyond.

Table 13. Overall Experience With KITE Client and Educator Portal

	Poor		Fair		Good		Excellent	
Interface	*n*	%	*n*	%	*n*	%	*n*	%
KITE Client	125	5.5	426	18.7	1,105	48.4	627	27.5
Educator Portal	223	9.8	593	25.9	1,082	47.3	389	17.0

IV.2.C.iii. Accessibility

Accessibility supports provided in 2016–2017 were the same as those available in 2015–2016. Accessibility guidance provided by the DLM system distinguishes between accessibility supports that can be used by selecting online features via the PNP, require additional tools or materials, and are provided by the test administrator outside the system. Table 14 shows selection rates for three categories of accessibility supports, sorted by rate of use within each category. For a complete description of available accessibility supports, see Chapter IV of the *2015–2016 Technical Manual – Science* (DLM Consortium, 2017b). Generally, the percentage of students for whom supports were selected in 2016–2017 was similar to that observed in 2015–2016.

Table 14. Personal Needs and Preferences Profile Supports Selected for Students (*N* = 20,242)

Supports	*n*	%
Supports provided in KITE Client via Access Profile		
Spoken audio	5,453	26.94
Magnification	3,233	15.97
Color contrast	2,821	13.94
Overlay color	2,527	12.48
Invert color choice	2,249	11.11
Supports requiring additional tools/materials		
Individualized manipulatives	8,408	41.54
Calculator	6,867	33.92
Single-switch system	2,704	13.36
Alternate form – visual impairment	2,126	10.50
Two-switch system	1,885	9.31
Uncontracted braille	1,684	8.32
Supports provided outside the system		
Human read aloud	18,212	89.97
Test administration enters responses for students	10,238	50.58
Partner-assisted scanning	3,036	15.00
Sign interpretation of text	1,946	9.61
Language translation of text	1,905	9.41

Table 15 summarizes teacher responses to survey items about the accessibility supports used during administration. Teachers were asked to respond to two items using a 4-point Likert-type scale (*strongly disagree*, *disagree*, *agree*, or *strongly agree*) or indicate if the item did not apply to the student. The majority of teachers agreed that students were able to effectively use accessibility supports (81.9%) and that accessibility supports were similar to ones the student used for instruction (83.4%). These data support the conclusions that the accessibility supports of the DLM alternate assessment were effectively used by students, emulated accessibility supports used during instruction, and met student needs for test administration. Additional data will be collected during the spring 2018 survey to determine whether results improve over time.

Table 15. Teacher Reports of Student Accessibility Experience

	SD		D		A		SA		A+SA		N/A	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Student was able to effectively use accessibility features.	66	2.8	95	4.0	1,039	43.6	912	38.3	1,951	81.9	269	11.3
Accessibility features were similar to ones student uses for instruction.	63	2.7	97	4.1	1,010	42.5	970	40.9	1,980	83.4	234	9.9

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

IV.3. Conclusion

During the 2016–2017 academic year, the DLM system was available for optional instructionally embedded use and during the operational spring window. Implementation evidence was collected in the forms of testlet adaptation analyses, a summary of students affected by incidents during operational testing, and teacher-survey responses regarding user experience and accessibility. Results indicated that teachers felt confident administering testlets in the system and found KITE Client easy to use but thought Educator Portal posed more challenges. Given the substantial improvement in teacher response rates, the teacher survey will continue to be distributed in KITE Client in 2017–2018 and beyond.

V. MODELING

Chapter V of the *2015–2016 Technical Manual Update – Science* (Dynamic Learning Maps® [DLM®] Consortium, 2017b) describes the psychometric model that underlies the Dynamic Learning Maps (DLM) assessment system and the process used to estimate item and student parameters from student assessment data. This chapter provides a high-level summary of the model used to calibrate and score assessments, along with a summary of updated modeling evidence from the 2016–2017 administration year. Additional evidence provided includes a description of model-fit analyses and results.

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

V.1. OVERVIEW OF THE PSYCHOMETRIC MODEL

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

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

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

Following calibration, results for each linkage level were combined to determine the highest linkage level mastered for each EE. Although the connections between linkage levels were not modeled empirically, they were used in the scoring procedures. In particular, if the latent class analysis determined a student had mastered a given linkage level within an EE, then the student was assumed to have mastered all lower levels within that EE.

In addition to the calculated posterior probability of mastery, students could demonstrate mastery of each EE in two additional ways: by correctly answering 80% of all items administered at the linkage level or by the two-down scoring rule which provides mastery status at two linkage levels down from a tested level that was not mastered. The two-down scoring rule was implemented to guard against students assessed at the highest linkage level being excessively penalized for incorrect responses.

V.2. Calibrated Parameters

As stated in the previous section, for diagnostic assessments, the comparable item parameters are conditional probabilities of providing a correct response to the item. Because of the assumption of fungibility, parameters are calculated for each of the 102 linkage levels in science.[2] Parameters include a conditional probability of providing a correct response for both non-masters and masters. Across all linkage levels, it is generally expected that the conditional probability of providing a correct response will be high for masters and low for non-masters. A summary of the operational parameters used to score the 2016–2017 assessment is provided in the following sections.

V.2.A. Probability of Master 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. Figure 5 depicts the conditional probability of masters providing a correct response to items measuring each of the 102 linkage levels based on the spring 2017 calibration. Because the point of maximum uncertainty is .5, masters should have a greater than .5 chance of providing a correct response. The results in Figure 5 demonstrate that all linkage levels performed as expected.

[2] The total of 102 includes all EEs and linkage levels measured by the science assessment. While there no states participated in the end-of-instruction biology assessment in spring 2017, data were accumulated across operational years for calibration and are therefore included in this chapter.

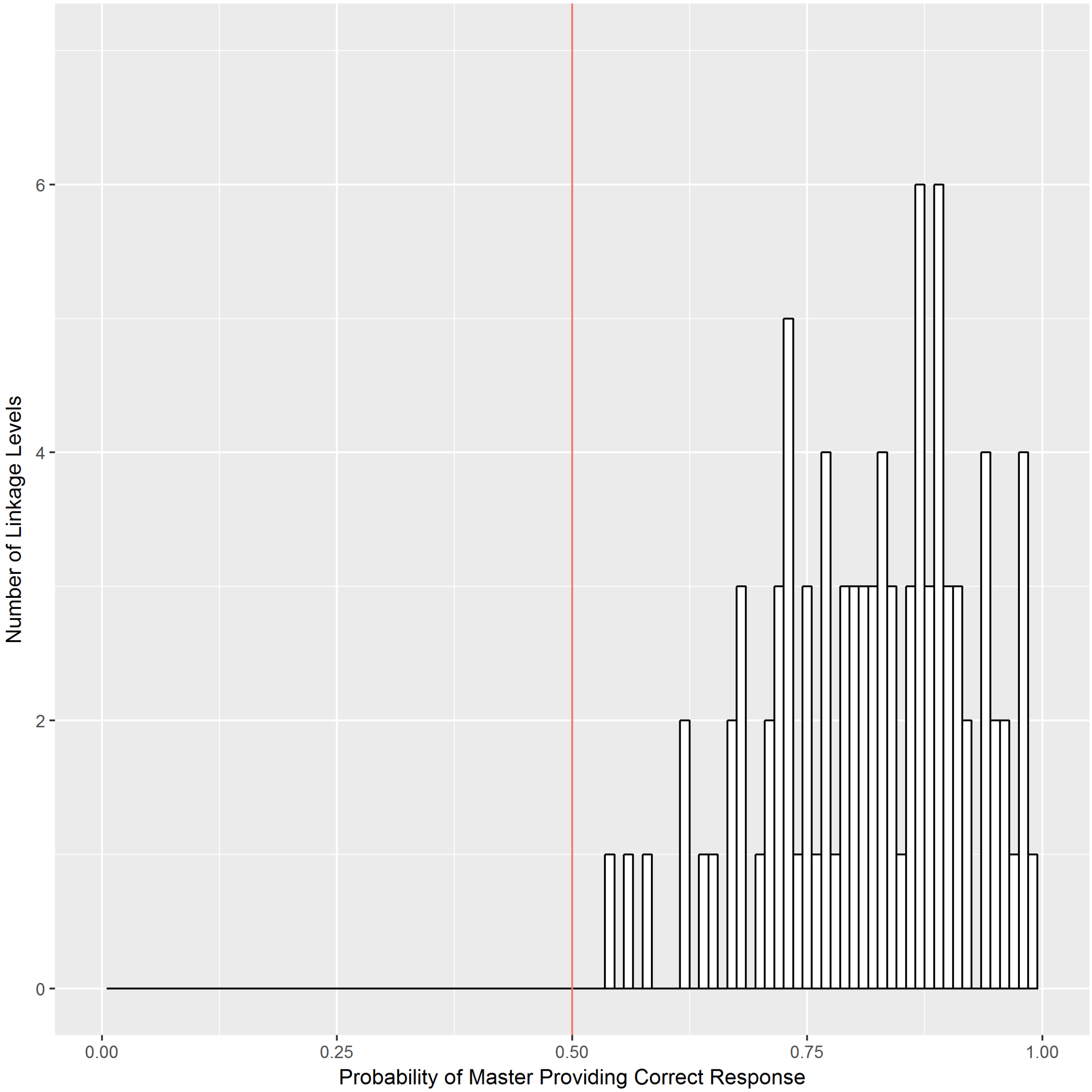


Figure 5. Probability of masters providing a correct response to items measuring each linkage level.
Note. Histogram bins are in increments of .01. Reference line indicates .5.

V.2.B. Probability of Non-Master Providing Correct Response

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

Figure 6 summarizes the probability of non-masters providing correct responses to items measuring each of the 102 linkage levels. There is greater variation in the probability of non-masters providing a correct response to items measuring each linkage level than was observed for masters; the histogram in Figure 6 indicates that non-masters sometimes have a greater than chance (>.5) likelihood of providing a correct response to items measuring the linkage level. This may indicate the items (and linkage level as a whole, since the item parameters are shared) are easily guessed or do not discriminate as well between the two groups of students.

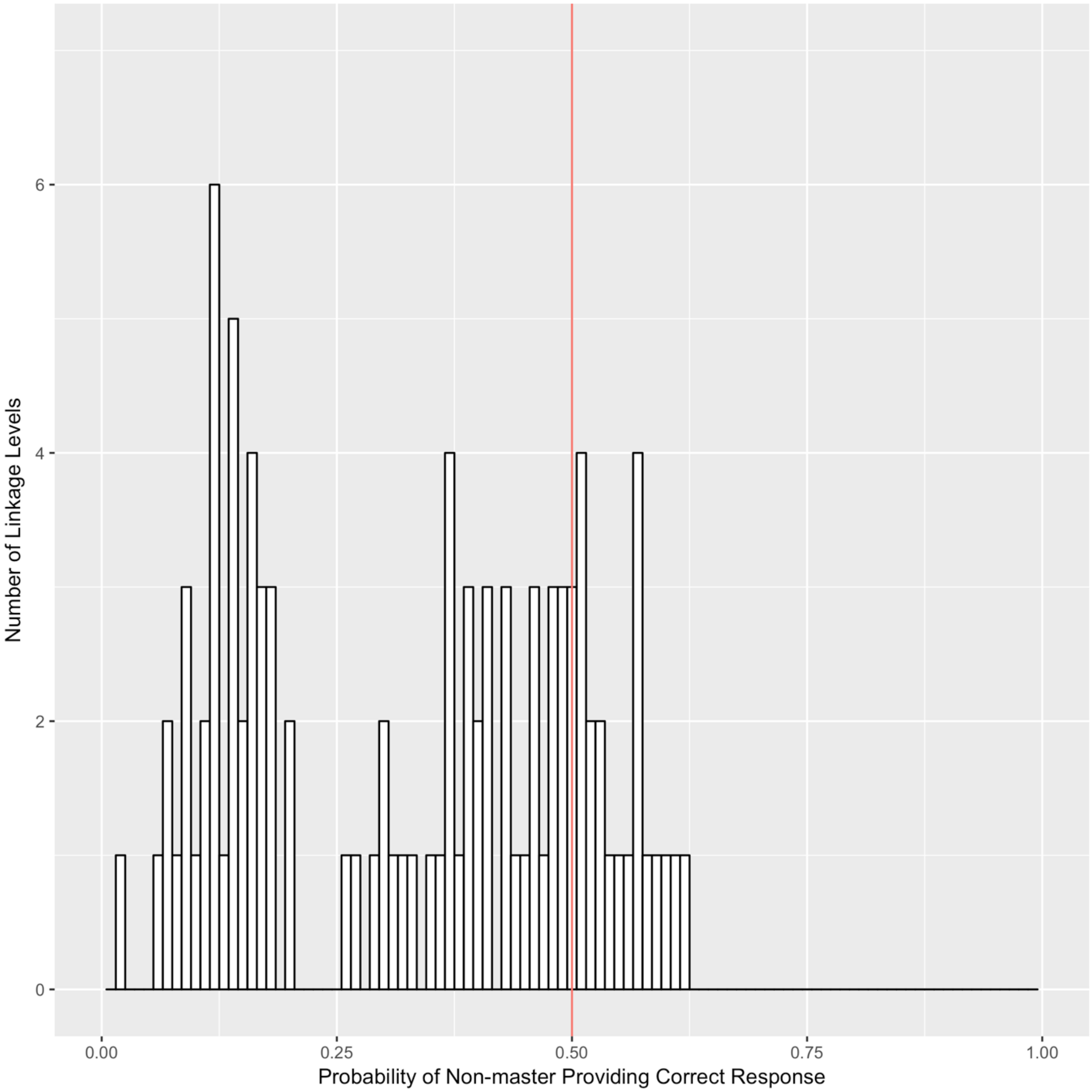


Figure 6. Probability of non-masters providing a correct response to items measuring each linkage level.

Note. Histogram bin size is in increments of 0.01. Reference line indicates .5.

V.3. Mastery Assignment

As mentioned, in addition to the calculated posterior probability of mastery, students could to demonstrate mastery of each EE in two additional ways: by correctly responding to 80% of all items administered at the linkage level or by the two-down scoring rule. To evaluate the degree to which each mastery assignment rule contributed to students' linkage-level mastery status during the 2016–2017 administration of DLM assessments, the percentage of both mastery statuses obtained by each scoring rule was calculated, as shown in Figure 7. Posterior probability was given first priority. If mastery was not demonstrated by meeting the posterior probability threshold, the other two scoring rules were imposed. Approximately 80% of mastered linkage levels were derived from the posterior probability obtained from the modeling procedure. The other linkage levels (approximately 10% to 15%) were assigned mastery status by the minimum mastery, or two-down rule, and the remaining percentages at each grade were determined by the percentage-correct rule. These results indicate that the percentage-correct rule likely had strong overlap (but was second in priority) with the posterior probabilities, in that correct responses to all items measuring the linkage level were likely necessary to achieve a posterior probability above the .8 threshold. The percentage-correct rule does, however, provide mastery status when providing correct responses to all or most items still resulted in a posterior probability below the mastery threshold.

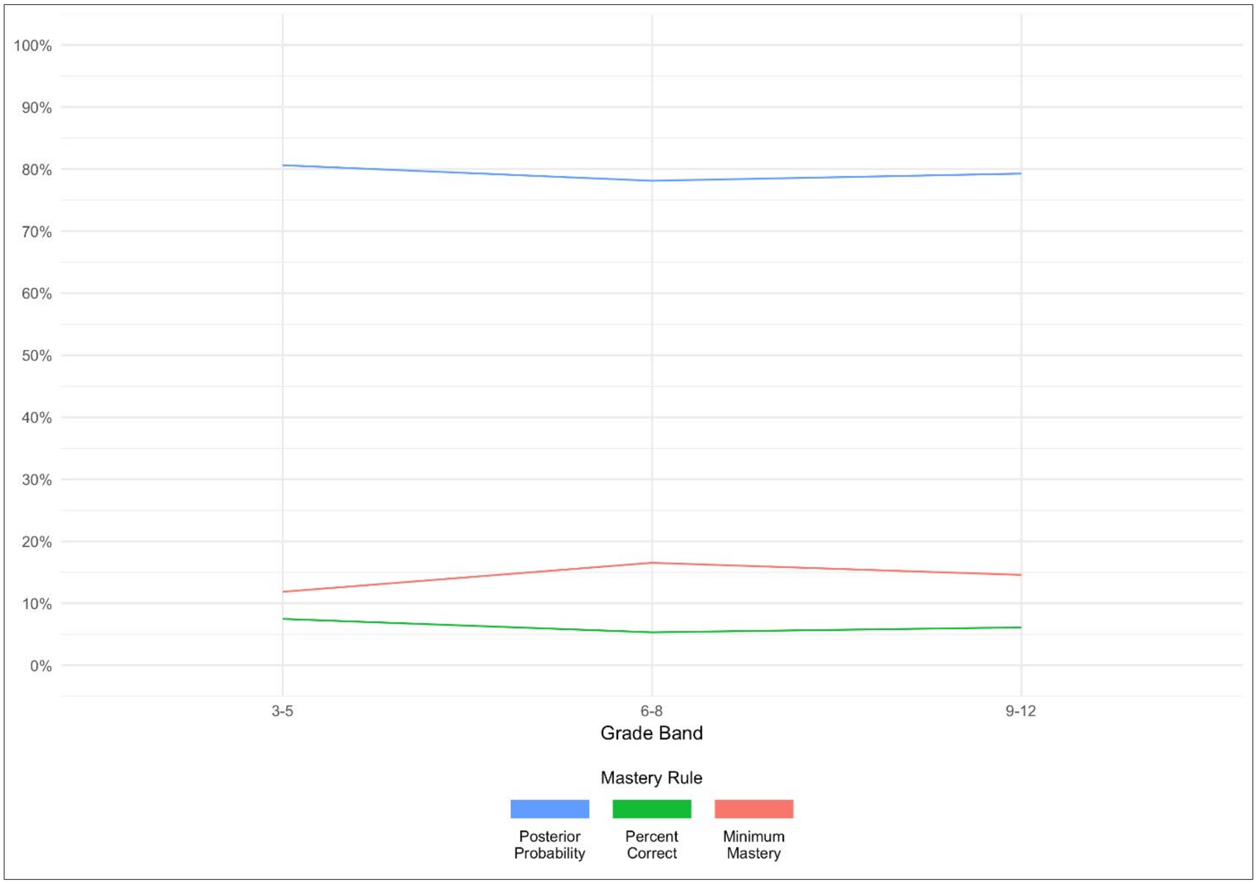


Figure 7. Linkage-level mastery assignment by mastery rule for each grade band.

V.4. Model Fit

Model fit has important implications for the validity of inferences that can be made from assessment results. If the model used to calibrate and score the assessment does not fit the data well, results from the assessment may not accurately reflect what students know and can do. Because one of the assumptions of the DLM assessment system is that items measuring the same linkage level are fungible, or exchangeable, evidence of the degree to which a fungible model fits the data must be evaluated. In addition, the fit of the fungible model should be compared to that of a nonfungible model to evaluate the models' relative fit.

The following sections provide a detailed description of the methodology used to evaluate model fit, using both relative and absolute indices. Results are summarized for the 102 linkage levels and 34 EEs measured by the assessment for science.

V.4.A. Description of Methods

To evaluate model fit for DLM assessments, two models were fit to each linkage level: a fungible and a nonfungible model. Definitions of each model follow, where π_{ij} is the probability of a respondent in class *j* providing a correct response to item *i*, η_j is the base-rate probability of class *j*, and respondents are subscripted as $h = \{1,2,3,...N\}$, items as $i = \{1,2,3,...I\}$, and classes as $j = \{1,2,...J\}$.

- ***Fungible Model***. In the fungible model, the conditional probabilities for non-masters and masters were held constant for all items measuring the same linkage level.

$$f(\mathbf{x}_h) = \sum_{j=0}^{J} \eta_j \prod_{i=1}^{I} \pi_j^{x_{ih}} (1 - \pi_j)^{1-x_{ih}} \qquad (1)$$

  In Equation 1, the probability of a correct response for a respondent in class j is denoted as π_j rather than π_{ij}, indicating that π is constant across items for all members of class j.

- ***Nonfungible Model***. In the nonfungible model, the conditional probabilities for non-masters and masters were allowed to vary across all items and linkage levels.

$$f(\mathbf{x}_h) = \sum_{j\,=\,0}^{J} \eta_j \prod_{i\,=\,1}^{I} \pi_{ij}^{x_{ih}} (1 - \pi_{ij})^{1-x_{ih}} \qquad (2)$$

  In Equation 2, the probability of a correct response for a respondent in class j is denoted as π_{ij}, indicating that π is specific to each item within class j.

Because of the conceptual basis for linkage levels measuring a single skill (see Chapter II of the *2015–2016 Technical Manual – Science,* [DLM Consortium, 2017b]), the fungible model has been used to calibrate and score DLM assessments to date. However, given that the item parameters are allowed to vary in the nonfungible model, it is expected that this model would demonstrate superior model fit. As is the case for the vast majority of statistical models, additional parameters will increase the fit to the data. However, the trade-off is that increasing the number of parameters also increases the risk of overfitting the model to the data. When this happens, the model is not generalizable to data outside of the sample used to estimate the model. Thus, if a more parsimonious model can provide adequate model fit, that simpler model would be preferred. Additionally, because there are fewer parameters to estimate, the fungible model allows for a faster calibration. That all items in the fungible model share the same parameter also means extreme parameter values are less likely. Parameters for items with low sample sizes are not allowed to vary freely but are instead pulled into the fungible parameter, which is calculated on the full sample of available data. This has important implications for scoring in operational assessment systems.

V.4.B. Background on Model Fit Calculation

To provide evidence of model fit for two competing models (e.g., fungible and nonfungible), model fit evidence can be provided in the form of both relative and absolute fit indices. Relative fit compares the fit of two competing models to determine which model better fits the data. However, to determine how well each individual model fits the data, absolute fit indices are necessary. The sections that follow describe considerations when calculating both relative and absolute model fit.

V.4.B.i. Relative Fit

The relative fit of two competing models can be evaluated by comparing two nested models in a likelihood ratio test (Neyman & Pearson, 1933). This test provides information about which of two competing models provides better fit to the data when summarizing results across all linkage levels. Relative fit is calculated based on the final loglikelihoods from the nested models and the number of parameters in each. Take, for example, a latent class analysis with five items. In the nonfungible model, 11 parameters are estimated: a conditional probability of a correct response for masters and non-masters (one each for all five items = 10) and one structural parameter that is the base-rate probability of mastery. The fungible model has three parameters: one conditional probability of masters providing a correct response shared by all items, one conditional probability of non-masters providing a correct response shared by all items, and one structural parameter. Because the nonfungible model has more parameters, it is expected to always have a larger loglikelihood (i.e., better fit). However, the likelihood ratio test tests whether this increase is large enough to justify the additional parameters. The likelihood ratio test is a χ^2 test defined as follows:

$$\chi^2 = 2\ln\left(\frac{\text{likelihood for alternative model}}{\text{likelihood for null model}}\right) \tag{3}$$

$$df = df_{\text{alt}} - df_{\text{null}} \tag{4}$$

In this notation, the null model is the more simplified, or nested, model (the fungible model for DLM scoring). If this test is significant, then the null model is rejected, and it is determined that the additional parameters in the alternative model provide a statistically significant increase in the likelihood.

V.4.B.ii. Absolute Fit

In item response theory, model goodness-of-fit is commonly assessed using residual analysis (see Hambleton, Swaminathan, & Rogers, 1991). At the item level, the continuous theta is split into quadrature nodes, and the fitted item-characteristic curve is used to determine the expected proportion of correct responses at each quadrature point. The observed data are then used to calculate the observed proportion correct for each quadrature node. The difference between these proportions (i.e., the residual) is then standardized by dividing by the standard error of the residual. Thus, the prediction errors are essentially turned into *z* scores, which can be summed across all quadrature points for an item. Summed *z* scores follow a χ^2 distribution, with degrees of freedom equal to the number of quadrature points. Thus, for each item, a χ^2 test can be conducted to determine item-level misfit. At the test level, item-characteristic curves can be aggregated into a test-characteristic curve, and a similar test can be done across quadrature points to assess test-level model fit.

Because DLM assessments use diagnostic models, in which the latent trait is categorical rather than continuous, it is not possible to create item-characteristic or test-characteristic curves. Nevertheless, a similar approach can be taken in that a χ^2 can be calculated for each item based on the residuals. However, because the latent trait is categorical, the expected proportion of respondents in each score category can be calculated directly from model parameters instead of by breaking the trait into quadrature points.

As an example, consider a dichotomous attribute, where the base-rate probability of mastery is .6, and an item that measures this attribute, where masters have a .8 probability and non-masters a .15 probability of providing a correct response. Given these parameters, the proportion of respondents expected to provide a correct response can be calculated as follows:

$$\begin{aligned} P(X_i = 1) &= \eta_1\pi_{i1} + \eta_2\pi_{i2} \\ &= (0.6)(0.8) + (0.4)(0.15) \\ &= 0.54 \end{aligned} \tag{5}$$

Similarly, the proportion of respondents expected to provide an incorrect response can be calculated as follows:

$$\begin{aligned} P(X_i = 0) &= \eta_1(1 - \pi_{i1}) + \eta_2(1 - \pi_{i2}) \\ &= (0.6)(0.2) + (0.4)(0.85) \\ &= 0.46 \end{aligned} \tag{6}$$

These proportions can be converted to frequencies by multiplying the expected proportions by the total number of respondents who took the item. For example, if 100 respondents had taken this item, a contingency table could be constructed showing the number of expected and observed respondents at each score point (Table 16).

Table 16. Univariate Contingency Table

Item 1 score	Expected *N*	Observed *N*
0	46	48
1	54	52

Using the data in Table 16, a χ^2 goodness-of-fit test can be calculated ($\chi^2_{(1)}$ = 0.16, p = .68). Because the p value is nonsignificant, this test would not indicate item-level misfit.

In addition to looking at a single item, it is also possible to look at the fit of multiple items simultaneously. For example, when using two items, a 2x2 contingency table can be constructed to show the observed and expected frequencies of each response pattern. Table 17 presents these contingency tables together (one for expected frequencies and one for observed frequencies) together in one long-format table for readability.

Table 17. Bivariate Contingency Table

Item 1 score	Item 2 score	Expected *N*	Observed *N*
0	0	26	30
0	1	20	18
1	0	10	7
1	1	44	45

As with the univariate example in Table 16, a χ^2 goodness-of-fit test can also be conducted on these expected and observed frequencies ($\chi^2_{(3)}$ = 1.74, p = .63). This family of tests is known as limited information goodness-of-fit tests (see Maydeu-Olivares & Joe, 2006), as they use only subsets of items. This approach can continue to add dimensions (e.g., trivariate tables); however, as more dimensions are added, the number of possible responses increases exponentially (number of response patterns = 2^{items}). Thus, the expected and observed counts at each possible response pattern start to become too small for there to be a stable χ^2 test.

Because these tests can use only a subset of items, they are unable to give an evaluation of fit for the entire model. Unlike in item response theory, no test-characteristic curve can be used to aggregate across items. Theoretically, this could be achieved by using the model parameters to calculate the expected sum score for each latent class (similar to the test-characteristic curve indicating the expected sum score for each theta value). However, this is not feasible for the DLM assessment because of the administration design. The number of items tested per linkage level, and thus the total possible sum score, varies by student, depending on which testlet or testlets were administered. Thus, the expected score for a master depends on which testlets the student received.

Therefore, the item-level indices have to be aggregated up to the model level using different methodology. This evaluation takes advantage of the additive properties of χ^2 distributions (Lancaster & Seneta, 2005), in that the sum of χ^2 values is also χ^2 distributed, with degrees of freedom equal to the sum of degrees of freedom from the component χ^2 values. Take, for example, an attribute measured by five items. As seen in Table 18, five univariate χ^2 values could be estimated (one for each item), and each test would have one degree of freedom. Aggregating to the model level, the univariate model-level fit could be assessed by a χ^2 value equal to the sum of the five item-level indices with five degrees of freedom, as illustrated. In Table 18, the χ^2 test at the model level is nonsignificant, indicating acceptable model fit.

Table 18. Example Model-Level Univariate Fit

Item	χ^2	*df*	***p* value**
1	0.16	1	.68
2	1.20	1	.27
3	0.87	1	.35
4	0.98	1	.32
5	1.03	1	.31
Model	4.24	5	.52

Note. *df* = degrees of freedom

There are several limitations to this approach. First, χ^2 is perfectly additive only asymptotically. Given the low sample sizes on many of the items, this assumption is unlikely to hold. Also, for χ^2 to be additive asymptotically, each χ^2 value must be independent of one another. For the univariate index, the assumption holds, as item responses are assumed to be independent, conditional upon mastery status. However, this is clearly not met when the bivariate or trivariate indices are aggregated in a manner similar to Table 18. This result is because the same item would be included in multiple item-level bivariate and trivariate indices. Therefore, the true sampling distribution of the aggregated χ^2 is unclear. Because of these limitations, Rupp, Templin, and Henson (2010) suggest using only the value of the aggregated χ^2 as an overall index of model fit, with larger values indicating worse fit.

Because of the number of linkage levels that must be estimated for each model (i.e., fungible and nonfungible), it is difficult to summarize the aggregated χ^2 in any meaningful way. This is largely because the magnitude of χ^2 is dependent on the number of indices that contributed to the sum. For example, an aggregated χ^2 of 8.00 that came from 10 univariate tests, each with one degree of freedom, seems much more reasonable than if that number came from only two univariate tests. Therefore, to help summarize the findings in a useful way, *p* values are calculated for the model level χ^2 values, even though the asymptotic distribution is likely incorrect. This *p* value is used only as a flagging criterion to give a general idea of how much misfit exists across multiple linkage levels (i.e., content area or level), and not to make decisions about individual linkage levels specifically. The literature suggests that *p* values calculated from this reference asymptotic distribution are overly conservative, leading to the rejection of correctly specified models (Maydeu-Olivares & Joe, 2014). Therefore, when using this *p* value, it is likely that more misfit is identified than is actually present. Maydeu-Olivares and Joe (2014) proposed the use of the M_2 statistic, which combines information from multiple indices (e.g., univariate, bivariate, trivariate) in a way that allows hypothesis tests with expected Type I error rates. However, given the sparseness of DLM data, bivariate and trivariate indices cannot be calculated for many linkage levels, making this approach unfeasible.

V.4.C. Procedure for Evaluating Model Fit

V.4.C.i. Data

The estimation of the models used data from the 2015–2016 and 2016–2017 spring science assessment windows. Field-test testlets from previous years were not included.

V.4.C.ii. Method

To evaluate model fit, a *k*-fold cross validation procedure was used, which is also known as *v*-fold cross validation (see Arlot, 2010; Hastie, Tibshirani, & Friedman, 2009). The specific method was a stratified, fivefold procedure, whereby the data were divided into five sections and both the fungible and nonfungible models were estimated on four of the five sections. Model fit was then evaluated using the 20% of the data excluded from calibration. This process was repeated five times so that each subsection of the data was used as the validation set once (as demonstrated in Table 19). Before creating the five samples, the data were stratified at the item level to ensure that some data from all items were included in each of the subsamples. This process controlled variation that occurred because of item exclusion, which required a more vigorous investigation using a methodology similar to jackknife resampling (see Tukey, 1958).

Table 19. Specification of *k*-Fold Estimation Procedure

Calibration sets	Validation set
2, 3, 4, 5	1
1, 3, 4, 5	2
1, 2, 4, 5	3
1, 2, 3, 5	4
1, 2, 3, 4	5

For each validation set, both absolute and relative fit were evaluated as described above. The results were then averaged across all five validation sets. This approach has the advantages of using all of the data for both estimation and validation, while still evaluating model fit using different data than were used for estimation.

V.4.D. Results

V.4.D.i. Relative Fit

To assess relative fit, a fungible and a nonfungible model were estimated for each of the 102 linkage levels. For each linkage level, a likelihood ratio test was computed for the comparison of fungible (null) to nonfungible (alternative) model. For each test, if the *p* value of the likelihood ratio test was less than .05, the null model was rejected, meaning that the nonfungible model demonstrated better fit. The number of linkage levels that performed better in each model was calculated for each of the validation sets and then averaged across the five sets of results. These findings are summarized in Table 20.

Table 20. Average Number of Linkage Levels That Performed Better Over Five Validation Sets

	Fungible vs. Nonfungible	
Linkage level	**Fungible**	**Nonfungible**
Initial	7.0 (0.7)	27.0 (0.7)
Precursor	2.0 (1.0)	32.0 (1.0)
Target	1.2 (0.8)	32.8 (0.8)

Note. Parentheses indicate the standard deviation across the five validation sets.

The results summarized in Table 20 indicate that the nonfungible model fit the data better than the fungible model for nearly all linkage levels across all subjects. Furthermore, these analyses provide evidence that, as expected, the increase in model fit provided by the extra parameters in the nonfungible model was statistically significant. This is shown by the large discrepancy in the number of linkage levels in which the nonfungible model was preferred to the fungible model across content areas and linkage levels.

V.4.D.ii. Absolute Fit

When using the limited information indices of model fit, χ^2 is calculated for each item or set of items within a linkage level (see Table 16 and Table 17). To calculate fit for the entire linkage level, the χ^2 values for each item or set of items are summed (as shown in Table 18). A p value for the linkage level is then calculated for the summed χ^2 values, with degrees of freedom equal to the sum of degrees of freedom from each of the item-level tests (Lancaster & Seneta, 2005). If the p value is less than .05, then the expected counts are significantly different from the observed counts, indicating poor model fit. Therefore, nonsignificant values are desired and significant values are flagged for evidence of poor model fit. Because assumptions of the reference asymptotic distribution are likely not met, the p value likely results in more misfit being identified than is actually present and is therefore only used for flagging to give a general summary of the amount of misfit that could be present in each model.

Because results from the χ^2 test can be unreliable when cell counts are low, a minimum cell count of five was specified for each test. For example, in a linkage level measured by four items, there are six unique combinations of two items, meaning there are six possible bivariate indices. If any of the observed or expected counts for a response pattern in a given index were less than five, that index was not computed. Thus, it is possible that only four of the six possible bivariate indices would be computed. This means that when aggregating the item-level indices to the linkage level, only the four computed indices would be used. For DLM assessments, because of the sparseness of the data and the further sparseness introduced by the k-fold procedure, there were some linkage levels where no indices could be computed for the bivariate indices. However, there were no linkage levels for which trivariate indices could be computed, and therefore they are not included in these results. The *k*-fold procedure has the benefit of using different data for the estimation and analysis of model fit. However, when using five folds, the analysis of model fit is limited to only 20% of the total data. This reduced sample size vastly limits the ability to calculate the higher order fit indices.

Table 21 shows the number of linkage levels that were flagged for having poor model fit using each of the methods (univariate and bivariate), as well as the total number of linkage levels for which the index was computed. Results were averaged across all five validation sets. For example, in looking at the bivariate fit for the Initial linkage levels under the fungible model, the bivariate index could be calculated for 27 linkage levels on average, and of those, an average of six linkage levels showed poor model fit.

There are several things to note from Table 21. First, as expected, the number of indices that could be computed decreases with added dimensions to the χ^2 (i.e., univariate to bivariate indices). This is because with more dimensions, there are more possible response patterns, making it more difficult to obtain the sample-size threshold for each. Overall, given the noted constraints, and based on the results that are calculable, the nonfungible model provides the best model fit. The nonfungible model results in the lowest rates of flags across content areas and linkage levels.

In the fungible model, a large proportion of the indices computed were flagged for poor model fit, with an average of 60% flagged across linkage levels in the univariate index and 42% in the bivariate index. The nonfungible model, on the other hand, showed a fairly low percentage of linkage levels flagged for misfit (10% in the univariate index and 5% in the bivariate index). Additionally, it appears that misfit substantially decreased at the lower linkage levels for the nonfungible model (3% to 8% were flagged across both indices for Initial and Precursor linkage levels, compared to 22% to 64% for the fungible model). There is a clear increase in the percentage of indices flagged for misfit when moving from the Initial to Target levels.

Table 21. Average Number of Flagged Linkage Levels Using Limited Information Indices Over Five Validation Sets

	Fungible				Nonfungible			
	Univariate		Bivariate		Univariate		Bivariate	
Linkage level	**Flags**	***N***	**Flags**	***N***	**Flags**	***N***	**Flags**	***N***
Initial	12.4 (2.3)	34.0 (0.0)	6.0 (1.2)	27.0 (0.0)	2.4 (2.1)	34.0 (0.0)	1.2 (1.3)	27.0 (0.0)
Precursor	21.6 (1.5)	34.0 (0.0)	13.0 (1.9)	26.8 (0.4)	2.8 (1.3)	34.0 (0.0)	0.8 (0.8)	26.8 (0.4)
Target	27.4 (1.9)	34.0 (0.0)	13.0 (1.4)	23.6 (0.9)	4.6 (2.4)	34.0 (0.0)	1.8 (0.8)	23.8 (1.1)

Note. Parentheses indicate the standard deviation across the five validation sets. There are 34 Essential Elements for science.

V.4.E. Operational Evaluation of Model Fit

Statistical significance should not be the only deciding factor when evaluating the appropriateness of a psychometric model; practical significance should also be considered. Specifically, for DLM assessments, the practical significance of model-fit results can be gauged by how much performance varies according to the scoring model used. Although *k*-fold cross validation is suggested for model building and evaluation, Hastie et al. (2009) suggested using all the data for the final model to be used operationally. Thus, following the best practices, all five subsets were used to create an operational fungible and nonfungible calibration to assess practical significance.

One way to evaluate the effect on student results is to examine the structural parameter from each linkage level. This represents the base-rate probability of mastery for the linkage level and thus provides information about the proportion of students that are being classified as masters. If students are classified as masters at similar rates across models, then there is preliminary evidence that student results are not significantly affected by the choice of model.

Figure 8 shows the difference in base-rate mastery probabilities across models by linkage level. Generally, science shows consistency in mastery rates across the three levels with more variability occurring at the Precursor level and very little variability at the Initial level. Overall, performance in science is expected to be fairly consistent across models.

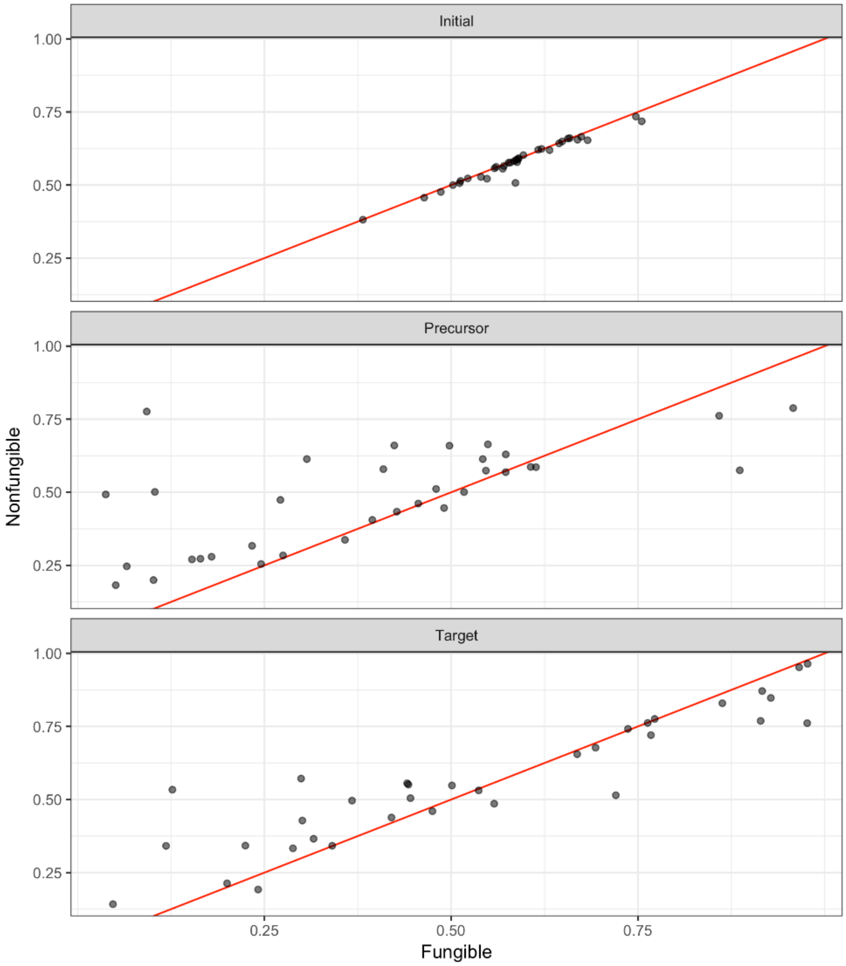


Figure 8. Comparison of base-rate mastery probability in the fungible and nonfungible models.

The consistency in results can be examined by comparing the number of linkage levels mastered by students when the fungible or the nonfungible model is used to score the assessment. As a natural extension to this analysis, the consortium-level impact data can also be compared across the two scoring models. For this analysis, each of the estimated models (i.e., fungible and nonfungible) was used to score the 2016–2017 operational assessment. For each model, the total linkage levels mastered by each student was calculated, and the percentage of students at each performance level for each grade and content area was determined using current operational cut points.[3] Figure 9 shows the comparison of total linkage levels mastered, including correlations, for science assessments. Figure 9 demonstrates that student results are extremely consistent across scoring models, with all correlations ranging from .98 to .99.

[3] Cut points were set for each tested grade level within the elementary and middle school grade bands; one set of cut points was used for the high school band.

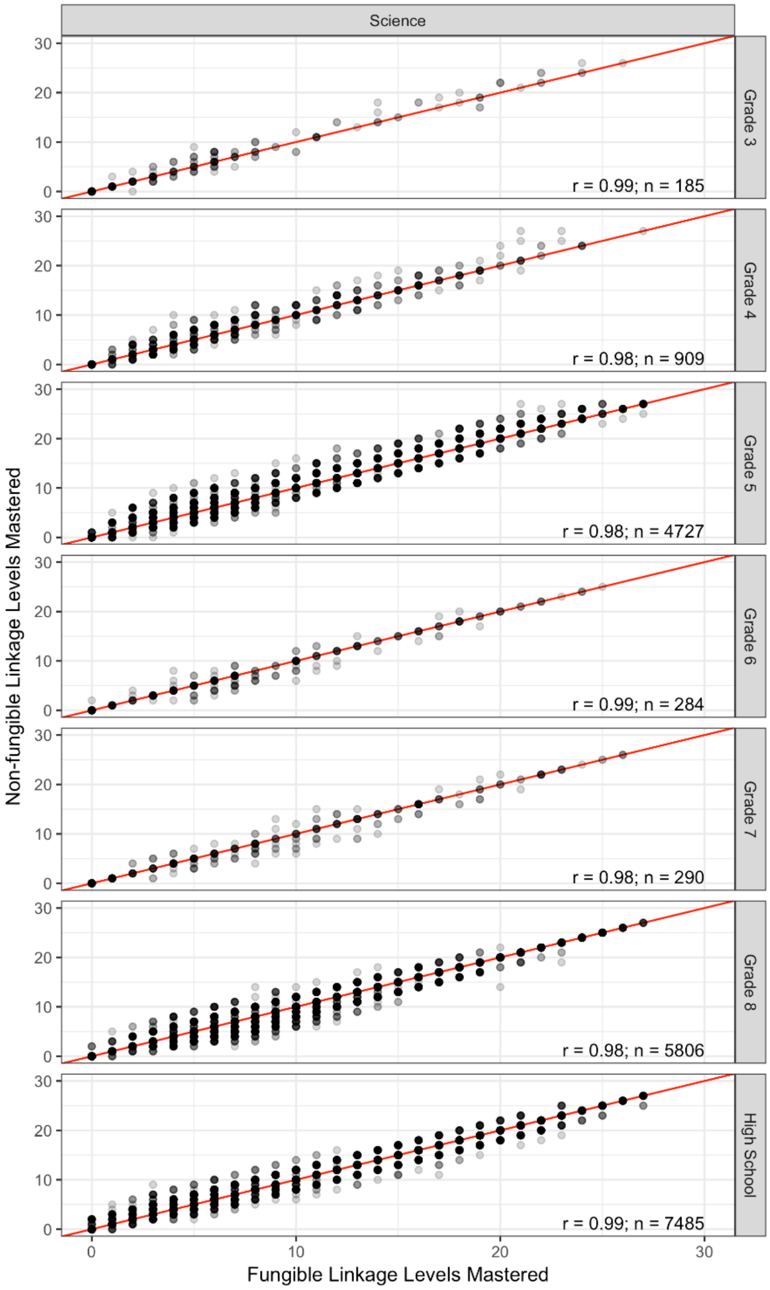
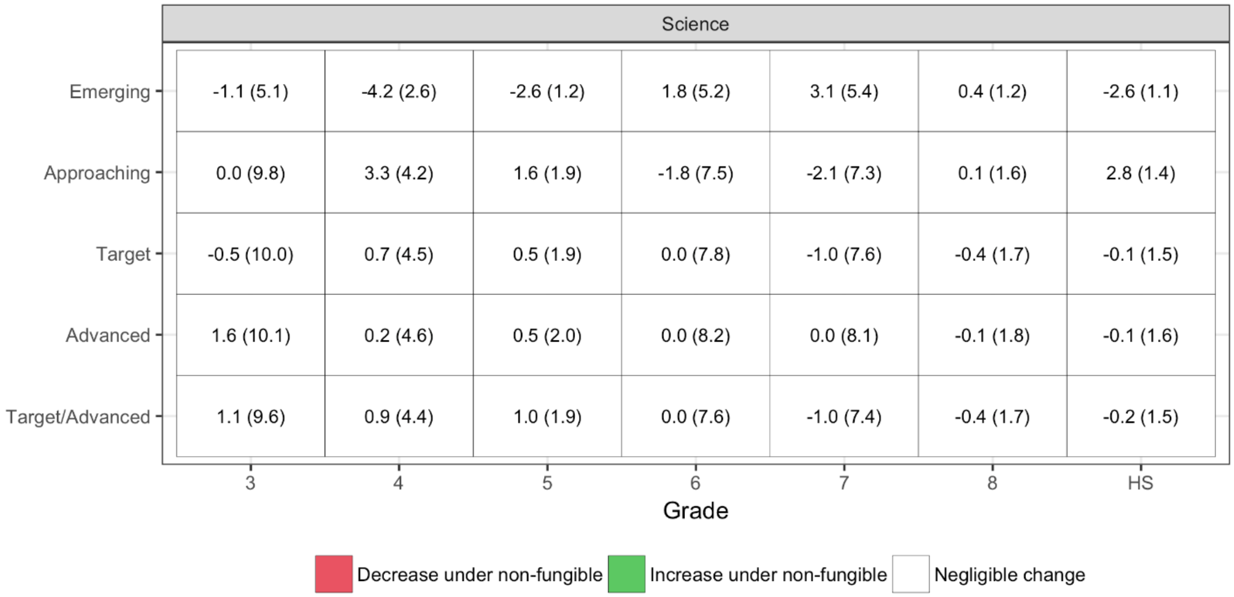


Figure 9. Total linkage levels mastered comparison.

A comparison of the percentage of students achieving at each performance level is also provided. Figure 10 shows the change in the percentage of students at each grade. The combined standard error of the difference is shown in parentheses, as calculated by $\sqrt{\sigma_1^2 + \sigma_2^2}$. For example, in fifth-grade science, 67.4% of students achieved at the Emerging category with the fungible model, compared to 64.9% with the nonfungible model, resulting in a change of 2.6 percentage points and a standard error of 1.2 that is interpreted on the scale of the percentages rather than for the difference value itself.



	3	4	5	6	7	8	HS
Emerging	-1.1 (5.1)	-4.2 (2.6)	-2.6 (1.2)	1.8 (5.2)	3.1 (5.4)	0.4 (1.2)	-2.6 (1.1)
Approaching	0.0 (9.8)	3.3 (4.2)	1.6 (1.9)	-1.8 (7.5)	-2.1 (7.3)	0.1 (1.6)	2.8 (1.4)
Target	-0.5 (10.0)	0.7 (4.5)	0.5 (1.9)	0.0 (7.8)	-1.0 (7.6)	-0.4 (1.7)	-0.1 (1.5)
Advanced	1.6 (10.1)	0.2 (4.6)	0.5 (2.0)	0.0 (8.2)	0.0 (8.1)	-0.1 (1.8)	-0.1 (1.6)
Target/Advanced	1.1 (9.6)	0.9 (4.4)	1.0 (1.9)	0.0 (7.6)	-1.0 (7.4)	-0.4 (1.7)	-0.2 (1.5)

Figure 10. Change in percentage of students achieving at each performance level.
Note. Where applicable, highlighted cells indicate a change of more than 5 percentage points. The standard error of the difference is shown in parentheses.

Similar to the comparison of the structural parameters, the science results are extremely consistent, with no performance levels flagged for changes of more than 5 percentage points.

V.4.F. Summary of Model Fit Analyses

This chapter presents two methods for evaluating model fit, along with comparisons of the operational effect of results obtained from the competing models. This included a relative fit analysis comparing model-to-model fit of the fungible and nonfungible models, and the absolute fit of each model was summarized via univariate and bivariate indices.

Overall, the combination of relative and absolute fit from the limited information tests indicates that the data best support use of a nonfungible model. The nonfungible model showed significantly better fit on the majority of linkage levels when compared to the fungible model, and showed the fewest flags in the univariate and bivariate indices. However, a number of methodological constraints were noted, including using p values to evaluate the model-level χ^2 values and limited sample sizes using the k-fold validation approach that call into question their use for operational decision-making purposes. Furthermore, the operational comparison of student results showed that the choice of model had little effect on student results. Additionally, there are practical benefits to using a more parsimonious model, including simpler and faster estimation for delivering student results on the timeline needed by states for accountability decision-making purposes. Finally, the recommendations of the DLM Technical Advisory Committee (TAC) have focused on exploring a Bayesian estimation procedure to help address some of the methodological issues with the current approach to assessing model fit. Specific next steps in the research agenda are to implement a Bayesian estimation technique and reevaluate model fit for both the fungible and nonfungible models. Although the current evidence suggests that the nonfungible model fits the data better than the fungible model, methodological constraints of the current evaluation, the limited and varied effect of model choice on students' results, and the practical benefits of the fungible model have led to the decision to retain the fungible model for operational scoring for the 2017–2018 academic year. Ongoing research is planned to identify an improved modeling strategy and corresponding assessment design. The plan to continue calibrating and scoring DLM assessments using a fungible model for the 2017–2018 administration was discussed with the DLM TAC during the August 2017 partner call, and the TAC indicated support for the plan.

V.5. Conclusion

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

A review of model parameters indicates that for most linkage levels, the conditional probability of masters providing a correct response falls above .5, and for most linkage levels the conditional probability of non-masters providing a correct response falls below .5. Beginning in spring 2018, test-development teams will begin reviewing model-based flagging to identify potential areas that may introduce construct-irrelevant variance into the calculation of student results.

Preliminary model-fit results indicated mixed support for the use of the current fungible scoring model. Because new modeling strategies may provide better alternatives for the assessment of model fit, current work focuses on developing a Bayesian estimation process for the fungible, nonfungible, and partially fungible models, whereby a partial equivalency model can be estimated. This approach would support improved methods for the assessment of model fit. Specifically, using Markov chain Monte Carlo estimation would allow the evaluation of model fit using posterior-predictive model checking (Gelman & Hill, 2006; Gelman, Meng, & Stern, 1996). The development of this procedure is underway; upon its completion, it will be disseminated for review to the DLM TAC modeling subcommittee, a subgroup of TAC members focused on reviewing modeling-specific topic guides.

VI. STANDARD SETTING

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

VII. ASSESSMENT RESULTS

Chapter VII of the *2016–2017 Technical Manual – Science* (Dynamic Learning Maps® [DLM®] Consortium, 2017b) describes assessment results for the 2016–2017 academic year, including student participation and performance summaries and an overview of data files and score reports delivered to state partners. This chapter presents 2016–2017 student participation data; final results in terms of the percentage of students at each performance level; and subgroup performance by gender, race, ethnicity, and English learner (EL) status for the 2016–2017 administration year. This chapter also reports the distribution of students by the highest linkage level mastered during 2016–2017. Finally, this chapter describes updates made to Individual Student Score Reports, data files, and quality control procedures during the 2016–2017 operational year. For a complete description of and interpretive guides, see Chapter VII of the *2015–2016 Technical Manual – Science* (DLM Consortium, 2017b).

VII.1. Student Participation

The spring 2017 assessments were administered to 19,686 students in seven states and one Bureau of Indian Education school. Counts of students tested in each state are displayed in Table 22. The assessment sessions were administered by 7,841 educators in 5,577 schools and 1,925 school districts.

Table 22. Student Participation by State (*N* = 19,686)

State	Students (*n*)
Alaska	214
Illinois	4,812
Iowa	1,042
Kansas	1,200
Maryland	1,674
Miccosukee Indian School	8
Missouri	3,676
Oklahoma	2,327
West Virginia	852
Wisconsin	3,881

Table 23 summarizes the number of students tested in each grade during spring 2017. More than 6,000 students participated in each of the elementary and the middle school grade bands.[4] In high school, almost 7,500 students participated. The differences in grade-level participation within each band can be traced to differing state-level policies about the grade in which students are assessed.

Table 23. Student Participation by Grade (*N* = 19,686)

Grade	Students (*n*)
3	185
4	909
5	4,727
6	284
7	290
8	5,806
9	958
10	2,005
11	4,253
12	269

Table 24 summarizes the demographic characteristics of students who participated in the spring 2017 administration. The majority of participants were male (65%) and white (63%). Only 3.6% of students were reported to be eligible for or monitored for EL services. Because teachers were not required to complete all of the student demographic information, some variables in the following tables are missing data.

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

Table 24. Demographic Characteristics of Participants

Subgroup	*n*	%
Gender		
Female	6,866	34.88
Male	12,816	65.10
Missing	4	0.02
Race		
White	12,371	62.84
African American	3,893	19.78
Asian	635	3.23
American Indian	624	3.17
Alaska Native	106	0.54
Two or more races	1,961	9.96
Native Hawaiian or Pacific Islander	66	0.34
Missing	30	0.15
Hispanic ethnicity		
No	17,014	86.43
Yes	2,606	13.24
Missing	66	0.34
English learner (EL) participation		
Not EL eligible or monitored	18,963	96.33
EL eligible or monitored	719	3.65
Missing	4	0.02

VII.2. Student Performance

Student performance on Dynamic Learning Maps (DLM) assessments is interpreted using cut points, determined during standard setting (see Chapter VI in the *2015–2016 Technical Manual – Science* [DLM Consortium, 2017b]), which separate student scores into four performance levels. A student receives a performance level based on the total number of linkage levels mastered across the assessed Essential Elements (EEs).

For the 2016–2017 administration, student performance was reported using the same four performance levels approved by the DLM Consortium for the 2015–2016 year.

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

VII.2.A. Overall Performance

Table 25 reports the performance distributions from the 2016–2017 spring administration for science.

The 2016–2017 results were fairly consistent with 2015–2016 performance distributions with the majority of students categorized as either Emerging or Approaching the Target performance levels. At the elementary level, the percentage of students who demonstrated performance at the At Target or Advanced levels ranged from approximately 11% to 15%; in middle school the range was 18% to 21%; and in high school the percentages ranged from 8% to 23%.

Table 25. Percentage of Students by Grade and Performance Level

Grade	Emerging (%)	Approaching (%)	Target (%)	Advanced (%)	Target+Advanced (%)
3 (*n* = 185)	76.2	10.3	8.7	4.9	13.6
4 (*n* = 909)	72.1	17.1	8.6	2.3	10.9
5 (*n* = 4,727)	67.4	17.9	13.6	1.1	14.7
6 (*n* = 284)	61.3	20.4	14.1	4.2	18.3
7 (*n* = 290)	56.6	22.8	16.9	3.8	20.7
8 (*n* = 5,806)	58.1	23.7	16.0	2.2	18.2
9 (*n* = 958)	62.4	24.4	11.4	1.8	13.2
10 (*n* = 2,005)	46.4	30.4	16.7	6.5	23.2
11 (*n* = 4,253)	58.7	26.6	12.3	2.4	14.7
12 (*n* = 269)	76.2	15.6	7.1	1.1	8.2

VII.2.B. Subgroup Performance

Performance-level results for subgroups, including groups based on gender, race, ethnicity, and EL status, were computed.

The distribution of students across performance levels was examined using demographic subgroups. Table 26 summarizes the disaggregated frequency distributions for science collapsed across all assessed grade levels. Although each state has its own rules for minimum student counts needed to support public reporting of results, small counts are not suppressed here because results are aggregated across states and individual students cannot be identified. Rows labeled *Missing* indicate the student's demographic data were not entered into the system. Overall, fewer demographic data were missing in 2016–2017 than in the previous year.

Table 26. Students at Each Performance Level by Demographic Subgroup (*N* = 19,686)

	Emerging		Approaching		Target		Advanced	
Subgroup	*n*	%	*n*	%	*n*	%	*n*	%
Gender								
Female	4,267	62.1	1,676	24.4	793	11.5	130	1.9
Male	7,655	59.7	2,857	22.3	1,950	15.2	354	2.8
Missing	1	25.0	2	50.0	1	25.0	n/a	n/a

	Emerging		**Approaching**		**Target**		**Advanced**	
Subgroup	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Race								
White	7,281	58.9	2,901	23.5	1,847	14.9	342	2.8
African American	2,439	62.7	902	23.2	484	12.4	68	1.7
Asian	468	73.7	127	20.0	36	5.7	4	0.6
American Indian	296	47.4	136	21.8	152	24.4	40	6.4
Alaska Native	69	65.1	24	22.6	13	12.3	n/a	n/a
Two or more races	1,319	67.3	416	21.2	197	10.0	29	1.5
Native Hawaiian or Pacific Islander	41	62.1	17	25.8	7	10.6	1	1.5
Missing	10	33.3	12	40.0	8	26.7	n/a	n/a
Hispanic ethnicity								
No	10,140	59.6	3,944	23.2	2,480	14.6	450	2.6
Yes	1,746	67.0	574	22.0	257	9.9	29	1.1
Missing	37	56.1	17	25.8	7	10.6	5	7.6
English learner (EL) participation								
Not EL eligible or monitored	11,477	60.5	4,347	22.9	2,665	14.1	474	2.5
EL eligible or monitored	445	61.9	186	25.9	79	11.0	9	1.3
Missing	1	25.0	2	50.0	n/a	n/a	1	25.0

VII.2.C. LINKAGE-LEVEL MASTERY

As described earlier in the chapter, overall performance in the content area is calculated based on the number of linkage levels mastered across all EEs. Based on the scoring method, for each EE the highest linkage level the student mastered can be identified. This means that a student may be classified as a master of zero, one (Initial), two (Initial and Precursor), or three (Initial, Precursor, and Target) linkage levels. This section summarizes the distribution of students by highest linkage level mastered across all EEs in each grade. For each grade band, the numbers of students who showed no evidence of mastery, Initial-level mastery, Precursor-level mastery and Target-level mastery (as the highest level of mastery) were summed across all EEs and divided by the total number of students assessed to get the proportion of students who mastered each linkage level.

Table 27 reports the percentage of students who mastered each linkage level as the highest linkage level across all EEs for each grade. For example, across all third-grade EEs, 40.5% of the time the highest level that students mastered was the Initial level. The percentage of students who mastered as high as the Target linkage level ranged from approximately 26% in third grade to 47% in tenth grade.

Table 27. Percentage of Students Demonstrating Highest Linkage Level Mastered Across Essential Elements by Grade

	Linkage level			
Grade	**No evidence (%)**	**Initial (%)**	**Precursor (%)**	**Target (%)**
3 (*n* = 185)	14.6	40.5	18.9	25.9
4 (*n* = 909)	8.1	43.2	18.6	30.0
5 (*n* = 4,727)	4.2	40.6	18.3	36.9
6 (*n* = 284)	9.2	25.4	31.7	33.8
7 (*n* = 290)	8.6	23.1	24.5	43.8
8 (*n* = 5,806)	4.9	21.0	32.3	41.8
9 (*n* = 958)	8.1	34.9	26.7	30.3
10 (*n* = 2,005)	5.8	24.6	22.4	47.1
11 (*n* = 4,253)	6.2	32.5	26.4	35.0
12 (*n* = 269)	21.2	43.5	13.4	21.9

VII.3. DATA FILES

Four data files, made available to DLM state partners, summarized results from the 2016–2017 year. Similar to the previous year, the General Research File (GRF) contained student results, including each student's highest linkage level mastered for each EE and final performance level for the subject for all students who completed any testlets. During the 2016–2017 year, the GRF was restructured to include one row per student per subject, with a corresponding EE crosswalk provided to identify the EE reported in each column, generically named EE1 – EE26.

In addition to the GRF, several supplemental files were delivered. The Incident File listed students who were potentially affected by an administration incident during the spring window (see Chapter IV of this manual) using the same structure as the prior year. Similarly, the Special Circumstances File was retained in 2016–2017, which provided information about which students and EEs were affected by extenuating circumstances (e.g., chronic absences), as defined by each state. A new supplemental file was also delivered to identify exited students who did not reenroll for the remainder of the window.

Consistent with 2015–2016, state partners were provided with a 2-week review window following delivery of the GRF to invalidate student records. Once final GRFs were submitted back to DLM staff, the final GRF was uploaded to Educator Portal.

VII.4. SCORE REPORTS

The DLM Consortium provides assessment results to all member states to report to parents/guardians and to educators at state and local education agencies. Individual Student Score Reports were provided to educators and parents/guardians. 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 aggregated report structure during 2017; however, district and state reports were generated in Educator Portal following final GRF upload rather than being generated outside the system by the score-report program. Changes to the Individual Student Score Reports are summarized below. For a complete description of score reports, including aggregated reports, see Chapter VII of the *2014–2015 Technical Manual – Integrated Model* (DLM Consortium, 2016c).

VII.4.A. INDIVIDUAL STUDENT SCORE REPORTS

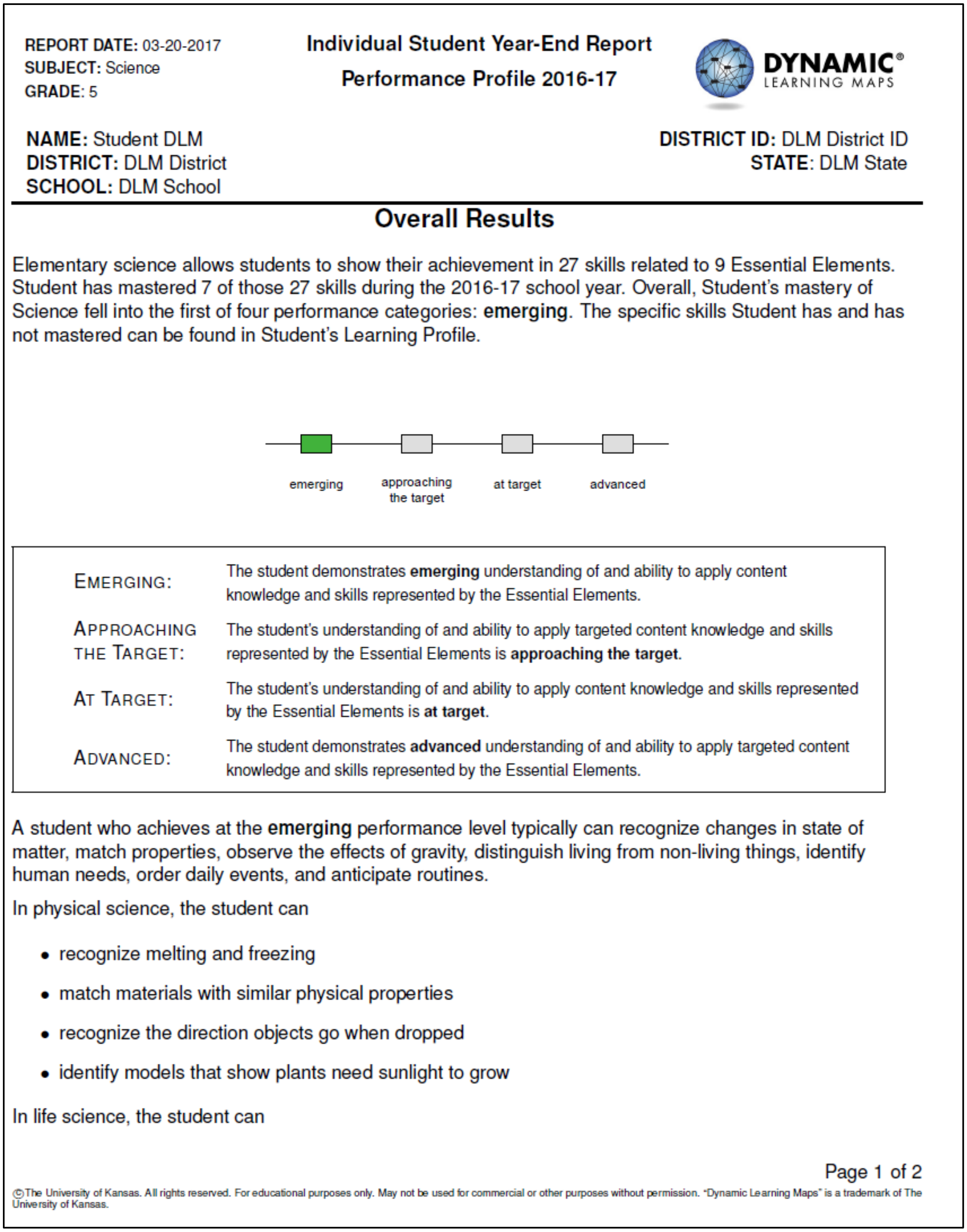
During the 2016–2017 year, minor changes were made to the Individual Student Score Reports.

One change to the content of the Performance Profile in science was the inclusion of grade-level performance level descriptors (PLDs). These grade-level PLDs replaced the bulleted list of skills mastered used in 2015–2016. The grade-level PLDs were developed after standard setting was conducted in 2016 to describe the types of skills typically mastered by students in a given performance level.

One change to the Learning Profile[5] was the adjustment of shading for "No evidence of mastery on this Essential Element" and "Essential Element not tested" to appear in the Essential Element column rather than the first-level cell to indicate the shading applied to the entire EE.

A sample Individual Student Score Report reflecting the 2017 changes is provided in Figure 11.

[5] Consistent with 2015–2016, only states that follow the integrated assessment model for DLM English language arts and mathematics receive the Learning Profile in all three subject areas. Year-end states requested this information be omitted for science to be consistent with their ELA and mathematics reports.



REPORT DATE: 03-20-2017
SUBJECT: Science
GRADE: 5

Individual Student Year-End Report
Performance Profile 2016-17

DYNAMIC® LEARNING MAPS

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

DISTRICT ID: DLM District ID
STATE: DLM State

Overall Results

Elementary science allows students to show their achievement in 27 skills related to 9 Essential Elements. Student has mastered 7 of those 27 skills during the 2016-17 school year. Overall, Student's mastery of Science fell into the first of four performance categories: **emerging**. The specific skills Student has and has not mastered can be found in Student's Learning Profile.

🟩	□	□	□
emerging	approaching the target	at target	advanced

EMERGING: The student demonstrates **emerging** understanding of and ability to apply content knowledge and skills represented by the Essential Elements.

APPROACHING THE TARGET: The student's understanding of and ability to apply targeted content knowledge and skills represented by the Essential Elements is **approaching the target**.

AT TARGET: The student's understanding of and ability to apply content knowledge and skills represented by the Essential Elements is **at target**.

ADVANCED: The student demonstrates **advanced** understanding of and ability to apply targeted content knowledge and skills represented by the Essential Elements.

A student who achieves at the **emerging** performance level typically can recognize changes in state of matter, match properties, observe the effects of gravity, distinguish living from non-living things, identify human needs, order daily events, and anticipate routines.

In physical science, the student can

- recognize melting and freezing
- match materials with similar physical properties
- recognize the direction objects go when dropped
- identify models that show plants need sunlight to grow

In life science, the student can

Page 1 of 2

Figure 11. Page 1 of the performance profile for 2016–2017.

VII.5. Quality Control Procedures for Data Files and Score Reports

Quality control (QC) procedures were updated in 2017 to include a score-report viewer tool to facilitate the manual quality control checks. No changes were made to automated QC checks for 2017. For a complete description of QC procedures, see Chapter VII in the *2015–2016 Technical Manual – Science* (DLM Consortium, 2017b).

VII.5.A. Manual Quality Control Checks

A PDF viewer tool was developed to increase the speed and efficiency of the manual QC process. Based on the data file fed to the program, reports were randomly selected from a relevant subset (model, grade, and content area) one at a time. When a report was selected, its data row from the GRF was displayed and the report was automatically opened by the tool so the two could be compared very quickly without having to navigate manually through folders where reports were stored. After a QC person finished reviewing the selected report, they clicked through and the tool then continued to the next report and its corresponding data row for review. This process was repeated until a minimum threshold for number of reports checked was met in the relevant subset.

VIII. RELIABILITY

Chapter VIII of the *2015–2016 Technical Manual – Science* (Dynamic Learning Maps® [DLM®] Consortium, 2017b) describes the methods used to calculate reliability and provides results at six reporting levels. This chapter provides a high-level summary of the methods used to calculate reliability, along with updated evidence from the 2016–2017 administration year for six levels, consistent with the levels of reporting.

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

VIII.1. Background Information on Reliability Methods

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

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

VIII.2. Methods of Obtaining Reliability Evidence

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

The simulation used to estimate reliabilities for DLM versions of scores and classifications considers the unique design and administration of DLM assessments. The use of simulation is necessitated by two factors: the assessment blueprint and the classification-based results that such administrations give. Because of the limited number of items students complete to cover the blueprint, students take only minimal items per EE. The reliability simulation replicates DLM versions of scores from actual examinees based upon the actual set of items each examinee took. Therefore, this simulation provides a replication of the administered items for the examinees. Because the simulation is based on a replication of the same items that were administered to examinees, the two administrations are perfectly parallel.

VIII.2.A. RELIABILITY SAMPLING PROCEDURE

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

1. Draw with replacement the student record of one student from the operational testing data. Use the student's originally scored pattern of linkage-level mastery and non-mastery as the true values for the simulated student data.
2. Simulate a new set of item responses to the set of items administered to the student in the operational testlet. Item responses are simulated from calibrated-model parameters[6] for the items of the testlet, conditional on the profile of linkage-level mastery or non-mastery for the student.
3. Score the simulated-item responses using the operational DLM scoring procedure (see Chapter V of the *2015–2016 Technical Manual – Science* (DLM Consortium, 2017b) for more information),[7] producing estimates of linkage-level mastery or non-mastery for the simulated student.
4. Compare the estimated linkage-level mastery or non-mastery to the known values from Step 2 for all linkage levels for which the student was administered items.
5. Repeat Steps 1 through 4 for 2,000,000 simulated students.

Figure 12 shows Steps 1 through 4 of the simulation process as a flow chart.

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

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

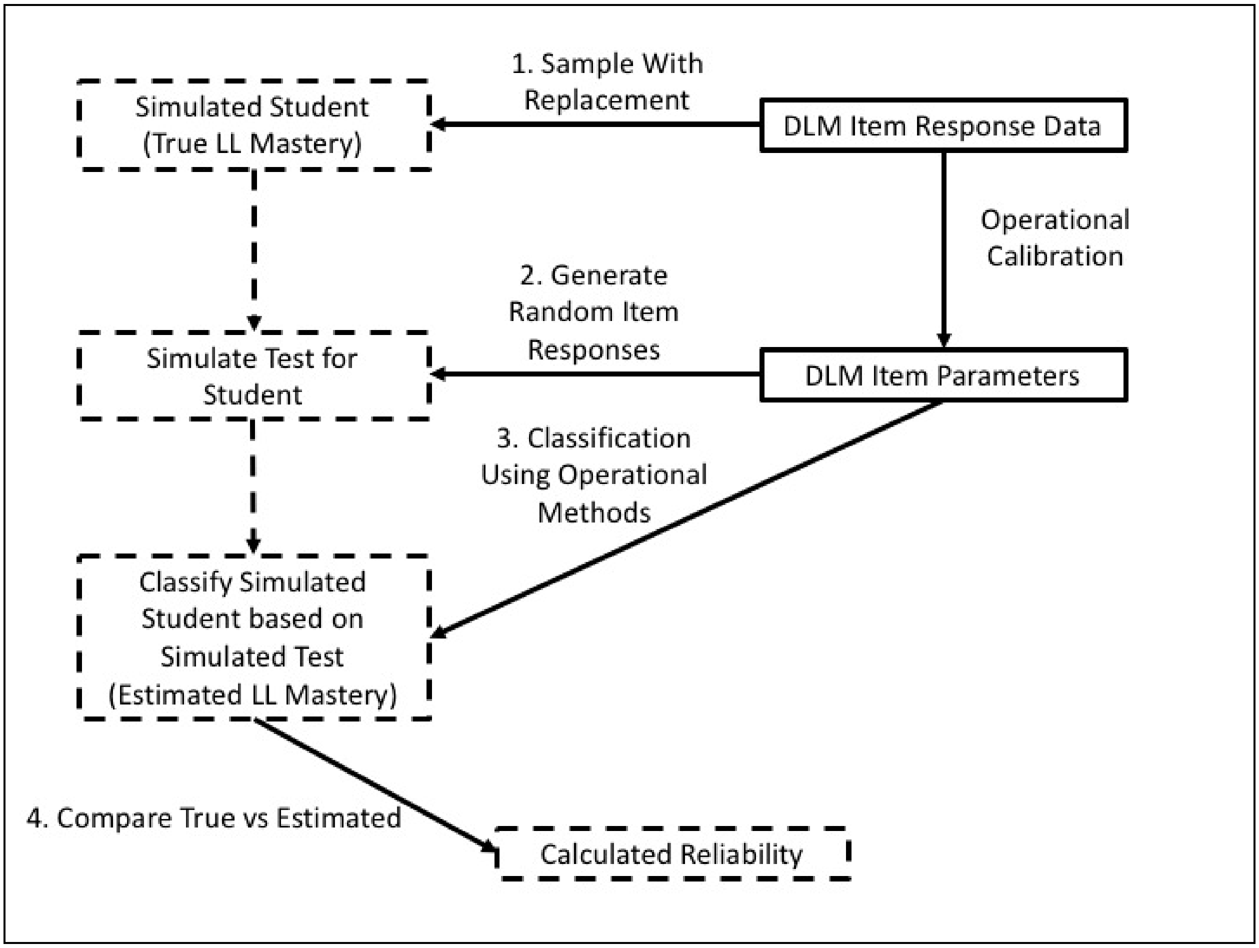

1. Sample With Replacement
Simulated Student (True LL Mastery)
DLM Item Response Data
Operational Calibration
2. Generate Random Item Responses
Simulate Test for Student
DLM Item Parameters
3. Classification Using Operational Methods
Classify Simulated Student based on Simulated Test (Estimated LL Mastery)
4. Compare True vs Estimated
Calculated Reliability

Figure 12. Simulation process for creating reliability evidence.
Note. LL = linkage level.

VIII.2.B. Reliability Evidence

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

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

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

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

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

Reliability evidence is given for six levels of data: (a) performance-level reliability, (b) content-area reliability, (c) domain reliability, (d) EE reliability, (e) linkage-level reliability, and (f) conditional reliability by linkage level. With 34 EEs, each with three linkage levels, 102 analyses were conducted to summarize reliability. Because of the number of analyses, the reported evidence will be summarized in this chapter. Full reporting of reliability evidence for all 102 linkage levels and 34 EEs is provided in an online appendix (http://dynamiclearningmaps.org/reliabevid). The full set of evidence is provided in accordance with Standard 2.12.

Reporting reliability at six levels ensures that the simulation and resulting reliability evidence were performed in accordance with Standard 2.2. Providing reliability evidence for each of the six levels also ensures that these reliability estimation procedures meet Standard 2.5.

VIII.2.B.i. Performance-Level Reliability Evidence

Four performance levels were used to report results from DLM assessments. The total linkage levels mastered in each content area is summed, and cut points are applied to distinguish between performance categories.

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

1. The polychoric correlation between the true and estimated performance level within a grade and content area
2. The correct classification rate between the true and estimated performance level within a grade and content area
3. The correct classification kappa between the true and estimated performance level within a grade and content area

Table 28 shows this information across all grades and content areas. Polychoric correlations between true and estimated performance levels ranged from .927 to .974. Correct classification rates ranged from .853 to .910 and Cohen's kappa values were between .740 and .873. These results indicate that the DLM scoring procedure of assigning and reporting performance levels based on total linkage levels mastered results in reliable classification of students to performance-level categories.

Table 28. Summary of Performance-Level Reliability Evidence

Grade	Polychoric correlation	Correct classification rate	Cohen’s kappa
3	.947	.910	.787
4	.930	.904	.741
5	.934	.909	.740
6	.933	.853	.774
7	.950	.888	.798
8	.927	.872	.758
9	.963	.881	.858
10	.965	.857	.868
11	.965	.876	.862
12	.974	.908	.873

VIII.2.B.ii. Content-Area Reliability Evidence

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

Content-area reliability evidence compares the true and estimated numbers of linkage levels mastered across all tested levels in science. Reliability is reported with three summary numbers.

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

Table 29 shows the three summary values for each grade. Classification-rate information is provided in accordance with Standard 2.16. The two summary statistics included in Table 29 also meet Standard 2.19. The correlation between true and estimated numbers of linkage levels mastered, ranged from .884 to .954. Average student correct classification rates ranged from .977

to .990 and average student Cohen's kappa values ranged from .954 to .980. These values indicate that the DLM scoring procedure of reporting the number of linkage levels mastered provides reliable results of student performance.

Table 29. Summary of Content-Area Reliability Evidence by Grade

Grade	Linkage levels mastered correlation	Average student correct classification	Average student Cohen's kappa
3	.914	.987	.975
4	.884	.985	.970
5	.899	.983	.966
6	.911	.979	.957
7	.907	.977	.954
8	.892	.978	.955
9	.943	.985	.972
10	.949	.980	.961
11	.945	.984	.970
12	.954	.990	.980

VIII.2.B.iii. Domain Reliability Evidence

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

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

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

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

3. The correct classification kappa for which linkage levels were mastered as averaged across all simulated students for each domain

Table 30 shows the three summary values for each domain by grade. Values ranged from .621 to .998, indicating that the DLM method of reporting the total and percentage of linkage levels mastered by domain generally results in values that can be reliably reproduced.

Table 30. Summary of Science-Domain Reliability Evidence by Grade

Grade	Domain	Linkage levels mastered correlation	Average student correct classification	Average student Cohen's kappa
3	ESS	.621	.998	.998
3	LS	.647	.996	.995
3	PS	.895	.996	.994
4	ESS	.634	.998	.997
4	LS	.647	.996	.995
4	PS	.868	.995	.992
5	ESS	.675	.998	.997
5	LS	.660	.996	.995
5	PS	.885	.995	.992
6	ESS	.807	.997	.996
6	LS	.710	.990	.985
6	PS	.916	.996	.995
7	ESS	.806	.997	.997
7	LS	.718	.989	.983
7	PS	.909	.996	.995
8	ESS	.810	.997	.997
8	LS	.682	.990	.985
8	PS	.905	.996	.995
9	ESS	.678	.996	.994
9	LS	.810	.996	.995
9	PS	.906	.997	.996
10	ESS	.679	.995	.994

Grade	Domain	Linkage levels mastered correlation	Average student correct classification	Average student Cohen's kappa
10	LS	.833	.995	.993
10	PS	.906	.996	.995
11	ESS	.688	.996	.994
11	LS	.822	.995	.994
11	PS	.898	.996	.95
12	ESS	.731	.997	.996
12	LS	.853	.996	.995
12	PS	.924	.997	.996

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

VIII.2.B.iv. Essential-Element Reliability Evidence

Moving from higher-level aggregation to EEs, the reliability evidence shifts slightly. That is, because EEs are collections of linkage levels with an implied order, EE-level results are reported as the highest linkage level mastered per EE. If one considers content-area scores as total scores from an entire test, evidence at the EE level is more fine-grained than reporting at a content-area strand level, which is commonly reported for other testing programs. EEs are the specific standards within the content area itself.

Three statistics are used to summarize reliability evidence for EEs.

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

Because there are 34 EEs, the summaries reported herein are based on the number and proportion of EEs that fall within a given range of an index value. Results are given in both tabular and graphical forms. Table 31 and Figure 13 provide proportions and the number of EEs, respectively, that fall within prespecified ranges of values for the three reliability summary statistics (i.e., correct classification rate, kappa, and correlation). In general, the reliability summaries for number of linkage levels mastered within EEs show strong evidence of reliability.

Table 31. Reliability Summaries Across All Essential Elements (EEs): Proportion of EEs Falling Within a Specified Index Range

Reliability index	Index range								
	< .60	.60–.64	.65–.69	.70–.74	.75–.79	.80–.84	.85–.89	.90–.94	.95–1.0
Polychoric correlation	.000	.000	.074	.037	.259	.037	.296	.111	.185
Correct classification rate	.000	.000	.000	.000	.000	.259	.407	.296	.037
Kappa	.111	.111	.148	.111	.074	.222	.111	.037	.074

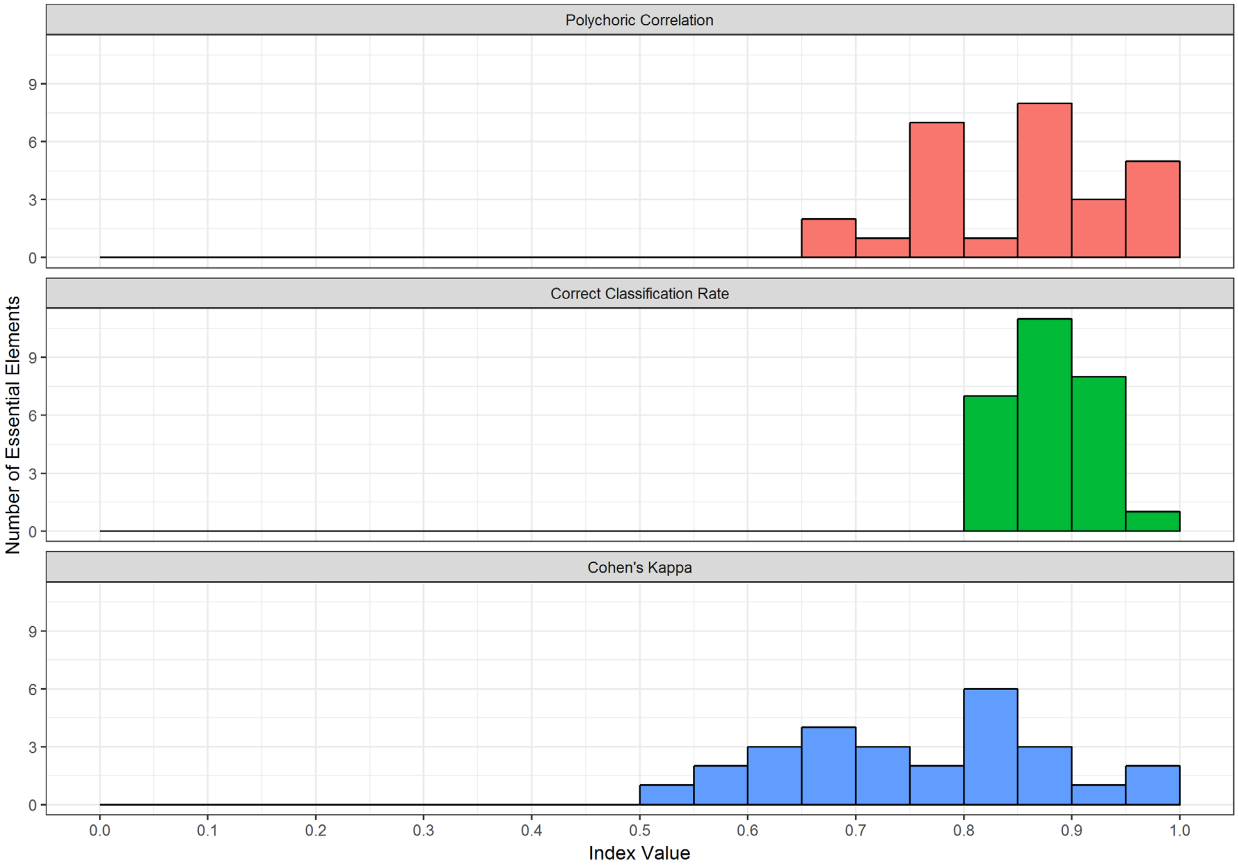


Figure 13. Number of linkage levels mastered within Essential Element reliability summaries.

VIII.2.B.v. Linkage-Level Reliability Evidence

Evidence at the linkage level comes from the comparison of true and estimated mastery statuses for each of the 102 linkage levels in the operational DLM assessment.[8] This level of reliability reporting is even more fine-grained than the EE level. While it does not have a comparable classical test theory or item response theory analog, its inclusion is important because it is the level where mastery classifications are made for DLM assessments.

As an example, Table 32 shows a simulated table from one linkage level of an EE.

Table 32. Example of True and Estimated Mastery Status From Reliability Simulation

		Estimated mastery status	
		Non-master	**Master**
True mastery status	**Non-master**	574	235
	Master	83	592

The summary statistics reported are all based on tables like this one: the comparison of true and estimated mastery statuses across all simulated examinees. As with any contingency table, a number of summary statistics are possible.

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

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

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

As there are 102 total linkage levels across all 34 EEs, the summaries reported herein are based on the proportion and number of linkage levels that fall within a given range of an index value. Results are given in both tabular and graphical form.

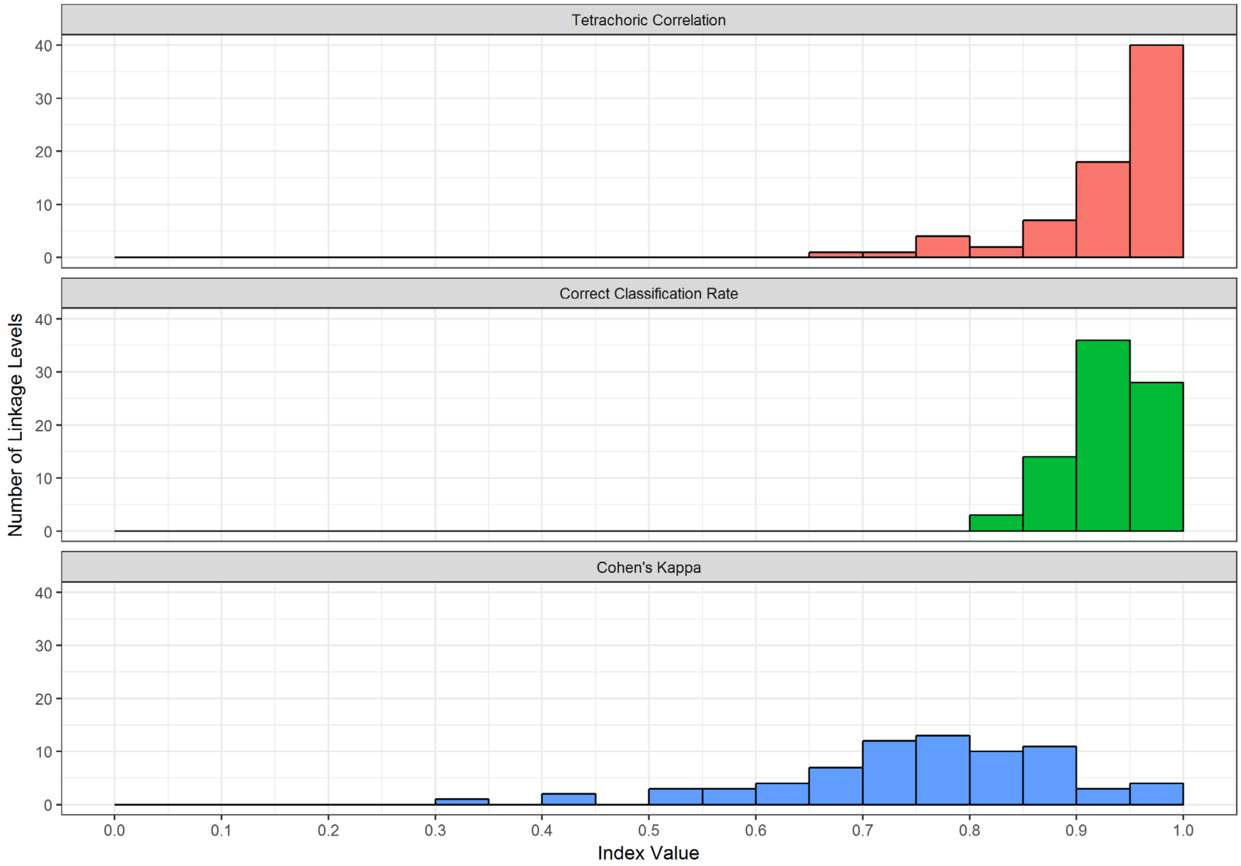


Table 33 and Figure 14 provide proportions and number of linkage levels, respectively, that fall within prespecified ranges of values for the three reliability summary statistics (i.e., correct classification rate, correlation, and kappa). The kappa value and tetrachoric correlation for eight linkage levels could not be computed because all students were labeled as masters of the linkage level.

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

Table 33. Reliability Summaries Across All Linkage Levels: Proportion of Linkage Levels Falling Within a Specified Index Range

Reliability index	Index range								
	< .60	.60–.64	.65–.69	.70–.74	.75–.79	.80–.84	.85–.89	.90–.94	.95–1.0
Tetrachoric correlation	.000	.000	.014	.014	.055	.027	.096	.247	.548
Correct classification rate	.000	.000	.000	.000	.000	.037	.173	.444	.346
Kappa	.123	.055	.096	.164	.178	.137	.151	.041	.055

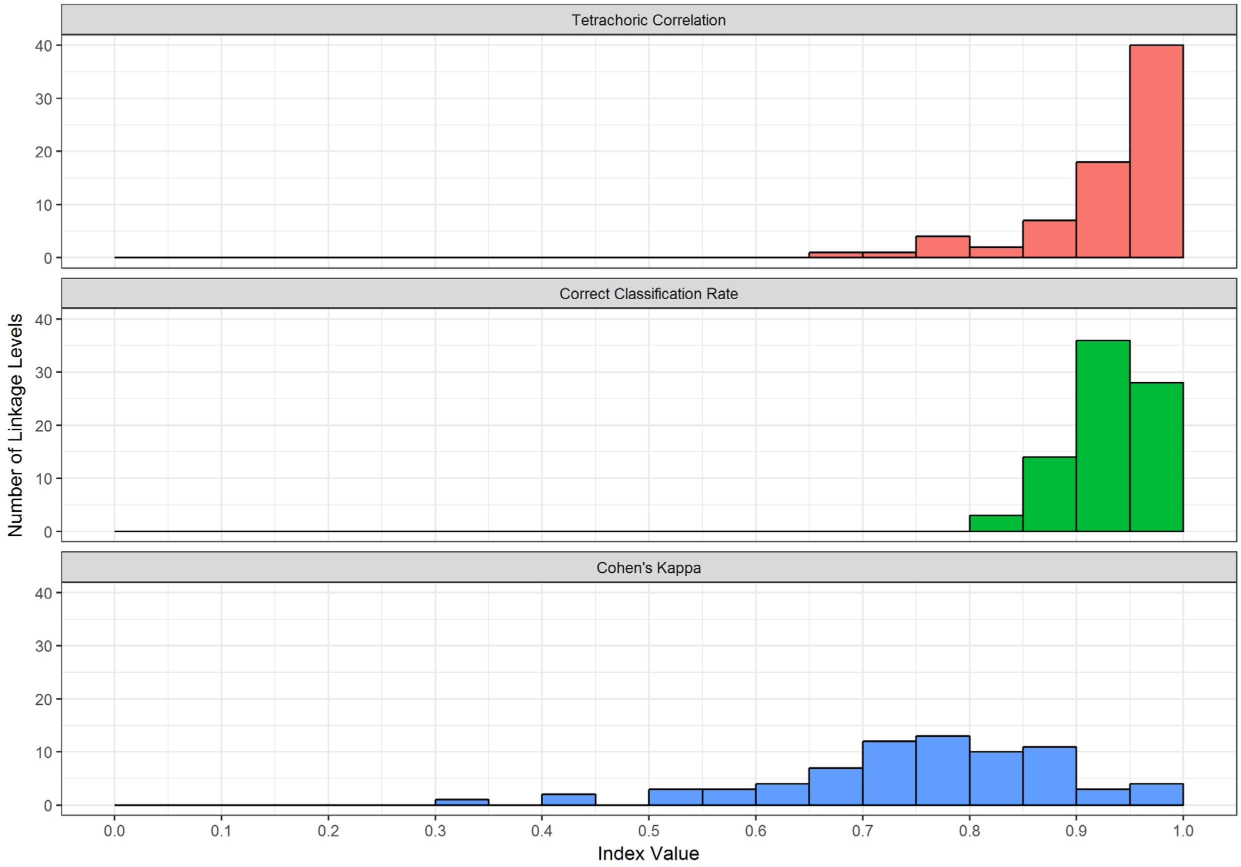


Figure 14. Linkage-level reliability summaries.

VIII.2.B.vi. Conditional Reliability Evidence by Linkage Level

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

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

Figure 15 provides the number of linkage levels that fall within prespecified ranges of values for the three reliability summary statistics (i.e., tetrachoric correlation, correct classification rate, and kappa). The correlations and correct classification rates generally indicate that all three linkage levels provide reliable classifications of student mastery; results are fairly consistent across all linkage levels for each of the three statistics reported.

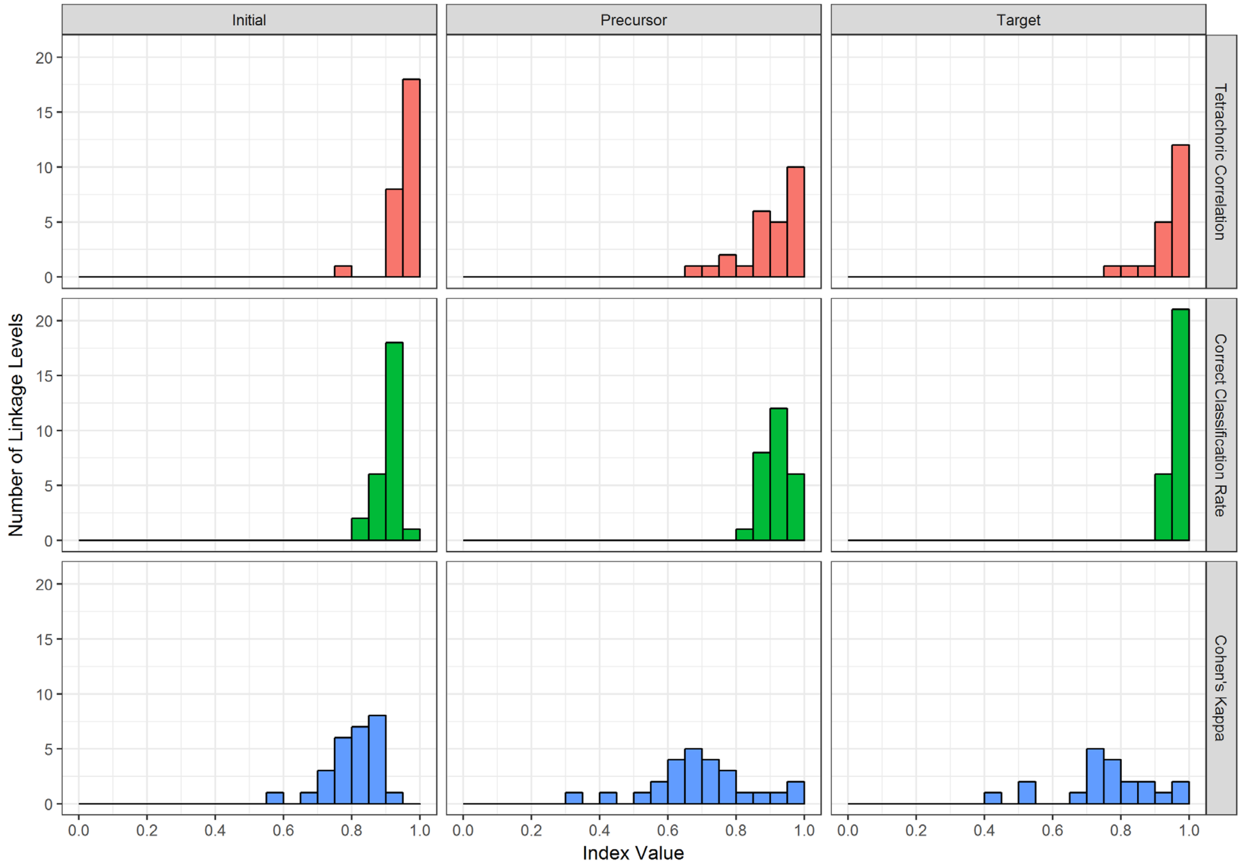


Figure 15. Conditional reliability evidence summarized by linkage level.

VIII.3. Conclusion

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

IX. VALIDITY STUDIES

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

IX.1. Evidence Based on Test Content

Evidence based on test content relates to the evidence "obtained from an analysis of the relationship between the content of the test and the construct it is intended to measure" (AERA et al., 2014, p. 14). The validity study presented in this section summarizes data collected during 2016–2017 regarding student opportunity to learn the assessed content. For additional evidence based on test content, including the alignment of test content to content standards, see Chapter IX of the *2015–2016 Technical Manual – Science* (DLM Consortium, 2017b).

IX.1.A. Opportunity to Learn

After completing administration of the spring 2017 operational assessments, teachers were invited to complete a survey about the assessment administration process (see Chapter IV of this manual for more information on recruitment and response rates). The survey included three blocks of items. The first and third blocks were fixed forms assigned to all teachers. For the second block, teachers received one randomly assigned section.

The survey served several purposes.[9] One item provided preliminary information about the relationship between students' learning opportunities before testing and the test content (i.e., testlets) they encountered on the assessment. The survey asked teachers to indicate the extent to which they judged test content to align with their instruction, across all testlets; Table 34 reports the results. Approximately 50% of teachers (*n* = 6,990) reported that most or all science testlets matched instruction. More specific measures of instructional alignment are planned.

Table 34. Teacher Ratings of Portion of Testlets That Matched Instruction

None		Some (< half)		Most (> half)		All		Did not administer	
n	%	*n*	%	*n*	%	*n*	%	*n*	%
1,454	10.4	4,569	32.6	4,744	33.8	2,246	16.0	1,019	7.3

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

The survey also asked teachers to indicate the approximate number of hours they spent instructing students on each of the DLM science domains and in the science and engineering practices. Teachers responded using a five-point scale: *none*, *1–10 hours*, *11–20 hours*, *21–30 hours*, or *more than 30 hours*.

Table 35 and Table 36 indicate the amount of instructional time spent on DLM science domains and science and engineering practices, respectively. For all three science domains, the most commonly selected responses were 1–10 hours and 11–20 hours. The most commonly selected response for the science and engineering practices was 1–10 hours.

Table 35. Instruction Time Spent on Science Domains, in Hours

	Number of hours									
	0		**1–10**		**11–20**		**21–30**		**>30**	
Domain	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Physical science	454	15.7	1,114	38.6	643	22.3	376	13.0	302	10.5
Life science	352	12.2	939	32.6	714	24.8	443	15.4	436	15.1
Earth and space science	305	10.6	979	34.0	744	25.8	481	16.7	372	12.9

Table 36. Instruction Time Spent on Science and Engineering Practices, in Hours

Science and engineering practices	**Number of hours**									
	0		**1–10**		**11–20**		**21–30**		**>30**	
	n	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Develop and use models	571	19.9	1,202	41.8	586	20.4	320	11.1	195	6.8
Plan and carry out investigations	522	18.3	1,173	41.1	620	21.7	326	11.4	215	7.5
Analyze and interpret data	442	15.5	1,075	37.6	675	23.6	401	14.0	267	9.3
Use mathematics and computational thinking	438	15.3	974	34.1	583	20.4	438	15.3	423	14.8
Construct explanations and design solutions	714	24.9	1,109	38.7	548	19.1	305	10.6	191	6.7
Engage in argument from evidence	951	33.2	1,041	36.4	450	15.7	244	8.5	175	6.1
Obtain, evaluate, and communicate information	477	16.6	1,048	36.6	655	22.9	376	13.1	309	10.8

Results from the teacher survey were also correlated with total linkage levels mastered by domain, as reported on Individual Student Score Reports. While a direct relationship between amount of instructional time and number of linkage levels mastered in the area is not expected, as some students may spend a large amount of time on an area and demonstrate mastery at the lowest linkage level for each Essential Element, it is generally expected that students who mastered more linkage levels in the area would also have spent more instructional time in the area.

Table 37 summarizes the Pearson correlations between domain instructional time and linkage levels mastered in the science domain. Based on guidelines from Cohen (1988), the observed correlations were small.

Table 37. Correlation Between Instruction Time in Science Domain and Linkage Levels Mastered in That Domain

Domain	Correlation with instruction time
Physical science	.20
Life science	.20
Earth and space science	.21

IX.2. Evidence Based on Response Processes

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

IX.2.A. Evaluation of Test Administration

After administering spring operational assessments in 2017, teachers provided feedback via a teacher survey. Survey data that inform evaluations of assumptions regarding response processes include teacher perceptions of students' ability to respond as intended, free of barriers, and with necessary supports available.[10]

One of the fixed-form sections of the spring 2017 teacher survey included three items about students' ability to respond. Teachers were asked to use a 4-point scale (*strongly disagree*, *disagree*, *agree*, or *strongly agree*). Results were combined in the summary presented in Table 38. The majority of teachers agreed or strongly agreed that their students responded to items to the best of their knowledge and ability; were able to respond regardless of disability, behavior, or health concerns; and had access to all supports necessary to participate. The percentage of teachers who agreed or strongly agreed to each statement slightly increased from 2015–2016.

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

Table 38. Teacher Perceptions of Student Experience With Testlets

	Strongly disagree		Disagree		Agree		Strongly agree		Agree or strongly agree	
Statement	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
The student responded to items to the best of their knowledge and ability.	531	3.8	1,029	7.3	7,389	52.3	5,173	36.6	12,562	88.9
The student was able to respond regardless of disability, behavior, or health concerns.	1,001	7.1	1,280	9.1	7,611	53.9	4,231	30.0	11,842	83.9
The student had access to all supports necessary to participate.	373	2.6	479	3.4	7,399	52.3	5,883	41.6	13,282	93.9

IX.2.B. Test-Administration Observations

Test-administration observations were conducted in multiple states during 2016–2017 to further understand student response processes. Students’ typical test-administration process with their actual test administrator was observed. Administrations were observed for the full range of students eligible for DLM assessments (i.e., students with the most significant cognitive disabilities [SCD]). Test-administration observations were collected by DLM project staff, as well as state and local education agency staff.

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

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

Test-administration observations were collected in three states during the 2016–2017 academic year. Table 39 shows the number of observations collected by state.

Table 39. Distribution of Teacher Observations by State (*N* = 32)

State	*n*	%
Kansas	1	3.1
Missouri	15	46.9
West Virginia	16	50.0

Of the 32 test-administration observations collected, 22 (68.8%) were of computer-delivered assessments and 10 (31.3%) were of teacher-administered testlets. All 32 observations were for science testlets; four observations were made for multiple subjects within a single observation.

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 40; behaviors were identified as supporting, neutral, or nonsupporting. For example, clarifying directions (77% 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 such as hand-over-hand guidance (0% of observations) clearly indicates that the teacher directly influenced the student's answer choice.

Table 40. Test Administrator Actions During Computer-Delivered Testlets (*n* = 22)

Evidence	Action	*n*	%
Supporting	Clarified directions or expectations for the student	17	77.3
	Read one or more screens aloud to the student	13	59.1
	Navigated one or more screens for the student	8	36.4
	Repeated question(s) before student responded	7	31.8
Neutral	Asked the student to clarify or confirm one or more responses	2	9.1
	Allowed student to take a break during the testlet	2	9.1
	Repeated question(s) after student responded (i.e., gave a second trial at the same item)	0	0.0
	Used verbal prompts to direct the student's attention or engagement (e.g., "look at this")	7	31.8
	Used pointing or gestures to direct student attention or engagement	7	31.8
	Used materials or manipulatives during the administration process	6	27.3
Nonsupporting	Reduced the number of answer choices available to the student	0	0.0
	Physically guided the student's hand to an answer choice	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 36% of the observations does not necessarily indicate the student was prevented from engaging with the assessment content as independently as possible. Depending on the student, test administrator navigation may either support or minimize students' independent, physical interaction with the assessment system. While not the same as interfering with students' interaction with the content of assessment, navigating for students who are able to do so independently conflicts with the assumption that students are able to interact with the system as intended. The observation protocol did not capture why the test administrator chose to navigate, and the reason was not always obvious from watching.

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 41. Independent response selection was observed in 91% of the cases. Nonindependent response selection may include allowable practices, such as test administrators entering responses for the student. The use of materials outside of KITE Client was seen in 14% 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 SCD, 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 41. Student Actions During Computer-Delivered Testlets (*n* = 22)

Action	***n***	**%**
Selected answers independently	20	90.9
Navigated the screens independently	14	63.6
Selected answers with verbal prompts	7	31.8
Navigated the screens with verbal prompts	4	18.2
Navigated screens after test administrator pointed or gestured	3	13.6
Used materials outside of KITE Client to indicate responses to testlet items	3	13.6
Independently revisited a question after answering it	2	9.1
Skipped one or more items	0	0.0
Revisited one or more questions after verbal prompt(s)	0	0.0

Note. Respondents could select multiple responses to this question.

Another assumption in the validity argument is that students are able to respond to tasks irrespective of sensory, mobility, health, communication, or behavioral constraints. This assumption was evaluated by having observers note whether there was difficulty with accessibility supports (including lack of appropriate, available supports) during observations of teacher-administered testlets. Of the 10 observations of teacher-administered testlets, observers did not note difficulty in any cases (0.0%). For computer-delivered testlets, evidence to evaluate this assumption was collected by noting how students indicated responses to items using multiple response modes, such as eye gaze (0.0%) and using manipulatives or materials outside of KITE Client (13.6%). Additional evidence for this assumption was gathered by observing whether students were able to complete testlets. Of the 32 test-administration observations, students completed the testlet in 31 cases (96.9%).

Another assumption underlying the validity argument is that test administrators enter student responses with fidelity. To record student responses with fidelity, test administrators needed to observe multiple modes of communication, such as verbal, gesture, and eye gaze. Table 42 summarizes students' response modes for teacher-administered testlets. The most frequently observed behavior was the student gestured to indicate a response to test administrator who selected answers.

Table 42. Primary Response Mode for Teacher-Administered Testlets (*N* = 10)

Response mode	***n***	**%**
Gestured to indicate response to test administrator who selected answers	5	50.0
Verbally indicated response to test administrator who selected answers	4	40.0
Used computer/device to respond independently	1	10.0
Eye-gaze system indication to test administrator who selected answers	0	0.0
Used switch system to respond independently	0	0.0
No response	4	40.0

Note. Respondents could select multiple responses to this question.

Computer-delivered testlets provided another opportunity to confirm fidelity of response entry when test administrators entered responses on behalf of students. This support is recorded on the Personal Needs & Preferences Profile and is recommended for a variety of situations (e.g., students who have limited motor skills and cannot interact directly with the testing device even though they can cognitively interact with the onscreen content). Observers recorded whether the response entered by the test administrator matched the student's response. In six of 22 (27.3%) observations of computer-delivered testlets, the test administrator entered responses on the student's behalf. In five (83.3%) of those cases, observers indicated that the entered response matched the student's response, while one observer could not tell. This evidence supports the assumption that test administrators entered student responses with fidelity.

IX.3. Evidence Based on Internal Structure

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

IX.3.A. Evaluation of Item-Level Bias

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

IX.3.A.i. Method

DIF analyses for 2017 followed the same procedure used in 2016, including data from 2015–2016 and 2016–2017 to flag items for evidence of DIF. As additional data are collected in subsequent operational years, the scope of DIF analyses will be expanded to include additional items, subgroups, and approaches to detecting DIF.

Items were selected for inclusion in the DIF analyses based on minimum sample-size requirements for the two gender subgroups: male and female. Within the DLM population, fewer female students responded to items than did male students, by a ratio of approximately 1:2; therefore, a threshold for item inclusion was retained from the previous 2 years whereby at least 100 students in the female group must respond to the item. The threshold of 100 was selected to balance the need for a sufficient sample size in the focal group with the relatively low number of students responding to many DLM items.

Consistent with 2016, additional criteria were included to prevent estimation errors. Items with an overall p value (or proportion correct) greater than .95 were removed from the analyses. Items for which the p value for one gender group was greater than .97 were also removed from the analyses.

Using the above criteria for inclusion, 361 (70%) items on science testlets were selected. In total, 112 items were evaluated in the elementary school grade band, 122 items were evaluated in the middle school grade band, and 127 items were evaluated in the high school grade band. Item sample sizes ranged from 294 to 3,504.

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

$$M_0: \mathit{logit}(\pi_i) = \alpha + \beta X + \gamma_I + \delta_I X$$

$$M_1: \mathit{logit}(\pi_i) = \alpha + \beta X + \gamma_I$$

$$M_2: \mathit{logit}(\pi_i) = \alpha + \beta X;$$

where π_i is the probability of a correct response to the item for group i, X is the matching criterion, α is the intercept, β is the slope, γ_I is the group-specific parameter, and $\delta_I X$ is the interaction term.

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

For each item, the change in the Nagelkerke pseudo R^2 measure of effect size was captured, from M_2 to M_1 or M_0, 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 0.13 and 0.26; values less than 0.13 have a negligible effect, values between 0.13 and 0.26 have a moderate effect, and values of 0.26 or greater have a large effect.

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

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

IX.3.A.ii. Results

IX.3.A.ii.a Uniform Differential Item Functioning Model

A total of 32 items were flagged for evidence of uniform DIF when comparing M_1 to M_2. Table 43 summarizes the total number of items flagged for evidence of uniform DIF by grade band for each model. The percentage of items flagged for uniform DIF for each grade band ranged from 7.1% to 10.7%.

Table 43. Items Flagged for Evidence of Uniform Differential Item Functioning by Grade Band

Grade band	Items flagged (*n*)	Total items (*N*)	Items flagged (%)	Items with moderate or large effect size (*n*)
Elementary	8	112	7.1	0
Middle	13	122	10.7	0
High	11	127	8.7	0

Using the Zumbo and Thomas (1997) effect-size classification criteria, all 32 items were found to have a negligible effect-size change after the gender term was added to the regression equation. Similarly, using the Jodoin and Gierl (2001) effect-size classification criteria, all 32 items were found to have a negligible effect-size change after the gender term was added to the regression equation.

IX.3.A.ii.b Combined Model

A total of 31 items were flagged for evidence of DIF when both the gender and interaction terms were included in the regression equation. Table 44 summarizes the number of items flagged by content area and grade. The percentage of items flagged for each grade and content area ranged from 7.4% to 9.4%.

Table 44. Items Flagged for Evidence of Differential Item Functioning for the Combined Model by Grade Band

Grade band	Items flagged (*n*)	Total items (*N*)	Items flagged (%)	Items with moderate or large effect size (*n*)
Elementary	10	112	8.9	0
Middle	9	122	7.4	0
High	12	127	9.4	0

Using the Zumbo and Thomas (1997) effect-size classification criteria, all 31 items had a negligible change in effect size after adding the gender and interaction terms to the regression equation. Likewise, using the Jodoin and Gierl (2001) effect-size classification criteria, all 31 items were found to have a negligible change in effect size after adding the gender and interaction terms to the regression equation.

While not found in the 2016–2017 administration, if items are flagged for evidence of DIF with either a moderate or large effect size, they are further reviewed by the test-development and psychometric teams. Depending on their review, items may be subject to further analysis (e.g., cognitive labs, panel reviews). Decisions to revise or remove items or testlets are not made based on the results of flagging alone.

IX.3.B. Internal Structure Across Linkage Levels

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

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

When linkage levels perform as expected, masters should have a high probability of providing a correct response and non-masters should have a low probability of providing a correct response. As indicated in Chapter V of this manual, for 102 (100.0%) linkage levels, masters had a greater than .5 chance of providing a correct response to items. Similarly, for 81 (79.4%) linkage levels, non-masters had a less than .5 chance of providing a correct response to items. This finding provides support for how well the linkage levels measured the construct and for the overall validity of inferences that can be made from mastery classifications for the linkage levels.

Chapter III of this manual includes additional evidence of internal consistency in the form of standardized difference figures. Standardized difference values are calculated for operational and field-test items to indicate how far from the linkage-level mean each item's p value falls. Across all linkage levels, 497 (97%) of items fell within two standard deviations of the mean for the linkage level.

These sources of evidence indicate that overall, the linkage levels provide consistent measures of what students know and can do. When linkage levels and the items measuring them do not perform as expected, test-development teams review flags to ensure the content measures the construct as expected.

IX.4. Evidence Based on Consequences of Testing

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

One source of evidence was collected in spring 2017 via teacher-survey responses regarding teacher perceptions of assessment content. An additional study was conducted based on a score-report tutorial to evaluate teachers' interpretation of report contents. Additional consequential evidence, including teacher focus groups on using score-report contents in the subsequent academic year, will be collected in subsequent years.

IX.4.A. Teacher Perception of Assessment Contents

On the spring 2017 survey,[11] teachers were asked three questions about their perceptions of the assessment contents; Table 45 summarizes their responses. Teachers generally responded that content reflected high expectations for their students (82% agreed or strongly agreed), measured important academic skills (70% agreed or strongly agreed), and was similar to instructional activities used in the classroom (70% agreed or strongly agreed). While the majority of teachers agreed with these statements, approximately 20%–30% disagreed. DLM assessments represent a departure from the breadth of academic skills assessed by many states' previous alternate assessments. Given the short history of general curriculum access for this population and the tendency to prioritize the instruction of functional academic skills (Karvonen, Wakeman, Browder, Rogers, & Flowers, 2011), teachers' responses may reflect awareness that DLM assessments contain challenging content. However, teachers were divided on its importance in the educational programs of students with SCD.

[11] Recruitment and sampling are described in Chapter IV of this manual.

Table 45. Teacher Perceptions of Assessment Content

	Strongly disagree		Disagree		Agree		Strongly agree	
Statement	*n*	%	*n*	%	*n*	%	*n*	%
Content measured important academic skills and knowledge for this student.	1,607	11.3	2,623	18.5	8,152	57.4	1,813	12.8
Content reflected high expectations for this student.	752	5.3	1,684	11.9	8,570	60.7	3,120	22.1
Activities in testlets were similar to instructional activities used in the classroom.	1,287	9.1	3,018	21.4	8,081	57.3	1,715	12.2

IX.4.B. Score-Report Interpretation Tutorial

To evaluate teacher interpretation and use of DLM score reports, a study was conducted based on an online tutorial created to support teacher interpretation of score-report contents (Karvonen, Swinburne Romine, Clark, Brussow, & Kingston, 2017). The tutorial included an informed consent portion, followed by pre-test items, the training video, evaluation questions, and a post-test. The video incorporated concepts from the interpretation guide and addressed misconceptions identified in score-report interpretation interviews with teachers. Researchers and DLM item writers familiar with DLM score reports wrote the pre- and post-test questions in the tutorial. Researchers wrote the evaluation questions, which included four Likert-scale items and two open-ended items.

Participating teachers reported a range of confidence in their ability to interpret and use DLM score reports before completing the tutorial; Table 46 summarizes the results. The greatest number of teachers reported being somewhat confident, while the fewest reported being not at all confident.

Table 46. Teacher Confidence in Ability to Interpret and Use Dynamic Learning Maps Score Reports Prior to Tutorial (*N* = 92)

Level of teacher confidence	*n*	%
Very confident	11	12.0
Somewhat confident	33	35.9
Neither confident nor unconfident	25	27.2
Somewhat unconfident	13	14.1
Not at all confident	10	10.9

Following the training video, evaluation questions were presented to the participants; 55 participants responded to these questions. All respondents either strongly agreed (40%) or agreed (60%) that the tutorial covered important information. Most respondents strongly agreed (25%) or agreed (64%) that explanations provided in the tutorial were clear. Most respondents also reported that they felt prepared to explain DLM score-report information to parents (87% agreed or strongly agreed) and to use DLM score reports to inform instruction (80% agreed or strongly agreed).

The evaluation included two open-ended items. The first asked teachers whether they had remaining questions about interpreting DLM score reports. The second asked teachers to indicate additional resources that would help with interpretation and use of DLM score reports. Most teachers reported that they did not have remaining questions about the score reports. Additional feedback included requests for local training and supplemental materials to support instructional planning and decision-making. One participant requested a repository of training videos on different aspects of DLM, which is already available; this request indicates a need to better inform teachers about the resources available. Several participants also requested transcripts and hard copies of the sample reports used in the video, which will be made available online.

Post-test items were included following the evaluation section of the tutorial to prevent performance on the quiz from influencing participant evaluation of the tutorial. Forty-two participants took the post-test. Of those, 18 participants (42.9%) passed (at least 80% accuracy) on their first try. If participants did not respond correctly to 80% of the items, the tutorial was presented again for retaking. Twenty-four participants (57.1%) completed the post-test a second time, two of whom reached the 80% threshold on their second attempt. Ten participants (23.8%) completed the tutorial a third time, but none achieved the passing threshold.

IX.5. CONCLUSION

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

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

X. TRAINING AND INSTRUCTIONAL ACTIVITIES

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

XI. CONCLUSION AND DISCUSSION

The Dynamic Learning Maps® (DLM®) Alternate Assessment System is based on the core belief that all students should have access to challenging, grade-level academic content. Therefore, the DLM assessments provide students with the most significant cognitive disabilities the opportunity to demonstrate what they know and can do.

The DLM science assessment system completed its second operational administration year in 2016–2017. This technical manual update provides updated evidence from the 2016–2017 year intended to support 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 47 and build on the original evidence included in the *2015–2016 Technical Manual – Science* (Dynamic Learning Maps [DLM] Consortium, 2017b). Together, the two documents summarize the validity evidence collected to date.

Table 47. Review of Technical Manual Update Contents

Chapter(s)	Contents
I	Provides an overview of information updated for the 2016–2017 year.
II	Provides an overview of the purpose of the Essential Elements for science, including the intended coverage within the selected organizing structure.
III, IV	Provide procedural evidence collected during 2016–2017 of test content development and administration, including field-test information and teacher-survey results.
V	Describes the statistical model used to produce results based on student responses, along with evidence of model fit.
VI	Not updated for 2016–2017.
VII, VIII	Describe results and analysis of the second operational administration's data, evaluating how students performed on the assessment, the distributions of those results, including aggregated and disaggregated results, and analysis of the internal consistency of student responses.
IX	Provides additional studies from 2016–2017 focused on specific topics related to validity and in support of the score propositions and assessment purposes.
X	Not updated for 2016–2017.

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.

XI.1. Validity Evidence Summary

The accumulated evidence available by the end of the 2016–2017 year provides additional support for the validity argument. Each proposition is addressed by evidence in one or more of the categories of validity evidence, as summarized in Table 48. While many sources of evidence contribute to multiple propositions, Table 48 lists the primary associations. For example, Proposition 4 is indirectly supported by content-related evidence described for Propositions 1 through 3. Table 49 provides the titles and sections for the chapters cited in Table 48.

Table 48. Dynamic Learning Maps Science Alternate Assessment System Propositions and Sources of Updated Evidence for 2016–2017

	Sources of evidence*				
Proposition	**Test content**	**Response processes**	**Internal structure**	**Relations to other variables**	**Consequences of testing**
1. Scores represent what students know and can do.	2, 3, 4, 5, 6, 8, 9, 10, 12	6, 13	3, 4, 7, 11, 14		8, 9, 16
2. Achievement-level descriptors provide useful information about student achievement.	8, 9		11		8, 9, 16
3. Inferences regarding student achievement, progress, and growth can be drawn at the conceptual-area level.	9, 12		11	12	9, 16
4. Assessment scores provide useful information to guide instructional decisions.					16

*See Table 49 for a list of evidence sources. Only direct sources of evidence are included. Some propositions are also supported indirectly by evidence presented for other propositions.

Table 49. Evidence Sources Cited in Previous Table

Evidence no.	Chapter	Section
1	III	English Language Arts Writing Testlets*
2	III	External Reviews
3	III	Operational Assessment Items for 2015–2016
4	III	Field Testing
5	IV	Administration Incidents
6	IV	User Experience with DLM System
7	V	All
8	VII	Student Performance
9	VII	Score Reports
10	VII	Quality Control Procedures for Data Files and Score Reports
11	VIII	All
12	IX	Evidence Based on Test Content
13	IX	Evidence Based on Response Process
14	IX	Evidence Based on Internal Structure
15	IX	Evidence Based on Relation to Other Variables
16	IX	Evidence Based on Consequences of Testing

*Reference relevant only to the DLM English language arts assessment and retained here to preserve common numbering of evidence across DLM technical manuals.

XI.2. Continuous Improvement

XI.2.A. Operational Assessment

As noted previously in this manual, 2016–2017 was the second year the DLM Science Alternate Assessment System was operational. While the 2016–2017 assessments were carried out in a manner that supports the validity of inferences made from results for the intended purposes, the DLM Science Alternate Assessment Consortium is committed to continual improvement of assessments, teacher and student experiences, and technological delivery of the assessment system. Through formal research and evaluation as well as informal feedback, some improvements have already been implemented for 2017–2018. This section describes significant changes from the first to second years of operational administration, as well as examples of improvements to be made during the 2017–2018 year.

Overall, there were no changes to the Essential Elements and linkage levels, item-writing procedures, item flagging outcomes, test administration, or the modeling procedure used to calibrate and score assessments from the previous year to 2016–2017.

Results from the 2016–2017 administration indicated that the majority of students were categorized as either Emerging or Approaching the Target, which was consistent with the results from the 2015–2016 administration. Results will be examined again following the 2017–2018 administration.

Based on an ongoing effort to improve KITE® system functionality, several changes are being implemented during 2017–2018. For instance, new science testlets will be available for use during the instructionally embedded window (similar to the English language arts and mathematics assessments). The spring 2018 administration will also expand availability of braille forms to include Unified English Braille (UEB) in addition to English Braille American Edition (EBAE), which is currently available. Educator Portal will also be enhanced to support creation and delivery of data files and score reports to allow faster delivery timelines. These enhancements include automated creation of all aggregated reports provided at the class, school, district, and state levels; delivery and 2-week review of General Research Files in the interface; on-demand Special Circumstance supplemental files; system-generated exited student files; and, in the event of administration incidents, Incident Files indicating actual impact on students, not potential impact.

The validity evidence collected in 2016–2017 expands upon the data compiled in the first operational year for each of the critical sources of evidence as described in *Standards for Educational and Psychological Testing* (AERA et al., 2014): evidence based on test content, internal structure, response process, relation to other variables, and consequences of testing. Specifically, opportunity to learn contributed to the evidence collected based on test content. Teacher-survey responses on test administration further contributed to the body of evidence collected based on response process, in addition to test administration observations. Evaluation of item-level bias via differential item functioning (DIF) analysis, along with item-pool statistics and model parameters, provided additional evidence collected based on internal structure. Teacher-survey responses provided evidence based on consequences of testing, as well as a score-report interpretation tutorial. Planned studies to provide additional validity evidence for 2017–2018 are summarized in the following section.

XI.2.B. Future Research

The continuous improvement process also leads to future directions for research to inform and improve the DLM Alternate Assessment System in 2017–2018 and beyond. This manual identifies some areas for further investigation.

DLM staff members are planning several studies for spring 2018 to collect data from teachers in the DLM Science Consortium states. The consortium plans to form a set of score-report interpretation focus groups to collect information about how teachers use the 2017 summative score reports to inform instruction in the subsequent academic year. DLM staff will conduct interviews with teachers of students with significant cognitive disabilities who are also English learners to determine how teachers identify students needing services and how to support those students during instruction. Teachers will also be recruited to participate in a study to collect additional evidence based on other variables, whereby teacher ratings of student mastery will be correlated with model-derived mastery. Finally, teacher-survey data collection will also continue during spring 2018 to obtain the second year of data for longitudinal survey items as further validity evidence.

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

Other ongoing operational research is also anticipated to grow as more data become available. For example, DIF analyses will be expanded to include evaluating items across subgroups of interest, as identified by the First Contact survey. Studies on the comparability of results for students who use various combinations of accessibility supports are also dependent upon the availability of larger data sets. This line of research is expected to begin in 2018.

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

REFERENCES

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

Arlot, C. (2010). "A survey of cross-validation procedures for model selection." *Statistics Surveys*. 4, 40-79.

Camilli, G, & Shepard, L.A. (1994). *Methods for identifying biased test items* (4th ed.). Thousand Oaks, CA: Sage.

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

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

Dynamic Learning Maps Consortium. (2016a). *Test Administration Manual 2016-2017*. Lawrence, KS: University of Kansas.

Dynamic Learning Maps Science Consortium. (2016b). *2014-2015 Technical Manual Update – Integrated Model*. Lawrence, KS: University of Kansas.

Dynamic Learning Maps Science Consortium. (2016c). *2016-2017 Technical Manual Update – Integrated Model*. Lawrence, KS: University of Kansas.

Dynamic Learning Maps Science Consortium. (2016d). *2016-2017 Technical Manual Update – Year-End Model*. Lawrence, KS: University of Kansas.

Dynamic Learning Maps Consortium. (2017a). *Educator Portal User Guide*. Lawrence, KS: University of Kansas.

Dynamic Learning Maps Science Consortium. (2017b). *2015–2016 Technical Manual – Science*. Lawrence, KS: University of Kansas.

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

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

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: SAGE.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, New York: Springer-Verlag New York.

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

Karvonen, M., Swinburne Romine, R., Clark, A., Brussow, J., & Kingston, N. (2017). "Promoting accurate score report interpretation and use for instructional planning." National Council on Measurement in Education. San Antonio, TX.

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, ERIC database.

Lancaster, H. O., & Seneta, E. (2005). "Chi-square distribution." *Encyclopedia of Biostatistics*. New York, New York: John Wiley & Sons, LTD. doi: 10.1002/0470011815.b2a15018

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

Maydeu-Olivares, A. & Joe, H. (2006). "Limited information goodness-of-fit testing in multidimensional contingency tables." *Psychometrika*. 71 (713). doi: 10.1007/s11336-005-1295-9

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

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

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

Neyman, J. & Pearson, E.S. "On the problem of the most efficient tests of statistical hypothesis." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*. 231, 289-337.

Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York: Guilford.

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

Templin, J., & Bradshaw, L. P. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79,* 317-339.

Tukey, J.W. (1958). "Bias and confidence in not-quite large samples." *Annals of Mathematical Statistics*, 29, 614-623.

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