



**DYNAMIC**®  
LEARNING MAPS

*2015-2016 Technical Manual*

---

Science

July 2017

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

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

### **Acknowledgements**

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*

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

**Edward Roeber, Ph.D.**, *Assessment Solutions Group/Michigan Assessment Consortium*

**David Williamson, Ph.D.**, *Educational Testing Service*

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

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

**Sue Bechard, Ph.D.**, *Senior Advisor*

**Brooke Nash, Ph.D.**, *Psychometrician Senior*

**Meagan Karvonen, Ph.D.**, *Director*

**Russell Swinburne Romine, Ph.D.**, *Associate Director for Test Development and Production*

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 Andersen

Neal Kingston

Jonathan Templin

Brianna Beitling

Allison Lawrence

Susan K. Thomas

Jennifer Brussow

Lee Ann Mills

Jacob Thompson

Amy Clark

Michael Muenks

Lisa Weeks

Nora Fairchild

Lindsay Ruhter

Sheila Wells-Moreaux

Lisa Harkrader

Michelle Shipman

### Revision History

Date	Revision
2019-06-14	Several edits were made throughout this manual to clarify the distinction between the development work of the 2015-2016 science assessment (previously referred to as Phase I) and future development work that will eventually support a new science assessment (previously referred to as Phase II). Specifically, language was removed that may have incorrectly implied that the two distinct development efforts are two phases of work for the same assessment.

## Table of Contents

<b>I. Introduction .....</b>	<b>1</b>
I.1. Background .....	1
I.1.A. Student Population .....	4
I.1.B. Theory of Action .....	6
I.1.C. Key Elements .....	8
I.2. System Components .....	9
I.2.A. Essential Elements and Linkage Levels .....	10
I.2.B. Assessments .....	11
I.3. Technical Manual Overview .....	11
<b>II. Essential Element Development .....</b>	<b>14</b>
II.1. Purpose of Essential Elements for Science .....	14
II.1.A. Grade-Level Science Content Standards .....	14
II.1.B. Alternate Science Content Standards Crosswalk .....	16
II.2. Development of the Essential Elements for Science .....	20
II.2.A. Linkage Levels .....	20
II.2.B. Codes for Essential Elements .....	21
II.2.C. Development Process .....	21
II.3. Science Blueprint Development .....	30
II.3.A. Options Development and Selection .....	31
II.3.B. Final Science Blueprint .....	36
II.4. Conclusion .....	37
<b>III. Item and Test Development .....</b>	<b>38</b>
III.1. Review of Science Assessment Structure .....	38
III.1.A. Items and Testlets .....	40
III.2. Essential Element Concept Maps for Testlet Development .....	45
III.3. Item Writing .....	46
III.3.A. Recruitment and Selection .....	46
III.3.B. Item Writer Characteristics .....	47
III.3.C. Item Writer Training .....	49
III.3.D. Item Writing Resource Materials .....	50
III.3.E. Item Writing Process .....	50
III.3.F. Item Writer Evaluations .....	51
III.4. External Reviews .....	56
III.4.A. Overview of Review Process .....	57
III.4.B. Review Assignments and Training .....	58
III.4.C. Reviewer Responsibilities .....	58
III.4.D. Decisions and Criteria .....	58
III.4.E. Results of Reviews .....	63
III.5. The First Contact Survey .....	63
III.6. Pilot Administration .....	64
III.7. 2015 Fall Field Test .....	67
III.7.A. Field Test Survey .....	73
III.8. Operational Assessment Items for 2015-2016 .....	78

III.9. Conclusion.....	81
<b>IV. Test Administration.....</b>	<b>82</b>
IV.1. Overview of Key Administration Features.....	82
IV.1.A. The Year-end Assessment Model .....	83
IV.1.B. Assessment Delivery Modes .....	83
IV.1.C. The KITE System .....	100
IV.1.D. Adaptive Delivery .....	102
IV.1.E. Special Circumstance Codes .....	105
IV.2. Test Administration .....	105
IV.2.A. Test Windows.....	106
IV.2.B. Administration Time.....	106
IV.2.C. Resources and Materials .....	107
IV.2.D. Test Administrator Responsibilities and Procedures .....	111
IV.2.E. Monitoring Assessment Administration.....	112
IV.3. Accessibility Supports .....	114
IV.3.A. Overview of Accessibility Supports .....	115
IV.3.B. Additional Allowable Practices .....	119
IV.4. Security .....	121
IV.4.A. Training and Certification .....	122
IV.4.B. Maintaining Security During Test Administration.....	122
IV.4.C. Security in the KITE System .....	123
IV.4.D. Secure Test Content .....	124
IV.4.E. Data Security .....	124
IV.4.F. State-Specific Policies and Practices .....	124
IV.4.G. Forensic Analysis Plans.....	125
IV.5. Implementation Evidence from 2015–2016 Test Administration .....	126
IV.5.A. Adaptive Delivery Implementation Evidence .....	126
IV.5.B. Administration Errors.....	128
IV.5.C. User Experience with Assessment Administration and KITE System .....	128
IV.6. Conclusion .....	135
<b>V. Modeling.....</b>	<b>136</b>
V.1. Psychometric Background.....	136
V.2. Essential Elements and Linkage Levels.....	137
V.3. Overview of DLM Modeling Approach .....	138
V.3.A. DLM Model Specification .....	138
V.3.B. Model Calibration.....	139
V.4. DLM Scoring: Mastery Status Assignment .....	141
V.5. Conclusion.....	142
<b>VI. Standard Setting.....</b>	<b>144</b>
VI.1. Standard Setting Overview .....	144
VI.1.A. Standard Setting Approach: Rationale and Overview .....	144
VI.1.B. Policy Performance Level Descriptors.....	146
VI.1.C. Profile Development.....	147
VI.1.D. Panelists .....	149

VI.1.E. Meeting Procedures .....	150
VI.2. Results.....	153
VI.2.A. Panel-Recommended and Adjusted Cut Points .....	153
VI.2.B. Vertical Articulation Panel Process.....	154
VI.2.C. DLM Staff–Recommended Cut Points and Impact Data.....	154
VI.2.D. External Evaluation of Standard Setting Process and Results.....	157
VI.3. Grade-Level Performance Level Descriptors .....	158
<b>VII. Assessment Results .....</b>	<b>160</b>
VII.1. Student Participation.....	160
VII.2. Student Performance .....	162
VII.2.A. Overall Performance .....	163
VII.2.B. Subgroup Performance.....	164
VII.2.C. Linkage Level Mastery.....	165
VII.3. Data Files.....	166
VII.4. Score Reports .....	167
VII.4.A. Individual Reports .....	167
VII.4.B. Aggregated Reports.....	169
VII.4.C. Interpretation Resources.....	170
VII.4.D. Quality Control Procedures for Data Files and Score Reports.....	171
VII.5. Conclusion .....	174
<b>VIII. Reliability .....</b>	<b>175</b>
VIII.1. Background Information on Reliability Methods .....	175
VIII.1.A. Methods of Obtaining Reliability Evidence .....	178
VIII.2. Reliability Evidence.....	180
VIII.2.A. Performance Level Reliability Evidence .....	181
VIII.2.B. Content-Area Reliability Evidence.....	182
VIII.2.C. Domain Reliability Evidence .....	183
VIII.2.D. Essential-Element Reliability Evidence.....	185
VIII.2.E. Linkage-Level Reliability Evidence.....	187
VIII.2.F. Conditional Reliability Evidence by Linkage-Level.....	190
VIII.3. Conclusion.....	191
<b>IX. Validity Studies .....</b>	<b>192</b>
IX.1. Evidence Based on Test Content .....	192
IX.1.A. External Alignment Study .....	192
IX.1.B. Opportunity to Learn.....	199
IX.2. Evidence Based on Response Processes .....	201
IX.2.A. Evaluation of Test Administration .....	201
IX.3. Evidence Based on Internal Structure.....	206
IX.3.A. Evaluation of Item-Level Bias .....	206
IX.4. Evidence Based on Relations to Other Variables .....	209
IX.5. Evidence Based on Consequences of Testing .....	211
IX.5.A. DLM Score Report Design and Use.....	211
IX.5.B. Teacher Resources .....	217
IX.5.C. Baseline Test Administrator Survey Responses.....	220

IX.6. Conclusion.....	220
<b>X. Training and Instructional Activities.....</b>	<b>221</b>
X.1. Training for State Education Agency Staff.....	221
X.1.A. Training for Local Education Agency Staff.....	221
X.2. Required Training for Test Administrators .....	222
X.2.A. Facilitated Training .....	224
X.2.B. Self-Directed Training .....	224
X.2.C. Training Content.....	225
X.3. Instructional Activities.....	228
<b>XI. Conclusion and Discussion .....</b>	<b>230</b>
XI.1. Validity Framework.....	231
XI.2. Propositions for Score Interpretation and Use .....	232
XI.3. Summary and Evaluation of Validity Evidence.....	232
XI.3.A. Proposition 1: Scores represent what students know and can do .....	233
XI.3.B. Proposition 2: Achievement level descriptors provide useful information about student achievement.....	240
XI.3.C. Proposition 3: Inferences regarding student achievement, progress and growth can be drawn at the domain level.....	242
XI.3.D. Proposition 4: Assessment scores provide useful information to guide instructional decisions .....	243
XI.3.E. Evaluation Summary.....	245
XI.4. Continuous Improvement.....	249
XI.4.A. Operational Assessment .....	249
XI.4.B. Future Research .....	250

## List of Tables

Table 1. Dynamic Learning Maps Participation Guidelines .....	4
Table 2. Common Science Standards Assessed by DLM States Organized by Framework Disciplinary Core Ideas and Sub-Ideas.....	18
Table 3. Example of Next Generation Science Standards Performance Expectations Related to States’ Alternate Science Standards.....	19
Table 4. Count of Standards by Grade Band Addressed in Essential Element Development .....	20
Table 5. Comparison Between English Language Arts/Mathematics and Science Linkage Levels .....	21
Table 6. Timeline for the Development of the Science DLM Essential Elements.....	22
Table 7. Percentage of Participant Ratings by Level of Agreement for Evaluation Items (N = 33) .....	27
Table 8. Count of Essential Elements Included in Science Blueprints for 2014–2018.....	37
Table 9. Item Writers’ Years of Teaching Experience .....	47
Table 10. Item Writers’ Grade-Level Teaching Experience .....	48
Table 11. Item Writers’ Level of Degree.....	48
Table 12. Item Writers’ Experience with Disability Categories .....	49
Table 13. Perceived Effectiveness of Training for January 2015 Workshop (n = 39) .....	52
Table 14. Perceived Effectiveness of Training for July 2015 Workshop (n =15) .....	52
Table 15. Overall Experience for January 2015 Workshop (n = 39).....	53
Table 16. Overall Experience for July 2015 Workshop (n = 15).....	54
Table 17. General Review Decisions for External Reviews .....	59
Table 18. Number of Participants in the Spring 2015 Science Pilot Test by Grade Band.....	65
Table 19. Item Flags for Content Administered During the 2015 Science Spring Pilot Test .....	66
Table 20. Content Team Response to Item Flags for the 2015 Science Spring Pilot Test .....	67
Table 21. 2015 Science Fall Field Test Sampling Design Example – Life Science.....	68
Table 22. Number of Testlets by Grade Band for Fall 2015 Field Test.....	69
Table 23. Number of Participants in the Fall 2015 Science Field Test by Grade Band .....	70
Table 24. Demographic Summary of Students Participating in the Fall 2015 Science Field Test .	70
Table 25. Item Flags for Content Administered During the 2015 Science Fall Field Test.....	72
Table 26. Content Team Response to Item Flags for the 2015 Science Fall Field Test.....	72



Table 27. Average Proportion Correct on Testlets by Complexity Band for each Grade Band ....	73
Table 28. Demographic Summary of Students Whose Educators Participated in the Science Field Test Survey.....	74
Table 29. Perceived Consistency of Student Skill during Science Instruction.....	75
Table 30. Personal Needs and Preferences Profile (PNP) Features That Met Students' Accessibility Needs (N=837).....	76
Table 31. Factors That Negatively Impacted Students' Assessment Experience (N=837) .....	77
Table 32. Factors That Positively Impacted Students' Assessment Experience (N=837) .....	78
Table 33. Operational Window Participation.....	78
Table 34. 2015–16 Science Operational Testlets .....	79
Table 35. Correspondence Between Complexity Band and Assigned Linkage Level.....	103
Table 36. Distribution of Response Times in Minutes for Initial Level Testlets.....	106
Table 37. Distribution of Response Times in Minutes for Precursor and Target Level Testlets .	107
Table 38. DLM Resources for Test Administrators and States .....	108
Table 39. DLM Resources for Test Administration Monitoring Efforts.....	113
Table 40. Accessibility Supports in the DLM Assessment System.....	115
Table 41. Additional Allowable Practices.....	120
Table 42. Correspondence of Complexity Bands and Linkage Level .....	126
Table 43. Adaptation of Linkage Levels Between the First and Second Testlets.....	127
Table 44. Number of Students Affected by Each 2016 Incident .....	128
Table 45. Educator Responses Regarding Test Administration (N=1,407 unless otherwise stated) .....	129
Table 46. Ease of Using KITE Client (N = 1,407).....	130
Table 47. Ease of Using Educator Portal (N = 1,407) .....	131
Table 48. Overall Experience with KITE Client and Educator Portal (N = 1,407).....	132
Table 49. Personal Needs and Preferences (PNP) Supports Selected for Students, Spring 2016 (N = 22,010) .....	132
Table 50. Teacher Report of Student Accessibility Experience (Year-End Model).....	134
Table 51. Depiction of Fungible Item Parameters for Items Measuring a Single Linkage Level	139
Table 52. Final Performance Level Descriptors for the DLM Consortium .....	147
Table 53. Demographic Characteristics of Panelists .....	150

Table 54. Panelists’ Years of Experience .....	150
Table 55. Panel Cut-Point Recommendations .....	153
Table 56. Adjusted Cut-Point Recommendations .....	154
Table 57. DLM Staff–Recommended Cut Points for Science .....	155
Table 58. Demographic Information for Students Included in Impact Data .....	156
Table 59. Student Participation by State or Agency .....	160
Table 60. Student Participation by Grade .....	161
Table 61. Demographic Characteristics of Participants .....	162
Table 62. Percentage of Students by Grade and Performance Level (n = 20,214) .....	163
Table 63. Students at Each Performance Level by Demographic Group (n = 20,214) .....	164
Table 64. Percentage of Students Demonstrating Highest Linkage Level Mastered Across EEs, by Grade/Course .....	166
Table 65. Summary of Performance Level Reliability Evidence.....	182
Table 66. Summary of Content Area Reliability Evidence .....	183
Table 67. Summary of Science Domain Reliability Evidence.....	184
Table 68. Reliability Summaries Across All EEs: Proportion of EEs Falling Within a Specified Index Range .....	186
Table 69. Example of True and Estimated Mastery Status from Reliability Simulation.....	188
Table 70. Reliability Summaries Across All Linkage Levels: Proportion of Linkage Levels Falling Within a Specified Index Range .....	189
Table 71. Percentage of Essential Element Ratings Which Met Each Criterion .....	195
Table 72. Percentage of Linkage Level Transition Ratings Which Met the Criterion.....	196
Table 73. Percentage of Testlet Items Which Met Each Criterion .....	197
Table 74. Average Number of Hours Spent Instructing Science Topics .....	200
Table 75. Science Practices in Which the Student Was Instructed (N=837) .....	201
Table 76. Teacher Observations by State (N = 37).....	202
Table 77. Test Administrator Actions During Computer-Delivered Testlets (N = 29).....	203
Table 78. Student Actions during Computer-Delivered Testlets (N = 29) .....	204
Table 79. Primary Response Mode for Teacher-Administered Testlet (N = 8).....	205
Table 80. Test Administrator Perceptions of Student Experience with Assessments, Spring 2016 .....	206

Table 81. Items Flagged for Evidence of Uniform DIF .....	209
Table 82. Items Flagged for Evidence of DIF for the Combined Model.....	209
Table 83. Correlations of Total Linkage Levels Mastered in Science with English Language Arts and Mathematics .....	210
Table 84. Correlations of Total Linkage Levels Mastered with Selected Demographic Characteristics .....	210
Table 85. Components of the DLM Individual Student Score Report.....	213
Table 86. Review of Technical Manual Contents.....	230
Table 87. Dynamic Learning Maps Science Alternate Assessment System Propositions and Sources of Related Evidence for 2015-16 .....	246
Table 88. Evidence Sources Cited in Previous Table.....	247
Table 89. Evaluation of Evidence for Each Proposition.....	248

## List of Figures

Figure 1. Timeline for the DLM science development project.....	4
Figure 2. Dynamic Learning Maps theory of action for science.....	7
Figure 3. Design of the DLM science assessment.....	10
Figure 4. Example of four states' content standards for physical properties.....	17
Figure 5. Essential Element with linkage levels developed by expert panel and connections noted to ELA and mathematics.....	23
Figure 6. Essential Element review checklist.....	25
Figure 7. External review panel revisions to EE.5.PS1-3.....	26
Figure 8. Final approved EE.5.PS.1-3.....	30
Figure 9. Design of the DLM science assessment.....	39
Figure 10. Example science story (EE.HS.PS3-4).....	44
Figure 11. Overview of the item review processes prior to field testing for the DLM Alternate Assessment System.....	57
Figure 12. Content review criteria.....	60
Figure 13. Accessibility review criteria.....	61
Figure 14. Bias and sensitivity review criteria.....	62
Figure 15. P-value for science operational items.....	80
Figure 16. Standardized difference z scores for science operational items.....	81
Figure 17. Computer-delivered released testlet – Opening screen with test directions and navigation buttons.....	85
Figure 18. Computer-delivered released testlet – Science story 1.....	86
Figure 19. Computer-delivered released testlet – Science story 1 (continued).....	87
Figure 20. Computer-delivered released testlet – Item 1.....	88
Figure 21. Computer-delivered released testlet – Item 2.....	89
Figure 22. Computer-delivered released testlet – Science story 2.....	90
Figure 23. Computer-delivered released testlet – Science story 2 (continued).....	91
Figure 24. Computer-delivered released testlet – Item 3.....	92
Figure 25. Teacher-administered released testlet – General educator directions.....	94
Figure 26. Teacher-administered released testlet – Educator directions for Item 1.....	95
Figure 27. Teacher-administered released testlet – Student response record for Item 1.....	96

Figure 28. Teacher-administered released testlet – Educator directions for Item 2. ....	97
Figure 29. Teacher-administered released testlet – Student response record for Item 2. ....	98
Figure 30. Teacher-administered released testlet – Educator directions for Item 3. ....	99
Figure 31. Teacher-administered released testlet – Student response record for Item 3. ....	100
Figure 32. An example screen from the student interface in KITE Client. ....	102
Figure 33. Linkage level adaptations for a student who completed five testlets. ....	105
Figure 34. Test security agreement text. ....	122
Figure 35. EE and linkage levels for SCI.EE.5.LS1-1 (fifth grade science). ....	138
Figure 36. Linkage level mastery assignment by mastery rule for each science grade bands. ...	142
Figure 37. Steps of the DLM standard-setting process. ....	146
Figure 38. Example standard setting profile for a hypothetical student. ....	148
Figure 39. Science impact data using DLM staff–recommended cut points. ....	156
Figure 40. Page one of the performance profile for 2015-2016. ....	169
Figure 41. Simulation process for creating reliability evidence. ....	180
Figure 42. Number of linkage levels mastered within EE reliability summaries. ....	187
Figure 43. Linkage-level reliability summaries. ....	189
Figure 44. Conditional reliability evidence summarized by linkage level. ....	191
Figure 45. Design of the DLM science assessment. ....	193
Figure 46. Required training processes flows for facilitated and self-directed training. ....	223

## I. INTRODUCTION

The Dynamic Learning Maps® (DLM®) Alternate Assessment System assesses student achievement in English language arts, mathematics, and science for students with the most significant cognitive disabilities (SCD) in grades 3-8 and high school. This manual describes the development and technical aspects of the DLM alternate assessment in science for the 2015-2016 school year. The purpose of the system is to improve academic experiences and outcomes for students with SCD by setting high and actionable academic expectations and providing appropriate and effective supports to educators.

Results from the DLM alternate assessment are intended to support interpretations about what students know and are able to do as well as 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 appropriate for state accountability programs to use. Results are not intended to support the determination of disability eligibility, placement, retention, graduation, or to directly compare with scores on general education assessments.

The DLM Alternate Assessment System is based on the core belief that all students should have access to challenging, grade-level content. The assessment system makes use of online adaptive assessments delivered directly to the student, teacher-administered assessments with online input of student responses, and a range of accessibility features and allowable practices to give students with SCD opportunities to demonstrate what they know in ways that traditional multiple-choice assessments cannot. A year-end summative assessment in science is administered in the spring, and results from that assessment are reported to states for their use in accountability programs and program improvement for the following school year.

This chapter describes the foundations of the DLM Alternate Assessment System, including the background, history, purpose, and key characteristics of the program. This chapter lays the groundwork for subsequent chapters on the DLM science assessment design, assessment development and administration, psychometric modeling, standard setting, reporting, reliability and validity, and overall evaluation. An overview of subsequent chapters is included at the end of this chapter. While these chapters focus on essential components of the assessment system separately, several key topics are included in multiple chapters throughout this manual, including accessibility and validity.

### I.1. BACKGROUND

The DLM science alternate assessment is a separate, state-funded addition to the DLM Alternate Assessment System that was created in 2010, when the U.S. Department of Education's Office of Special Education Programs awarded a five-year General Supervision Enhancement Grant to the DLM Consortium. The grant, which supported development of new alternate assessments in English language arts (ELA) and mathematics, was overseen by the Center for Educational Testing and Evaluation (CETE) in the Achievement and Assessment Institute at the University of Kansas.

The DLM science alternate assessment builds on the processes, products, and lessons learned from the development of the grant-funded ELA and mathematics assessments. These initial assessments were developed by a consortium of state education agencies (SEAs). In 2010, 13 SEAs were involved and by the end of the fifth year (2015), there were 16 member states in the DLM ELA and mathematics consortium.

In addition to CETE and partner states, other key partners during the grant-funded project included the Center for Literacy and Disability Studies at the University of North Carolina at Chapel Hill, which provided professional development materials; Edvantia, which merged with Mid-continent Research for Education and Learning during the project and served as the project’s external evaluator; The Arc, which assisted with gathering parent feedback to DLM student reports and parent materials; and the Center for Research Methods and Data Analysis at the University of Kansas, which provided programs for generating score reports. The project was also supported by a technical advisory committee and a special education advisory committee.

Detailed information about the development and operationalization of the ELA and mathematics assessments is available in the 2014-2015 technical manuals and numerous technical reports at <http://dynamiclearningmaps.org/about/research/publications>. The purpose of this 2015-2016 Technical Manual for Science is to document the development of the first DLM science operational assessment.

Alternate assessments in science are largely state-specific, which has resulted in large variations in science content for students with SCD across the U.S. (Rogers, Thurlow, and Lazarus, 2015). However, in 2014, five DLM member states—Kansas, Oklahoma, Missouri, Mississippi, and Iowa—decided to jointly self-fund the development of a science assessment following the DLM model with the intent of administering an operational assessment by 2016. Given the short timeline, the states decided to focus on the development and administration of a year-end operational assessment for three grade bands and an end-of-course high school biology assessment. These assessments were based on Essential Elements (EEs), which are alternate content standards in three grade bands: elementary, grades 3-5, which uses the grade 5 content standards; middle school, which uses the 6-8 grade band content standards; and high school, which uses the 9-12 grade band content standards (see Goal 1 below). Alternate achievement standards were set for each grade in which one or more states test science: 4, 5, 6, 8, and high school.

The science assessment system parallels the existing ELA and mathematics assessments in many ways, including testlet design and delivery, policy performance level descriptors (PLD), scoring and reporting, and reliability and validity. Yet there are some differences between science and existing ELA and mathematics systems (e.g., assessments available at three levels of cognitive complexity per content standard instead of five).

Future development for the DLM science assessment system includes development of a learning map model for science (which would be followed by additional assessments developed



based on the map), professional development products, and instructionally embedded assessments.

The focus of this 2015-2016 Technical Manual for Science is the development work, which resulted in the first operational assessment in 2016. There were three goals for the DLM Science project.

- **Goal 1:** To link the Science Consortium state partners’ alternate content standards in science to the National Research Council’s *Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (National Research Council, 2012; *Framework*) and the Next Generation Science Standards (2013; NGSS) as a framework for developing EEs. The science EEs were intended to reflect the concepts that were currently assessed in the partner states as well as reflect the multidimensional components of the *Framework*.
- **Goal 2:** To develop an operational adaptive computerized assessment system for science based on the EEs by spring 2016.
- **Goal 3:** To develop and apply cut points based on achievement level descriptors that describe what students with SCD should know and be able to do.

All three stated goals were met. Figure 1 summarizes the timeline and milestones of the DLM Science project.

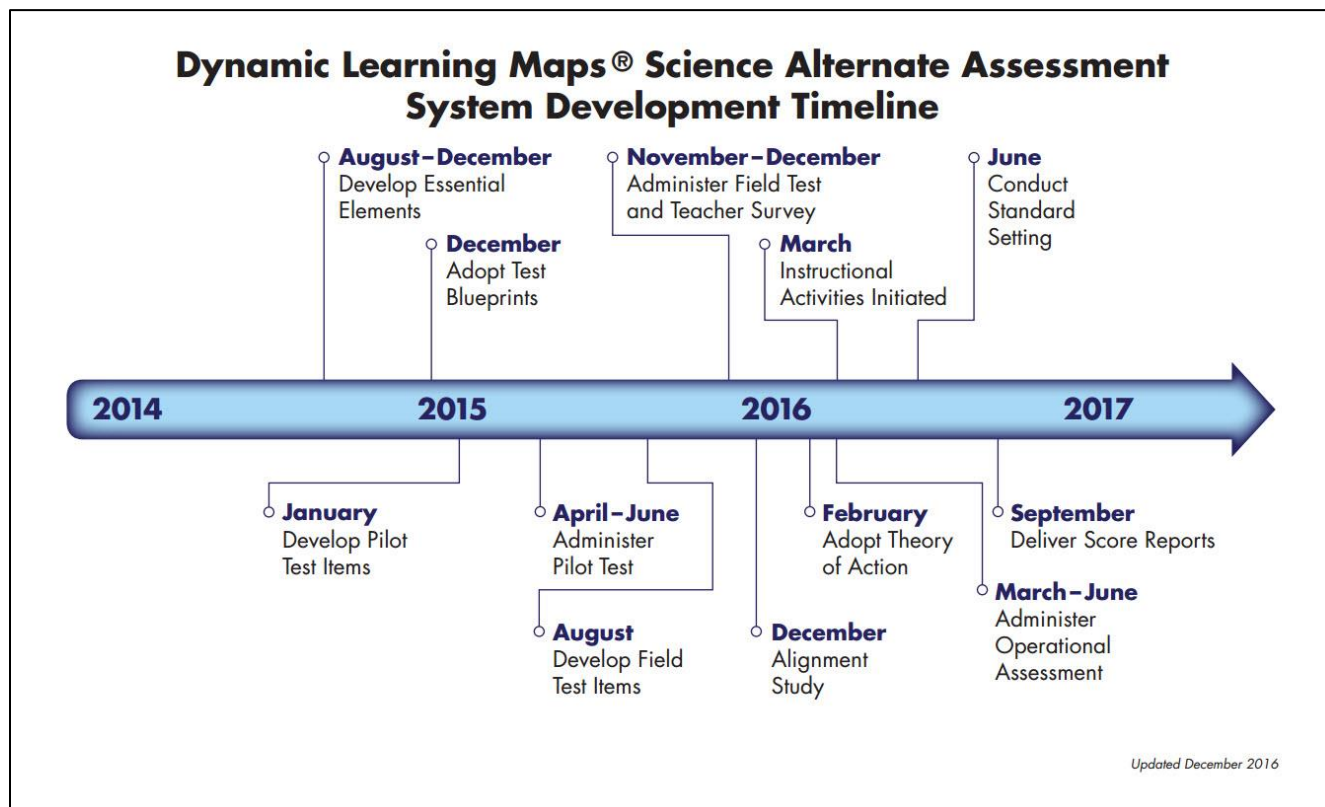




Figure 1. Timeline for the DLM science development project.

The subset of states in the DLM Consortium that administer the science assessments guide the DLM science assessment design and development. This group is described as the DLM Science Consortium throughout the manual.

### ***I.1.A. STUDENT POPULATION***

The DLM Alternate Assessment System serves students with SCD who are eligible to take their state’s alternate assessment based on alternate academic achievement standards. This population is, by nature, diverse in learning style, communication mode, support needs, and demographics. The participation guidelines adopted by DLM states in 2013 are used for all DLM assessments, as described below.

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

The DLM Alternate Assessment System provides the opportunity for students with SCD to show what they know. These are students for whom general education assessments, even with accessibility features or supports, are not appropriate. These students learn academic content aligned to grade-level content standards, but at reduced depth, breadth, and complexity. As described in Chapter II, the EEs, derived from the *Framework* and the NGSS, are the learning targets for the DLM assessments for the grade bands at the elementary, middle school, and high school levels, plus end-of-instruction high school biology.

While all states provide additional interpretation and guidance to their districts, three general participation guidelines are considered for a student to be eligible for the DLM alternate assessment. All three criteria must be met and are outlined in Table 1 below.

Table 1. Dynamic Learning Maps Participation Guidelines

<b>Participation Criterion</b>	<b>Participation Criterion Descriptors</b>
1. The student has a significant cognitive disability.	Review of student records indicate a disability or multiple disabilities that significantly impact intellectual functioning and adaptive behavior.*

Participation Criterion	Participation Criterion Descriptors
2. The student is primarily being instructed (or taught) using the DLM EEs as content standards.	Goals and instruction listed in the IEP for this student are linked to the enrolled grade-level DLM EEs and address knowledge and skills that are appropriate and challenging for this student.
3. The student requires extensive direct individualized instruction and substantial supports to achieve measureable gains in the grade- and age-appropriate curriculum.	<p>The student</p> <ul style="list-style-type: none"> <li>a. requires extensive, repeated, individualized instruction and support that is not of a temporary or transient nature, and</li> <li>b. uses substantially adapted materials and individualized methods of accessing information in alternative ways to acquire, maintain, generalize, demonstrate, and transfer skills across multiple settings.</li> </ul>

\*Note: Adaptive behavior is defined as essential for someone to live independently and to function safely in daily life.

The DLM Alternate Assessment System eligibility guidelines also specify characteristics that, on their own, are not sufficient for determining student participation in the alternate assessment, such as

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

### ***I.1.B. THEORY OF ACTION***

The theory of action that guided the design of the science assessment was similar to the DLM Alternate Assessment System for ELA and mathematics, finalized in December 2013. The original theory of action expresses the belief that high expectations for students with SCD, when combined with appropriate educational supports and diagnostic tools for educators, result in improved academic experiences and outcomes for students, educators, and parents or guardians.

The process of articulating the theory of action started with identifying critical problems that characterize large-scale assessment of students with SCD so that the DLM Alternate Assessment System design could help alleviate these problems. Critical problems included how best to capture the multidimensional nature of teaching and learning, how best to allow for non-linear approaches to demonstrating learning, how best to support best practices in instruction without replacing it with assessment preparation, and how best to avoid negative unintended consequences for students. The DLM theory of action expresses a commitment to provide students with SCD access to flexible cognitive and learning pathways and an assessment system that is capable of validly and reliably evaluating their progress and achievement. Ultimately, the goal is for educators, parents/guardians, and others to hold higher expectations of students and improve their educational experiences.

After identifying these overall guiding principles and anticipated outcomes, specific elements of the DLM Alternate Assessment System theory of action were articulated to inform assessment design and to highlight the associated validity arguments. The theory elements were organized around four main topics: precursors to assessment development and implementation, assessment features, score interpretation and use, and goals of the assessment system.

The DLM theory of action was modified and adopted by the Science Consortium in February, 2016 (see Figure 2). Modifications included the removal of references to the learning map models and instructionally embedded assessment.



THEORY OF ACTION: Science Assessment Design

**PRECURSORS**

- Alternate content standards, the Essential Elements, provide grade level access to NGSS and prepare students for college, career, and citizenship
- The system used to deliver DLM assessments is designed to maximize accessibility
- The linkage levels represent the Essential Elements at appropriate access points for SWSCD
- Educators understand the personal needs and preferences of their students and correctly document the students' needs within the assessment system
- Teachers provide instruction aligned with Essential Elements and at a level of complexity that provides an appropriate level of challenge
- Parents and teachers have high expectations regarding what students are able to achieve
- Students know how to interact with the assessment system

**ASSESSMENT**

- Testlets presented to the student align to the Essential Element and are free from construct irrelevant variance
- The end of year assessments have been designed to allow students to demonstrate their knowledge and skills in relation to academic expectations
- The combination of testlets administered at the end of the year measure knowledge and skills at the appropriate breadth, depth, and complexity of the content
- Teachers administer the end of year assessments with fidelity so that students can respond to the items as intended

**SCORE INTERPRETATION AND USE**

- Scores represent what students know and can do
- Achievement level descriptors provide useful information about student achievement
- Inferences regarding student achievement, progress, and growth can be drawn at the Domain level
- Assessment scores provide information that can be used to guide instructional decisions

**GOALS**

- Students with significant cognitive disabilities are able to show what they know and can do through the end of year assessment tasks
- Parents, teachers, and students have high expectations for students' academic achievement
- Students achieve increasingly higher academic expectations
- Trajectory of student growth in academic knowledge and skills is improved

**UNINTENDED CONSEQUENCES**

Negative unintended consequences are minimized

Figure 2. Dynamic Learning Maps theory of action for science.

### ***I.1.C. KEY ELEMENTS***

Consistent with the theory of action, key elements were identified to guide the design of the DLM science alternate assessment. The list of key elements below mirrors the organization of this manual and provides chapter references. Terms are defined in the glossary (Appendix A).

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

The selection of learning targets is crucial to instruction and assessment development; teachers must be able to build the knowledge, skills, and understandings required to achieve the content standard expectations for each grade band and content area. This forms a local learning progression toward a specific learning target. The process for selecting learning targets and developing EEs with three linkage levels for assessment are described in Chapter II.

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

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

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

The DLM science alternate assessment is designed as an adaptive, computer-delivered assessment that is intended to measure knowledge, skills, and understandings at appropriate levels of complexity for the content. It consists of an end-of-year assessment that meets the requirements of accountability systems and provides detailed descriptions of what students know and can do. Assessment administration is described in Chapter IV.

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

Accessibility is a prerequisite to validity or the degree to which an assessment score interpretation is justifiable for a particular purpose and supported by evidence and theory (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Therefore, throughout all phases of development, the DLM Alternate Assessment System was designed with accessibility in mind to provide multiple means of representation, expression, action, and engagement. Students must understand what is being asked in an item or task and have the tools to respond in order to demonstrate what they know and can do (Karvonen, Bechard, & Wells-

Moreaux, 2015). The DLM alternate assessment provides accessible content, accessible delivery via technology, and adaptive routing. Since all students taking an alternate assessment based on alternate academic achievement standards are students with SCD, accessibility supports are universally available. The emphasis is on selecting the appropriate accessibility features and tools for each individual student. Accessibility considerations are described in Chapter II (linkage levels), Chapter III (testlet development), and Chapter IV (accessibility during assessment administration).

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

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

## **I.2. SYSTEM COMPONENTS**

The DLM Science Alternate Assessment System is based on EEs for science. The EEs are based on the general education grade-level content standards but exhibit reduced depth, breadth, and complexity. They link the general education content standards to grade band expectations that are at an appropriate level of rigor and challenge for students with SCD. The EEs specify the academic content standards and delineate three levels of cognitive complexity: Initial (I), Precursor (P), and Target (T). These levels represent knowledge, skills, and understandings in science that support a progression toward mastery associated with the grade band content standards. Assessment design is based on three key relationships between system elements (see Figure 3):

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

These relationships are further explained in Chapter III.

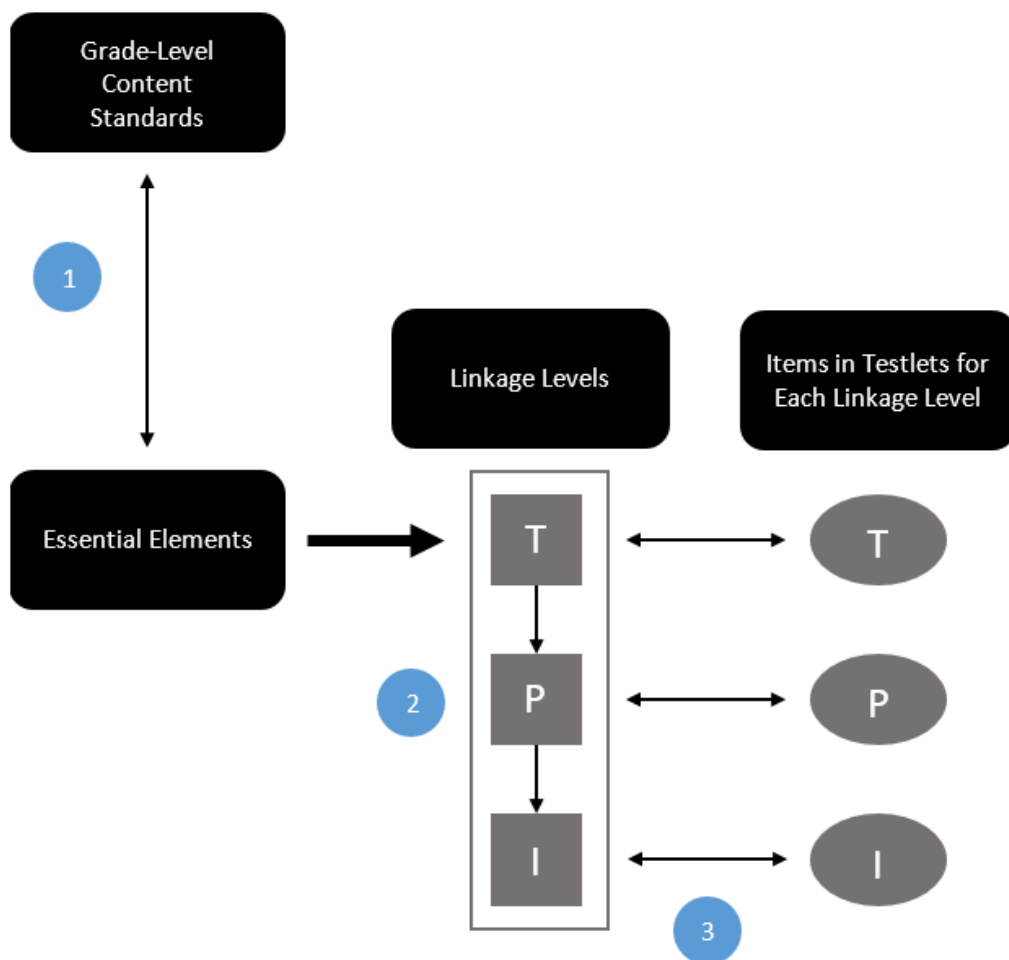


Figure 3. Design of the DLM science assessment.

*Note:* Linkage levels are Target (T), Precursor (P), and Initial (I).

### ***I.2.A. ESSENTIAL ELEMENTS AND LINKAGE LEVELS***

The DLM EEs are specific statements of knowledge and skills. The purpose of the EEs is to build a bridge from grade-level science content standards to academic expectations for student with SCD for both instruction and assessment. In other words, EEs are alternate versions of the content standards used for general education assessments. The *Framework* and subsequent NGSS performance expectations were used to develop the EEs and linkage levels, as described in Chapter II. The NGSS performance expectations are organized by grades within K-5 and in grade bands for middle school and high school. The content is organized into three domains or disciplines: physical science, life science, and Earth and space science. Within each discipline there are three to four core ideas, and within each of the core ideas are sub-ideas or topics. These sub-ideas are then elaborated into lists of what students should know and understand.



For each of the 11 core ideas, the NGSS developed performance expectations that combined a disciplinary core idea, a science and engineering practice, and a crosscutting concept.

The EEs specify alternate academic content standards aligned to grade-level content standards at reduced depth, breadth, and complexity in order to be appropriate for the DLM student population. The small collections of related knowledge, skills, and understandings are called linkage levels. The Target linkage level reflects the grade band-appropriate expectation in the EE—in other words, the expectation the student would reach by the end of that grade band. There are two linkage levels prior to the Target (Initial and Precursor).

The progression of content and skills across grade bands reflects the changing priorities for instruction and learning as students move from one grade band to the next. The differences between EEs at different grade bands are subtler than what is typically seen in content standards for general education due to the addition of linkage levels; the grade band standards expressed in the EEs consist of added prerequisite skills that are less complex than the Target. However, to the degree possible, the skills represented by the EEs increase in complexity across the grade bands, with clear links to the shifting emphases at each grade band in the general education content standards.

These three linkage levels are the basis for developing assessment items as shown above in Figure 3. Additionally, the linkage levels and their relationships are shown in visual mini-maps and described in Essential Element Concept Maps (EECMs) that item writers use during assessment development. Explanations of these tools and an example of an EECM are provided in Chapter III.

### ***I.2.B. ASSESSMENTS***

The DLM assessments are delivered as a series of testlets, each containing an unscored engagement activity and three to four items. Assessment items are written to align to one of the three linkage levels and are clustered into testlets as shown above in Figure 2. Therefore, each linkage level is available to be assessed. Students are initially placed in the assessment at the appropriate linkage level based on information collected in the First Contact survey about their expressive communication skills, as described in Chapter IV. Adaptive routing to the next appropriate testlet is provided by the system based on the student’s performance.

Assessment blueprints consist of EEs prioritized for assessment by the DLM Consortium. To achieve blueprint coverage, each student is administered a series of testlets. Each testlet is delivered through an online platform, the Kansas Interactive Testing Engine (KITE®), as described in Chapter IV. Student results are based on evidence of mastery of the linkage levels for every assessed EE as described in Chapter VI.

## **I.3. TECHNICAL MANUAL OVERVIEW**

This manual provides evidence to support the DLM Science Consortium’s assertion of technical quality and the validity of assessment claims. Because of similarities with the existing ELA and



mathematics systems, some evidence for science assessment overlaps with ELA and mathematics evidence presented in a separate manual.

Chapter I provides the theoretical underpinnings of the DLM Alternate Assessment System, including the background, purpose, rationale, target student population, problems addressed, and design. The chapter also describes how science assessment development fits within the DLM model for students with SCD and provides an overview of the components of the science assessment.

Chapter II describes the process by which the EEs and assessment blueprint were developed, guided by the *Framework for K-12 Science Education* and the needs of the student population. Based on input from experts and practitioners, the science EEs and assessment blueprint are the conceptual and content basis for the DLM science alternate assessment.

Chapter III outlines procedural evidence related to assessment content. It relates how evidence-centered design was used to develop testlets—the basic unit of assessment delivery for the DLM alternate assessment. Further, the chapter describes how the EEs were used to specify item and testlet development. Using principles of UDL, the entire development process accounted for the student population’s characteristics, including accessibility and bias considerations. Chapter III includes summaries of external reviews for content, bias, and accessibility. The final portion of the chapter describes the pilot and field tests.

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 describes how students are assigned their first testlet using the First Contact survey results and explains the assessment delivery modes (computer delivery and teacher delivery). The following sections briefly describe test administration protocols, accessibility tools and features, test security, and system usability.

Chapter V demonstrates how the DLM project draws upon a well-established research base in cognition and learning theory and uses operational psychometric methods that are relatively uncommon in large-scale assessments to provide feedback about student progress. This chapter describes the psychometric model that underlies the DLM project and describes the process used to estimate item and student parameters from student assessment data.

Chapter VI describes the methods, preparations, procedures, and results of the standard setting meeting and the follow-up evaluation of the impact data and cut points based on the 2015-2016 operational assessment administration. This chapter also explains the process of developing grade-specific PLDs for science.

Chapter VII reports the 2015-2016 operational results, including student participation data. The chapter details the percentage of students at each performance level (impact); subgroup performance by gender, race, ethnicity, and English language 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 interpretive guidance.

Chapter VIII focuses on reliability evidence, including a description of the methods used to evaluate assessment reliability and a summary of results by the linkage level, EE, and overall performance.

Chapter IX describes additional validity evidence not covered in previous chapters. It looks back at the intended score uses and interpretations as stated in the theory of action, and it details the evaluation of assessment content through review and alignment study results. The chapter relates how response processes were evaluated through review of assessment score integrity and how the internal structure of the assessment was evaluated through dimensionality and differential item functioning studies. Finally, the chapter examines the intended and unintended consequences with respect to the assessment.

Chapter X describes the training and instructional activities that were offered across the DLM Science Consortium, including the 2015–2016 training for state and local education agency staff, the required test administrator training, the optional science training, and the science instructional activities that were available to support instruction.

Chapter XI synthesizes the evidence provided in the previous chapters. It evaluates how the evidence supports the intended interpretations and uses of results from the 2015-2016 DLM science assessment and also describes ongoing and future development work of the science project.

## II. ESSENTIAL ELEMENT DEVELOPMENT

Chapter I provided an introductory description and illustration of the Dynamic Learning Maps (DLM) science alternate assessment as part of the full DLM Alternate Assessment System. In Chapter II, we describe the process for the development of the Essential Elements (EEs) for science with the overarching purpose of supporting students with the most significant cognitive disabilities (SCD) in their learning of science content standards. The EEs for science, which include three levels of cognitive complexity, are the conceptual and content basis for the DLM alternate assessments for science (Dynamic Learning Maps Science Consortium, 2015a).

The EEs were developed based on the organizing structure suggested by the *Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (National Research Council, 2012) and the Next Generation Science Standards (2013; NGSS). Guided by extensive input from experts and practitioners, the DLM EEs for science were developed in four iterations from July to December 2014. This chapter describes the development process. Pilot and field tests were designed to collect data on this content in 2015–2016 (discussed in Chapter III), and the first operational administration occurred during the spring of 2016.

The 2015–2016 alternate assessments for science were based on the EEs for science developed from the *Framework* and the NGSS in elementary, middle, and high school grade bands, as well as an end-of-course assessment in high school biology.

### 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 from the content in the *Framework* and NGSS to academic expectations for students with SCD.

Within this broad purpose, the DLM EEs for science serve three specific purposes:

1. Alignment to grade-level science standards promoting learning and development over time
2. Specification of learning targets for students with SCD based on the understanding that there are multiple ways that students can engage in instruction or demonstrate understanding through an assessment
3. Horizontal alignment with the grade-level standards and vertical alignment through the grades

#### II.1.A. GRADE-LEVEL SCIENCE CONTENT STANDARDS

The first task in EE development was to determine a common set of grade-level science standards that would support assessments across states and grades. The project began with seven states interested in developing a DLM science alternate assessment. Each state had

developed alternate science content standards and varying alternate assessments based on their own grade-level science content standards. While some of these states had already adopted the NGSS after their publication in 2013, others had not and did not intend to do so. Therefore, the *Framework*, which laid the foundation for the NGSS performance expectations but maintained a separate framework of ideas for science education, provided a more widely accepted approach that all states agreed was an appropriate common basis for the DLM science alternate assessment. This section provides an overview of the *Framework* and the NGSS and how it was used as the basis for organizing and identifying science content standards that were eventually developed into the DLM EEs.

The purpose of the NGSS was to implement the vision of the *Framework* by developing performance expectations or standards, which are the measurable statements of students' knowledge, skills, and understandings (NGSS Lead States, 2013). Following the *Framework*, the NGSS performance expectations were designed to incorporate three dimensions: disciplinary core ideas (DCIs, the content), science and engineering practices (SEPs), and crosscutting concepts (CCCs). To identify grade-level science standards, the descriptions of the NGSS performance expectations were used, and the *Framework's* coding structure, which is reflected in the DCI arrangement of the NGSS performance expectations (available at <http://www.nextgenscience.org/overview-dci>), was also incorporated. Using the NGSS performance expectations and the *Framework's* coding structure, a crosswalk with the seven interested DLM science states' current alternate science standards was conducted. This process will be discussed subsequently following a description of the *Framework* and NGSS dimensions.

The *Framework* and the NGSS described significant changes in science education and differed from previous science standards in two ways. First, they exhibited a new focus on gradual progressions of skill development in the eight SEPs rather than the more generic inquiry process that was the focus of previous standards. Second, as mentioned, each NGSS performance expectation is expressed as a combination of all three dimensions. In effect, these changes meant that students are expected to develop a deep understanding of content knowledge through application of one or more of the practices rather than just knowing science facts. The *Framework* architecture and the three NGSS dimensions are summarized here.

### **II.1.A.i. The Framework Architecture**

The *Framework* and subsequent NGSS performance expectations are organized by grades within K-5 as well as by grade band for the middle and high school grades. The content is organized into three domains or disciplines: physical science, life science, and Earth and space science. Within each discipline, there are three to four core ideas (see Table 2 below), and within each of the core ideas are sub-ideas. These sub-ideas are then elaborated into lists of what students should know and understand about the sub-idea; these lists are referred to as the DCIs. For each of the 11 core ideas, the NGSS developed performance expectations that combined a DCI, SEP, and CCC.

### **II.1.A.ii. Dimension 1: Science and Engineering Practices**

The eight SEPs listed in Table 2 are (a) the major practices that scientists employ as they investigate and build models and theories about the world and (b) a key set of engineering practices that engineers use as they design and build systems. Because the term *inquiry* has been interpreted in various ways by the science education community and expressed differently in previously developed standards documents, the DLM project articulated the SEPs and their progressions to adequately define *inquiry* as it is used in science fields and also specified the range of cognitive, social, and physical practices required for students with SCD. Because the NGSS identifies SEPs for each performance expectation, the subset of NGSS performance expectations selected for DLM resulted in the inclusion of all of the SEPs except one: asking questions and defining problems.

### **II.1.A.iii. Dimension 2: Crosscutting Concepts**

There are seven CCCs that have application across all domains of science and are meant to give students an organizational structure to understand the world. As such, they provide one way of linking across the domains in the DCIs and echo many of the unifying concepts and processes in the National Science Education Standards (National Science Teachers Association, 2010), the common themes in the Benchmarks for Science Literacy (American Association for the Advancement of Science, 1994), and the unifying concepts in the Science College Board Standards for College Success (College Board, 2009). The CCCs are 1) patterns; 2) cause and effect; 3) scale, proportion, and quantity; 4) systems and system models; 5) energy and matter in systems; 6) structure and function; and 7) stability and change of systems. All CCCs were retained for DLM assessments.

### **II.1.A.iv. Dimension 3: Disciplinary Core Ideas**

The continuing expansion of scientific knowledge makes it impossible to teach all the ideas related to a given discipline in exhaustive detail during the K-12 years, so the *Framework* identified only 11 DCIs, as shown in Table 2. DLM states value the coherent progression of DCIs across grade bands and support the idea that the goal of science education is to provide students with sufficient core knowledge so that they can later acquire additional information on their own.

### **II.1.B. ALTERNATE SCIENCE CONTENT STANDARDS CROSSWALK**

After selecting a foundation for the grade-level science standards, the next task was to determine whether there were alternate science standards that states had in common and if they could be linked to content in the *Framework*. The state partners chose not to develop EEs for every sub-idea in the *Framework*. Therefore, participating states' alternate science standards were reviewed rather than their grade-level science standards, as their alternate standards express their intended foci for students with SCD. DLM staff with expertise in science education and alternate assessments completed a crosswalk of the seven states' alternate science standards. This information allowed the DLM Science Consortium 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 EEs also map onto the eight science and engineering practices identified in the NGSS. Most states' alternate science standards included scientific inquiry practices, typically as a separate strand that was not integrated with the core content areas.

The states' alternate standards were expressed differently and at different grain sizes, but they contained common themes. Figure 4 provides several examples of different states' standards on a topic that most states addressed in their alternate standards for physical science across grade levels. The crosswalk was organized on a spreadsheet to group statements that were similar across states.

<ul style="list-style-type: none"> <li>• <u>State 1</u>: The student will observe, compare, and classify properties of matter.             <ul style="list-style-type: none"> <li>○ identifies the changes in the properties of solids, liquids, and/or gases</li> <li>○ demonstrates how one object reacts with another object or substance</li> </ul> </li> <li>• <u>State 2</u>: Students can understand and identify properties and changes of matter.</li> <li>• <u>State 3</u>: Objects, and the materials they are made of, have properties that can be used to describe and classify them.</li> <li>• <u>State 4</u>:             <ul style="list-style-type: none"> <li>○ Content Ia. Describe a physical property of matter.                 <ul style="list-style-type: none"> <li>▪ Example: Given an object, describe physical properties of the object. (e.g., color, shape, size).</li> </ul> </li> <li>○ Content Ib. Describe the appearance of a substance before and after a physical change.</li> </ul> </li> </ul> <p>Example: Given an object, describe the physical properties of the object before and after the change occurs.</p>
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure 4. Example of four states' content standards for physical properties.

The information from the cross-state review was then mapped to the DCIs in the *Framework*. The previous example was mapped to the domain of physical science (PS) under PS1: Matter and Its Interaction, PS1A: Structure and Properties of Matter.

The analysis of states' alternate content standards resulted in a list of common cross-grade DCIs and sub-ideas seen in the *Framework* in states' science standards, as shown in Table 2. The states' most commonly assessed sub-ideas and practices from the *Framework*.



Table 2. Common Science Standards Assessed by DLM States Organized by Framework  
Disciplinary Core Ideas and Sub-Ideas

Physical Science (PS)	Life Science (LS)	Earth and Space Science (ESS)
<p><b>PS1 Matter and Its Interactions</b></p> <p><i>PS1A Structure and Properties of Matter</i></p> <p>PS1B Chemical Reactions</p> <p>PS1C Nuclear Processes</p> <p><b>PS2 Motion and Stability: Forces and Interactions</b></p> <p><i>PS2A Forces and Motion</i></p> <p><i>PS2B Types of Interactions</i></p> <p>PS2C Stability and Instability in Physical Systems</p> <p><b>* PS3 Energy</b></p> <p>PS3A Definitions of Energy</p> <p>PS3B Conservation of Energy and Energy Transfer</p> <p>PS3C Relationship Between Energy and Forces</p> <p><i>PS3D Energy and Chemical Processes in Everyday Life</i></p> <p>PS4 Waves and Their Applications in Technologies for Information Transfer</p> <p><b>PS4 Waves and Their Applications in Technologies for Information Transfer</b></p> <p><i>PS4A Wave Properties</i></p> <p>PS4B Electromagnetic Radiation</p> <p>PS4C Information Technologies and Instrumentation</p>	<p><b>* LS1 From Molecules to Organisms: Structures and Processes</b></p> <p><i>LS1A Structure and Function</i></p> <p><i>LS1B Growth and Development of Organisms</i></p> <p><i>LS1C Organization for Matter and Energy Flow in Organisms</i></p> <p>LS1D Information Processing</p> <p><b>LS2 Ecosystems: Interactions, Energy, and Dynamics</b></p> <p><i>LS2A Interdependent Relationships in Ecosystems</i></p> <p>LS2B Cycles of Matter and Energy Transfer in Ecosystems</p> <p>LS2C Ecosystem Dynamics, Functioning, and Resilience</p> <p>LS2D Social Interactions and Group Behavior</p> <p><b>LS3 Heredity: Inheritance and Variation of Traits</b></p> <p>LS3A Inheritance of Traits</p> <p>LS3B Variation of Traits</p> <p><b>LS4 Biological Evolution: Unity and Diversity</b></p> <p>LS4A Evidence of Common Ancestry</p> <p>LS4B Natural Selection</p> <p>LS4C Adaptation</p> <p>LS4D Biodiversity and Humans</p>	<p><b>* ESS1 Earth's Place in the Universe</b></p> <p>ESS1A The Universe and Its Stars</p> <p><i>ESS1B Earth and the Solar System</i></p> <p>ESS1C The History of Planet Earth</p> <p><b>* ESS2 Earth's Systems</b></p> <p><i>ESS2A Earth Materials and Systems</i></p> <p>ESS2B Plate Tectonics and Large-Scale System Interactions</p> <p>ESS2C The Roles of Water in Earth's Surface Processes</p> <p><i>ESS2D Weather and Climate</i></p> <p>ESS2E Bio-geology</p> <p><b>* ESS3 Earth and Human Activity</b></p> <p>ESS3A Natural Resources</p> <p>ESS3B Natural Hazards</p> <p><i>ESS3C Human Impacts on Earth Systems</i></p> <p>ESS3D Global Climate Change</p>
<b>Science and Engineering Practices</b>		<b>Engineering, Technology, and Applications of Science (ETS)</b>
<ol style="list-style-type: none"> <li>Asking questions (for science) and defining problems (for engineering)</li> <li>Developing and using models</li> <li><i>Planning and carrying out investigations</i></li> <li><i>Analyzing and interpreting data</i></li> <li>Using mathematics and computational thinking</li> <li>Constructing explanations (for science) and designing solutions (for engineering)</li> <li>Engaging in argument from evidence</li> <li>Obtaining, evaluating, and communicating information</li> </ol>		<p>ETS1: Engineering design</p> <p>ETS2: Links among engineering, technology, science, and society</p>

Note. DLM states' most common disciplinary core ideas and science and engineering practices are *italicized*. \*These DCIs appear across all grades.

States reviewed the suggested core content for EE development that was both common across states and showed strong progressions across grades. They requested that at least one EE would be developed under each of the 11 DCIs, so the life science EEs LS3A, LS3B, and LS4C were added to the list per their recommendation. Their rationale included a desire for breadth of coverage and content that was most important for students with SCD to be prepared for college, career, and community life.

After the DCIs and sub-ideas were identified, DLM staff worked with two university science education experts, one in elementary and middle science education and one in secondary science education, to use the NGSS DCI Arrangement document to identify performance expectations that were most closely aligned to the content of the states’ alternate science standards by grade. For example, two of the four possible performance expectations for grade 5 physical science 1A (PS1A) were identified as shown in Table 3.

Table 3. Example of Next Generation Science Standards Performance Expectations Related to States’ Alternate Science Standards

<p><b>PS1 Matter and Its Interactions</b> (disciplinary core idea)</p> <p>PS1A Structure and Properties of Matter (sub-idea)</p>	<p><b>Performance Expectations</b></p> <p>5.PS.1.2: Measure and graph quantities to provide evidence that regardless of the type of change that occurs when heating, cooling, or mixing substances, the total weight of matter is conserved.</p> <p>5.PS.1.3: Make observations and measurements to identify materials based on their properties.</p>
-------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

A worksheet was prepared showing the NGSS performance expectations that best fit the states’ current alternate standards. States voted on their preferences based on their stated intent to develop EEs linked to a selected number of DCIs for the initial iteration of the assessment, with the understanding that further EE development will occur as the project matures. The operational assessment was anticipated to include approximately 30 items, with 3 to 4 items per EE. States then identified the initial set of performance expectations to use in the development of the EEs, resulting in 45 standards, as shown in Table 4. As such, this set of EEs addressed a relatively small number of science standards from the NGSS, representing a breadth, but not depth, of coverage across the entire *Framework* that corresponded to the most commonly assessed sub-ideas and practices in the states’ existing alternate standards.



Table 4. Count of Standards by Grade Band Addressed in Essential Element Development

Grade Band	Physical Science	Life Science	Earth & Space Science	Total
Elementary	4	2	3	9
Middle school	4	4	6	14
High school	4	5	6	15
High school biology	N/A	10	N/A	10

## II.2. DEVELOPMENT OF THE ESSENTIAL ELEMENTS FOR SCIENCE

The changes that the *Framework* and the NGSS prompted in science education represented a significant increase in expectations for students with SCD. Several challenges arose while considering the alternate assessments that would be built on the EEs. Alternate assessments typically constrain reliance on prior knowledge, abstract thinking, and generalization due to the cognitive characteristics of students with SCD. Therefore, it would be difficult to present science-based problem situations that accurately elicit evidence of student mastery on the multiple dimensions simultaneously. Due to the need to hold the cognitive complexity of the EEs to a rigorous but reasonable level, the EEs were drafted with the intention of maintaining two of the dimensions (DCIs and SEPs) in the expansion of the grade-level science standards for assessment purposes. Careful consideration of the SEPs was important to maintain a link to the performance expectations in the NGSS, and including the SEPs as an additional dimension in the assessments was a new feature for the design of DLM assessments. While the CCCs were not formally targeted as learning goals, they were included for instructional purposes in the EE documents.

The EEs for the DLM alternate assessments for ELA and mathematics were aligned to nodes in an overarching learning map cognitive model. In the case of science, based on the states' needs for an operational assessment developed with limited time and resources, the EEs were developed with a goal to eventually become aligned to an interconnected set of skills or nodes and assessment targets in a learning map model to be created in the future.

### II.2.A. LINKAGE LEVELS

In English language arts and mathematics, five linkage levels (LLs) were developed based on the nodes and pathways identified in the learning map models. Because science EEs were developed through a different process, in August 2014 the states discussed the number of LLs that would be appropriate for science EE development. State partners determined that three LLs of cognitive complexity would be appropriate, with the understanding that additional LLs could be added during future map development. The initial work on the development of the EEs focused on the description of the Target level. Once the Target level was created for all EEs,

two additional levels of complexity were developed within each EE, with the Target level as the highest complexity level. The lower adjacent levels, known as the Precursor level and the Initial level, clarified the knowledge, skills, and understandings students should develop to reach the Target levels. Therefore, subsequent test development steps were based on EEs with three LLs: Initial, Precursor, and Target. Table 5 illustrates the comparison of the science LLs to the English language arts and mathematics LLs.

Table 5. Comparison Between English Language Arts/Mathematics and Science Linkage Levels

Content Areas	Linkage Levels				
	ELA and Math	Initial	Distal Precursor	Proximal Precursor	Target
Science	Initial		Precursor	Target	N/A

ELA = English language arts.

### ***II.2.B. CODES FOR ESSENTIAL ELEMENTS***

The codes for the DLM EEs were derived from the *Framework* (see Table 1 above). The first part of the code indicates for which grade band the EE is intended: 5 (elementary school, which is represented by grade 5 in the coding schema), MS (middle school), or HS (high school). The next code specifies the discipline: PS (physical science), LS (life science), and ESS (Earth and space science). This is followed by codes for the core idea and sub-idea. Finally, the number at the end of each code indicates the order in which that statement appeared as a DCI in the *Framework*. In the final EE document, the code begins with the letters EE to indicate that the standard is an EE.

### ***II.2.C. DEVELOPMENT PROCESS***

Because the primary goal of the DLM Consortium is to assess grade-level academic expectations of what students with SCD know and can do, the EEs were created to accurately reflect the knowledge, skills, and understandings that are appropriately challenging grade-level targets for students with SCD.

The DLM EEs were developed in a four-step process from August to December 2014 (Table 6). The development of the first draft began with guidance from an expert panel to develop EEs for three grade bands: elementary school (represented by grade 5 standards), middle school, and high school (including EEs appropriate for end-of-course high school biology). We discuss the development of each of these drafts in detail below.

Table 6. Timeline for the Development of the Science DLM Essential Elements

Draft	Development	2014 Timeline
1	Essential Elements drafted by DLM Science Consortium and developed by expert panel	August 28 – August 29
2	DLM staff conducts in-person state educators review	October 14 – October 15
3	States conduct internal review	October 27 – November 7
4	Final state review	November 18 – December 3

### II.2.C.i. Draft 1: Expert Panel Development

The first draft began with guidance from the DLM Science Consortium states’ science experts and consultants with science and special education expertise. The expert panel started development of EEs for three grade bands: elementary school (represented by grade 5 standards), middle school, and high school (including EEs appropriate for end-of-course high school biology). The expert panel convened in August 2014 and consisted of seven expert panelists. These panelists had representative experience in the fields of special education, science education, English language arts and mathematics content, and measurement (Appendix B). The panel members represented five universities and two state departments of education and included persons who had been extensively involved in the development of the NGSS. Drs. Neal Kingston, Sue Bechard, Brooke Nash, and Jake Thompson from the DLM organization facilitated the meeting.

Using the selected core content for EE development that (a) was common across states, (b) demonstrated strong progressions across grades, or (c) was selected as being important for students with SCD to be prepared for college, career, and community life, DLM staff drafted a Target level EE for each identified NGSS standard in preparation for the August meeting. The purpose of the meeting was to have the panel review and revise all of the 45 EEs covering all of the grade bands and end-of-instruction biology. Finally, DLM staff provided guidance on fidelity to the NGSS grade-level performance expectations, vertical alignment of EEs across grade bands, ideas for LL statements, and horizontal alignment to connect EEs in English language arts and mathematics. The NGSS identified connections to the Common Core State Standards, so EEs related to those were identified. The Initial Precursor LL of those EEs were found to help the expert panel draft the Initial LL for science. The meeting started with a group introduction to the review processes; an overview of the 45 science standards that were the basis for EE development, including the process for selection (i.e., crosswalk); and presentation of the final selections. The training presentation included defining and explaining the draft EEs for science. The group worked as a whole to commence review and revision work for drafted EEs. On the second day, groups completed their tasks of review, revision, and development and

met together with English language arts and mathematics experts to make connections across content areas and to develop LL statements.

As an example, an EE was developed based on NGSS performance expectation 5-PS1-3, “Make observations and measurements to identify materials based on their properties.” For the DLM project, the EE LLs were developed such that the Target level included the same content and practice as the performance expectation. In this example, the Target level uses the same wording as the NGSS performance expectation, with some clarifying examples added to aid interpretation. Precursor and Initial level descriptions were developed to show knowledge, skills, and understandings students should develop to reach the Target, and possible connections to English language arts and mathematics map nodes at the Initial level (F-75 and M-76) were identified. Figure 5 shows the resulting LLs for this example EE.

<p><b>Essential Element: EE.5.PS.1.3</b></p> <p><b>Target Level:</b> Make observations and measurements to identify materials based on their properties (e.g., weight, shape, texture).</p>
<p><b>Precursor Level:</b></p> <p>Match materials with similar physical properties.</p>
<p><b>Initial Level:</b></p> <p>Recognize same. Recognize different.</p>
<p><b>Initial Precursor ELA/Math EE Connections</b></p> <p>F-75 Demonstrate an understanding of property words.</p> <p>M-76 Classify</p>

Figure 5. Essential Element with linkage levels developed by expert panel and connections noted to ELA and mathematics.

Extensive notes were taken during the expert panel meeting to reflect the discussions and the issues considered. To demonstrate the types of dialogue that occurred, the key discussion points from the review of EE.5.PS.1.3 are listed below.

- There was a long discussion about how to create a Target level description that was less complex than the performance expectation in the NGSS. The group decided the grade-level standard verbs were okay to use in the Target EE as long as there was a clarification of which physical properties should be included in the example list, thus reducing the complexity of the context.
- The remainder of the discussion of this EE focused on which properties to include. Color was left out, as this would be difficult for students who are blind or have visual impairments.

These notes were compiled and provided to the next group of reviewers along with Draft 1.

The expert panel review resulted in a revision of all but one of the original draft Target level EEs and the creation of the additional two LLs for each EE.

### **II.2.C.ii. Draft 2: In-Person External Review**

Draft 1 EEs and the Draft 1 notes were presented to representatives from each state education agency and their selected educators and content specialists. Sixteen experts in science and 17 individuals with expertise in instruction for students with SCD from five states reviewed the draft documents in a two-day, in-person meeting in October 2014. Participating reviewers had a variety of backgrounds and experiences, but most had some classroom experience in teaching science and/or had experience teaching students with SCD. Some reviewers held leadership roles, including work at the district level on special education or curriculum. Many reviewers had worked on other statewide assessments as item writers or reviewers.

This review process used a standardized checklist to determine whether the EEs and related LLs were acceptable or needed revision (Figure 6). If revisions were recommended, panelists were asked to identify the issue of concern and to provide specific wording for the recommended changes. Panelists were first organized into grade band panels that represented special education and science experts who had experience teaching within each of the grade bands. After the grade band panels completed their review and discussion of each of the relevant EEs and LLs, panelists were re-organized into science domain-focused groups (life science, physical science, and Earth and space science). The science domain-focused groups consisted of science education teachers who had experience teaching within the domain and special education teachers who were matched to a domain, where appropriate, depending on experience. The domain-focused groups reviewed the recommendations of the grade band groups for all of the EEs and LLs that measured the science domain across all grade bands and made recommended changes accordingly. DLM staff facilitated the discussions at each table.

<b>Essential Element Review Checklist</b>	
<input type="checkbox"/>	Does it align to the standard?
<input type="checkbox"/>	Does it reflect a high but reasonable expectation of what a student with the most significant cognitive disabilities can do?
<input type="checkbox"/>	Does it reflect what the student needs for post-secondary life?
<input type="checkbox"/>	Is the scope appropriate and manageable?
<input type="checkbox"/>	Is it written in universal terms so students can demonstrate knowledge and skills in a variety of ways?
<input type="checkbox"/>	Does it use terms that are consistent across EEs?
<input type="checkbox"/>	Is it similar in complexity with other EEs that are written for the same grade band level?

Figure 6. Essential Element review checklist.

This review resulted in significant changes that

- clarified the science concepts that are the essential targets for measurement,
- revised verbs to convey clear statement of what the student should demonstrate related to scientific and engineering practices,
- focused on universal access issues,
- revised the EEs to be more measurable,
- aligned the LLs with the Target EEs across the grade band and refined Initial and Precursor levels, and
- provided examples within the EE statements.

As an illustrative example, Figure 7 demonstrates the revisions made by the external review panel to EE.PS.1.3. In this revision, panelists recommended changes in the examples at the Target level and changes in wording to the Precursor and Initial levels. Also, possible connections to nodes on the English language arts and mathematics map at the LL were adjusted, referencing three foundational nodes (F-2, F-75, F-76) and one mathematics node (M-76).

<p><b>Essential Element: EE.5-PS1-3</b></p> <p><b>Target Level:</b> Make observations and measurements to identify materials based on their properties (e.g., weight, shape, texture, buoyancy, or magnetism).</p>
<p><b>Precursor Level:</b></p> <p>Classify materials by physical properties (e.g., weight, shape, texture, buoyancy, or magnetism).</p>
<p><b>Initial Level:</b> Match materials with similar physical properties.</p>
<p><b>Initial Precursor ELA/Math EE Connections</b></p> <p>F-2 Recognize same</p> <p>F-75 Can demonstrate an understanding of property words.</p> <p>F-76 Recognize different</p> <p>M-76 Classify</p>

Figure 7. External review panel revisions to EE.5.PS1-3.

Again, notes were taken of the discussion and made available to the next set of reviewers. In this example, discussion of EE.5.PS1-3 included

- Target: Kids love buoyancy and magnetism. Whatever ways comparisons are made, there should be options to use (e.g., student with autism may not want to touch objects). Measurement does not require numbers.
- Precursor: Draft 2 Precursor should be the Initial level—move down. The understanding of property words is necessary before students can observe and measure.
- Initial: Students at this level can match.

DLM staff asked participants to complete surveys at the end of the October review meeting. Overall, results of the surveys showed agreement or strong agreement for every evaluation item (Table 7).



Table 7. Percentage of Participant Ratings by Level of Agreement for Evaluation Items (N = 33)

<b>Evaluation Item</b>	<b>Strongly Disagree</b>	<b>Disagree</b>	<b>Agree</b>	<b>Strongly Agree</b>	<b>Missing</b>
The overall goals and objectives for this review meeting were clear.	3.0	0.0	30.3	66.7	0.0
The contents of the presentation were effective at helping me participate in the review process.	3.0	6.1	42.4	48.5	0.0
The resource materials provided by DLM staff were effective at helping me participate in the review process.	3.0	6.1	45.5	45.5	0.0
I had enough time to review and discuss each Essential Element.	3.0	0.0	42.4	51.5	3.0
I felt comfortable providing feedback and suggestions on the Essential Elements.	3.0	0.0	33.3	60.6	3.0
The DLM staff were knowledgeable about the review process and goals.	3.0	0.0	30.3	66.7	0.0
I am confident that the feedback and suggestions to the Essential Elements will benefit students with the most significant cognitive disabilities and their teachers.	6.1	3.0	33.3	54.5	3.0
I valued the DLM Essential Element review process as a professional development experience.	3.0	0.0	18.2	78.8	0.0

In general, review participants expressed satisfaction with the workshop proceedings, with emphasis on the professional development afforded by working on the EE reviews. For example, one participant wrote, “I always learn so much from interacting with other teachers. I have a much better understanding of special education and the problems faced in that area.” Comments from participants were generally positive about the review process. They expressed having enjoyed the process, with appreciation for the opportunity to learn, to contribute, and to collaborate with other educators. Some participants stated that the meeting was well organized



and informative. One person expressed appreciation for being able to ask the DLM staff questions about testing and procedures.

One person noted concerns regarding “the severely disabled students as opposed to the significantly disabled and the incredible differences between students in their ability and intellectual range.” In addition, the same participant expressed the concern that “[t]eachers feel huge pressure when students are severely intellectually limited in their growth potential.” Some participants stated that noise in the room was distracting at times. One wrote, “A larger room or smaller separate rooms would help for noise problems. Great discussion but it made it hard to hear our group. Also roundtables not as easy to get large groups around. It might have helped when we compared the subject matter over grades it would have been easier to see the flow from one to another if they were side by side...” In addition, one participant expressed, “I was unaware until the second day that most initial level questions would be given by the teacher.”

### **II.2.C.iii. Draft 3: State Internal Review**

The third round of EE reviews occurred in November 2014, and the DLM Science Consortium states of Iowa, Kansas, Missouri, and Oklahoma each facilitated their own review process. They did not provide DLM staff with data on the number or the experience of the reviewers they selected, as the internal review process was intended to be completely state-driven. State representatives and experts selected by the state reviewed the Draft 2 EEs.

DLM staff prepared materials for states to use, including a PowerPoint training video, copies of each Draft 2 EE with notes from the first two review panels, a feedback spreadsheet, and a list of guiding questions:

1. Do the Essential Elements fit within the topics and core ideas that are the framework for the DLM system?
2. Do the Essential Elements in each topic support student learning over time?
3. Are the Essential Elements and linkage level learning targets clearly defined?
4. Do the linkage levels represent the learning target content at appropriately reduced levels of breadth and depth?

The reviewers used a rating form that captured each DCI and sub-idea as well as the EE and corresponding LLs. Reviewers logged their decision to accept the EE as is or to revise it, noting problems and recommendations for revision if appropriate. State representatives were asked to compile a single set of reviewer responses to submit to DLM staff.

The states’ responses were compiled and reviewed by DLM staff. Although there was overall acceptance of the EEs, there were suggestions that required further discussion by the states. DLM staff compiled the comments into Draft 3 and indicated which items needed further discussion. For example, the Target level wording for EE.5.PS.1.3 was rewritten as seen below (to add “mass”) accompanied by the following State Comment:

- **Target Level:** Make observations and measurements to identify materials based on their properties (e.g., mass, weight, shape, texture, buoyancy, color, or magnetism).
- **State Comment:** Using the term “mass” will ensure accurate content is taught and the concept is understood. Weight and mass are often misconceptions in science because the term “weight” is often misused. Mass is easily shown on a balance scale for these students and easily manipulated by adding or subtracting to the different sides of the balance. Calling the difference a change in weight would be incorrect. Weight is much more abstract in that it involves gravitational pull (e.g., a person weighs less in an airplane than in Death Valley).

A state call was held on November 18, 2014 to consider the comments and reach a consensus on how to proceed. In the case of the above comment, minutes indicate that “verbal consensus was to remove the term ‘mass’ from examples.” Following the call, DLM staff revised Draft 3 of the EE document and produced Draft 4 for final state review and vote.

#### II.2.C.iv. Draft 4: DLM Science Consortium Review and Vote

A discussion and consensus vote by participating states in December 2014 resulted in the final EEs for science (*Dynamic Learning Maps Essential Elements for Science*, 2015). These EEs were then used to develop the test blueprints and the DLM science assessments. The EEs are presented with tables by grade band (i.e., elementary, middle school, high school) and domain (i.e., physical science, life science, Earth and space science) in a format that contains core idea, sub-idea (topic), state standard for the general education group (using NGSS language), the description of the EE by LL (Target, Precursor, Initial), as well as connections to SEPs, CCCs, and DLM English language arts and mathematics EEs. The connections to specific English language arts and mathematics map nodes were excluded (i.e., the Initial Precursor linkage level), but connections to English language arts and mathematics EEs identified from the NGSS Connections are included.

The result of the vote by the states was the set of final EEs

([http://dynamiclearningmaps.org/sites/default/files/documents/Science/Science\\_EEs\\_Combined\\_final\\_Jan2017.pdf](http://dynamiclearningmaps.org/sites/default/files/documents/Science/Science_EEs_Combined_final_Jan2017.pdf)). Figure 8 shows EE.5.PS.1.3 as approved in December 2014.

<b>Domain:</b> Physical
<b>Core Idea:</b> PS1: Matter and Its Interactions
<b>Topic:</b> PS1.A: Structure and Properties of Matter

<p><b>State Standard for General Education:</b></p> <p><b>5.PS1-3:</b> Make observations and measurement to identify materials based on their properties.</p>
<p><b>Essential Element: EE.5.PS1-3</b></p> <p><b>Target Level:</b> Make observations and measurements to identify materials based on their properties (e.g., weight, shape, texture, buoyancy, color, or magnetism).</p>
<p><b>Precursor Level:</b> Classify materials by physical properties. (e.g., weight, shape, texture, buoyancy, color, or magnetism).</p>
<p><b>Initial Level:</b> Match materials with similar physical properties.</p>
<p><b>Connections to Science Practices</b></p> <p>Planning and Carrying out Investigations</p>
<p><b>Connections to Crosscutting Concepts</b></p> <p>Scale, Proportion, and Quantity</p>
<p><b>Connections to English Language Arts Essential Elements</b></p> <p><b>EE.W.5.7:</b> Conduct short research projects using 2 or more sources.</p> <p><b>EE.W.5.8:</b> Gather and sort relevant information on a topic from print or digital sources into given categories.</p>
<p><b>Connections to Mathematics Essential Elements</b></p> <p><b>EE.5.MD.A.1:</b> Use standard units to measure weight and length.</p>

Figure 8. Final approved EE.5.PS.1-3.

### II.3. SCIENCE BLUEPRINT DEVELOPMENT

The summative DLM science test blueprint was developed in late 2014. A total of 45 standards approved for EE development (9 EEs at the elementary level, 14 EEs at the middle school level, 15 EEs at the high school level, and 10 life science EEs for end-of-course high school biology), as shown in Table 4.

Despite a commitment initially expressed by states to a blueprint that would maximize the breadth of content coverage, given the number of EEs at each grade level, it was necessary to select and weigh the EEs to meet the test length requirement of approximately 10 EEs per grade level. States desired a summative (Year-End) assessment for which students take a 25- to 30-item test. The assessment would be designed in the form of testlets containing three or four items written to assess a single EE. The elementary level already contained only nine EEs, and high school biology had 10 EEs, so these blueprints could accommodate a test with

approximately 25–30 items. Therefore, the focus of the blueprint decisions was on the middle and high school levels where a reduced number of EEs was required. The resulting blueprint options for these grade bands covered content in all three science domains, but with different emphases.

### ***II.3.A. OPTIONS DEVELOPMENT AND SELECTION***

The development of the middle and high school blueprints followed a three-step process: (1) a group of educators rated EEs using an Excel spreadsheet sent via email, (2) ratings were compiled and used to assemble options for blueprints, and (3) member states met to vote on a final blueprint. The principles that guided the development of four blueprint options for each grade band were

- use the feedback from the educator survey to prioritize content that has the potential to maximize student growth in academic skills across grades.
- use knowledge of academic content and instructional methods to prioritize content that is considered important by stakeholders and central to the construct.
- prioritize content that can be applied to real-world or workplace problems.
- maximize the breadth of coverage of EEs, given the time needed to administer an assessment to students in the alternate assessment population.

In the first step, 10 of the 31 educators (32%) who attended the EE review meeting in October 2014 (see Section II.4.A) rated all of the EEs via electronic survey. The 10 educators who responded had a range of experiences in science education ( $n = 4$ ) or special education ( $n = 6$ ) and represented the participating science partner states of Iowa, Missouri, Mississippi, Kansas, and Oklahoma.

Ratings were based on three criteria using a 4-point agreeability scale (4 = agree, 3 = somewhat agree, 2 = somewhat disagree, 1 = disagree). The three criteria were

- The EE reflects a high but reasonable expectation for a student with the most significant cognitive disabilities at this grade band.
- The EE is important for learning what the student will need in post-secondary life.
- The EE is relevant to current science instruction in the classroom.

Results were aggregated for middle school and high school, and two aggregate variables were calculated for each EE (see Appendix B). The average agreement rate was the average proportion of the respondents who chose the “top box” (agree) across the three criteria. The overall average rating was the average of the mean rating across the three criteria.

In step 2, the educator survey ratings were compiled and used to develop two different blueprint options for each middle school and high school grade band. In each case, blueprint option 1 used the highest overall average ratings, and option 2 used the highest average

agreement rates to organize the EEs. Blueprint options 3 and 4 suggested additional breadth of content considerations. These blueprint options were prepared to be reviewed by states between November 24 and December 9, 2014, and they were discussed at an in-person meeting on December 9, 2014. States received the following information

1. Process for Arriving at Science Blueprint Options. This document contained a description of the process used to derive the blueprint options, including the results of the educator survey that was used to create the first two blueprint options at the middle and high school levels.
2. Blueprint Options. This document contained the four options for blueprints, with the specific EEs to be included in each option and the resulting number of EEs in each domain. Considerations were described in the following manner and also included lists of specific EEs included in each option:

#### **Middle School Option 1**

<b>Physical</b>	<b>Life</b>	<b>Earth &amp; Space</b>	<b>Total</b>
1	4	5	10

#### Considerations:

- Accounts for EEs that had the highest overall ratings across the three criteria
- Highest overall ratings accounts for how teachers rated the EEs across the scale (i.e., how much they agreed **AND** how much they disagreed that the EE met the criteria)
- Lack of coverage in Physical Science
- Possible over-coverage in Earth & Space Science
- Purely data-driven approach to EE selection

#### **Middle School Option 2**

<b>Physical</b>	<b>Life</b>	<b>Earth &amp; Space</b>	<b>Total</b>
2	3	5	10

#### Considerations:

- Accounts for EEs that had the highest average agreement ratings across the three criteria
- Highest average agreement ratings accounts for how teachers rated the EEs at the “top box” (i.e., how much they agreed that the EE met the criteria)
- Lack of coverage in Physical Science
- Possible over-coverage in Earth & Space Science

- Ten EEs total, so all testlets will have three items
- Another purely data-driven approach to EE selection

### Middle School Option 3

Physical	Life	Earth & Space	Total
3	3	3	9

#### Considerations:

- Accounts for EEs that had the highest overall average ratings within each domain
- Ensures balanced coverage across domains
- Includes some EEs that were not as highly rated relative to other EEs from different domains
- Nine EEs total, so three testlets will have four items
- Data- and content-driven approach to EE selection

### Middle School Option 4

Physical	Life	Earth & Space	Total
3	3	4	10

Considerations:

- Accounts for EEs that had the highest overall average ratings within each domain
- Ensures balanced coverage across domains, with additional weight in the domain that was most highly rated
- Includes some EEs that were not as highly rated relative to other EEs from different domains
- Ten EEs total, so all testlets will have three items
- Data- and content-driven approach to EE selection

### High School Option 1

Physical	Life	Earth & Space	Total
2	3	5	10

Considerations:

- Accounts for EEs that had the highest overall ratings across the three criteria
- Highest overall ratings accounts for how teachers rated the EEs across the scale (i.e., how much they agreed **AND** how much they disagreed that the EE met the criteria)
- Lack of coverage in Physical Science
- Possible over-coverage in Earth & Space Science
- Purely data-driven approach to EE selection



### High School Option 2

Physical	Life	Earth & Space	Total
1	3	6	10

Considerations:

- Accounts for EEs that had the highest average agreement ratings across the three criteria
- Highest average agreement ratings accounts for how teachers rated the EEs at the “top box” (i.e., how much they agreed that the EE met the criteria)
- Lack of coverage in Physical Science
- Possible over-coverage in Earth & Space Science
- Ten EEs total, so all testlets will have three items
- Another purely data-driven approach to EE selection

### High School Option 3

Physical	Life	Earth & Space	Total
3	3	3	9

Considerations:

- Accounts for EEs that had the highest overall average ratings within each domain
- Ensures balanced coverage across domains
- Includes some EEs that were not as highly rated relative to other EEs from different domains
- Nine EEs total, so three testlets will have four items
- Data- and content-driven approach to EE selection

#### High School Option 4

Physical	Life	Earth & Space	Total
3	3	4	10

#### Considerations:

- Accounts for EEs that had the highest overall average ratings within each domain
- Ensures balanced coverage across domains, with additional weight in the domain that was most highly rated
- Includes some EEs that were not as highly rated relative to other EEs from different domains
- Ten EEs total, so all testlets will have three items
- Data- and content-driven approach to EE selection

The third and final step of the science blueprint development process involved states reviewing the blueprint option documentation internally and discussing as a consortium prior to voting for the final blueprint options in middle and high school grade bands. Again, the elementary set of EEs and End-of-Instruction biology set of EEs were not included in blueprint development process because they already consisted of the desired number of EEs to be assessed.

### ***II.3.B. FINAL SCIENCE BLUEPRINT***

The result of the state vote was to select the blueprint option that consisted of nine EEs at each grade band. The consensus decision was that a smaller scope of standards was desirable for the new science content standards for students with SCD. The rationale for this decision included perceived limited opportunity to learn science content for students with SCD and the desire to minimize the breadth of content educators would need to focus on for the new administration of DLM science assessments. The final blueprint included a total of 37 EEs: nine at each grade band and 10 EEs for End-of-Instruction biology, as shown in Table 8. The final set of EEs included on the blueprint represent a breadth of content coverage across 10 DCIs, 14 topics, and 7 SEPs. Appendix B provides the final blueprint for each grade and course.

Table 8. Count of Essential Elements Included in Science Blueprints for 2014–2018

Level	Physical Science	Life Science	Earth & Space Science
Elementary	4	2	3
Middle school	3	3	3
High school	3	3	3
High school biology	N/A	10	N/A

#### II.4. CONCLUSION

The DLM EEs for science were carefully developed with multiple rounds of stakeholder input to reflect high expectations for students with SCD. Priorities in participating states' current science content standards and the *Framework* and NGSS informed development of the EEs. The three linkage levels provide access to the EE with varying cognitive complexity. Blueprints were developed using several criteria to prioritize EEs that are valued for the student population and that have the potential to support high student attainment and growth.

### III. ITEM AND TEST DEVELOPMENT

Chapter II described the development of the Essential Elements (EEs) for science with the overarching purpose of supporting students with the most significant cognitive disabilities (SCD) in their learning of the content standards. Following from the discussion in Chapter II, Chapter III presents the rationale and processes that DLM staff used to develop the items and test content for the DLM alternate assessment in science.

EEs are specific statements of knowledge and skills, analogous to alternate or extended content standards. The EEs were developed (see Chapters I and II) by linking to the grade-level expectations identified in *Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (National Research Council, 2012; *Framework*) and Next Generation Science Standards (2013; NGSS). The purpose of the EEs is to build a bridge from the *Framework* and NGSS to academic expectations for students with SCD.

For each EE, three linkage levels were identified: Initial, Precursor, and Target. A linkage level is an incremental level of complexity toward the learning target for an EE. The EEs specify the learning target (Target linkage level), with the Initial and Precursor linkage levels clarifying how students can reach those targets. The Target linkage level reflects the grade-level expectation linked directly to the NGSS performance expectation. For each EE, the two linkage levels preceding the Target, represent important knowledge, skills and understandings on the way to the target level skill. Assessment items were grouped into testlets and developed based on each of the three linkage levels.

#### III.1. REVIEW OF SCIENCE ASSESSMENT STRUCTURE

The DLM EEs for science are the basis upon which all content was developed. As described in Chapter II, the framework for the system was adapted from the National Research Council's *Framework* and the NGSS. The final blueprint included a total of 37 EEs: nine at each grade band and 10 EEs for End-of-Instruction.

As discussed in Chapter II, seven science practices were incorporated into the DLM EEs for science. A document developed by DLM staff, the *DLM Adapted Science and Engineering Practices*, details information on each practice, including component skills and grade-level progressions appropriate for students with SCD (Appendix C).

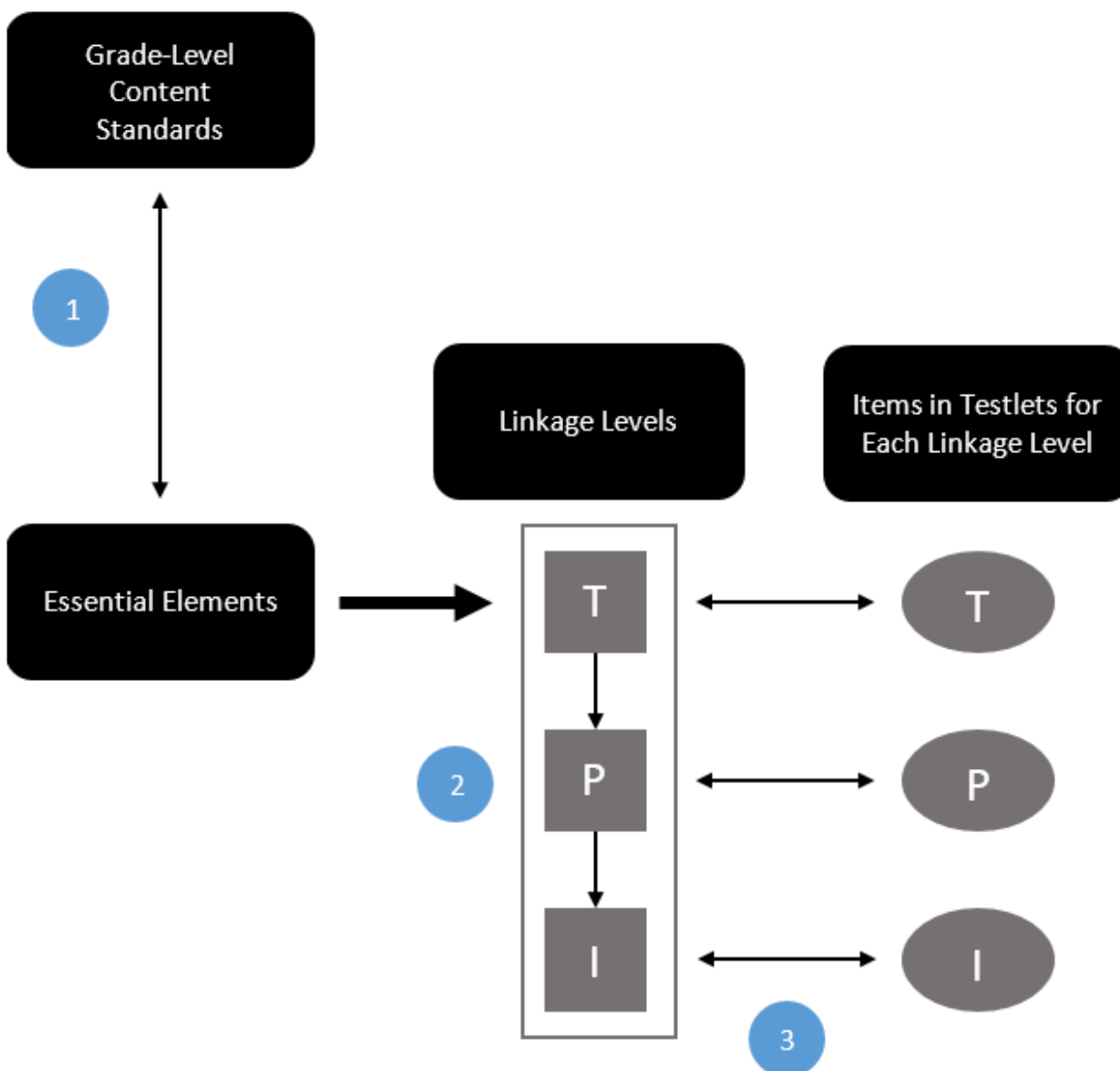


Figure 9. Design of the DLM science assessment.

Note: Linkage levels are Target (T), Precursor (P), and Initial (I).

Figure 9 depicts the development flow from standards to items in testlet, for the DLM science assessment. Overall, the relationship of test items and testlets to the grade-level content standards is mediated by the EEs and linkage levels. Therefore, test design is based on three linkages, depicted as blue circles in Figure 9: (1) the links of the content standards (*Framework* and NGSS) and the DLM EEs for science, (2) the links of the EEs and linkage levels, and (3) the links of the linkage levels and items/testlets.

### **III.1.A. ITEMS AND TESTLETS**

Testlets are the basic units of the DLM Alternate Assessment System. These testlets are short, instructionally relevant measures of student knowledge, skills, and understandings that are designed to provide results that can inform instructional planning. Each testlet begins with an engagement activity—a stimulus related to the assessment designed to help the student focus on the task at hand, or become involved in a science activity—followed by three to four items. There is one testlet per EE and linkage level (e.g., three testlets for each EE). Students take a series of testlets to achieve blueprint coverage according to the test’s design. An example of a testlet can be seen in Chapter IV.

#### **III.1.A.i. Overview of the Testlet Development Process**

The testlet was the focus of DLM assessment development. Item writers wrote all items for assigned testlets following an evidence-centered design (ECD) approach. Every testlet went through multiple rounds of development, reviews by DLM staff for content and accessibility, editorial reviews, external reviews by educators in DLM states, and revisions. The full set of test development steps are outlined below.

1. Item writer is trained.
2. Item writer is assigned testlet specifications articulated by the Essential Element Content Map (EECM) with other supporting materials, as described in section III.2.
3. Item writer develops a draft testlet and associated metadata.
4. Content team completes first internal quality control review.
5. Testlet receives first editorial review. Where applicable, graphics needed for engagement activities and items are inserted.
6. Content and accessibility specialists complete internal quality control review.
7. Content team completes second internal quality control review.
8. Testlet is entered into the content management system.
9. Testlet receives second editorial review.
10. Content team completes third internal quality control review.
11. External reviewers review testlet for content, accessibility, and bias and sensitivity.
12. Synthetic read aloud tagging is applied to the testlet.
13. Test production team completes first quality control review.
14. Testlet is prepared for delivery in the content management system.

15. Testlet receives testing window delivery quality control checks by test production, content, and psychometric teams for accessibility, display, content, and associated test delivery resources.
16. The testlet is delivered for field testing.
17. Field test data is reviewed by psychometric and content teams.
18. Testlets and items that do not require revision are made operational.
19. Prior to operational use, step 15 is repeated.

Each review group was carefully trained to look for potential problems with the academic content, accessibility issues, and concerns about bias or sensitive topics. After testlets were externally reviewed and then revised, they were scheduled for field testing. DLM staff reviewed results from field tests to determine which testlets met quality standards and were ready for operational assessment. Security of materials was maintained through the test development process. Paper materials were kept in locked facilities. Electronic transfers were made on a secure network drive or within the secure content management system.

### **III.1.A.ii. General Testlet Structure and Item Types**

Testlets are based on learning targets for one linkage level of one EE. Each testlet contains a non-scored engagement activity and three to four items.

There are two general modes for DLM testlet delivery: computer-delivered and teacher-administered (see Chapter IV). Computer-delivered assessments are designed so students can interact independently with the computer using special assistive technology devices such as alternate keyboards, touch screens, or switches as necessary. Computer-delivered testlets emphasize student interaction with the content of the testlet, regardless of the means of physical access to the computer. Therefore, the contents of testlets, including directions, engagement activities, and items, are presented directly to the student. Educators may assist students during these testlets using procedures described in Chapter IV.

Teacher-administered testlets are designed for educators to administer outside the system, with the test administrator recording responses in the system rather than the student recording his or her own responses. These teacher-administered testlets include onscreen content for the test administrator that begins by telling, in a general way, what will happen in the testlet. Directions for the test administrator then specify the materials that need to be collected for administration. After the educator directions screen(s), teacher-administered testlets include instructions for the engagement activity. After the engagement activity, items are presented. All teacher-administered testlets have some common features:

- Directions and scripted statements guide the test administrator through the administration process.
- The engagement activity involves the test administrator and student interacting directly, usually with objects or manipulatives.



- The test administrator enters responses based on observation of the student’s behavior.

Testlet organization, the type of engagement activity, and the type and position of items vary depending on the intended delivery mode (computer-delivered or teacher-administered) and content being assessed. Specific descriptions and examples of the structure of testlets, engagement activities, and different item types are included in the following sections.

DLM computer-delivered testlets used only multiple-choice, single-select item formats for the 2016 science operational assessment. All items within the testlets have three answer options presented in a multiple-choice format using either text or images. Teacher-administered testlets contain items with five options that describe anticipated student behaviors. Test administrators select the description that most closely matches their observations of student response when the item is administered.

### **III.1.A.iii. Science Testlet Development**

Science testlets begin with an engagement activity. The purpose of the engagement activity is to increase access for this student population by setting the context, activating prior knowledge, and increasing student interest. The engagement activity in science may also present a science story that describes an experiment or science activity. Three to four test items follow or are embedded in the engagement activity.

Test content developers used specific guidelines in writing the engagement activities and subsequent items to ensure alignment to the EEs and adherence to the same item writing specifications used in ELA and mathematics (see subsection III.2 for a description of item writing; see Appendix C for a list of all materials used by item writers). These item writing specifications have been refined over time to effectively produce items and testlets based on principles of evidence-centered design and Universal Design for Learning.

- Item writers considered the linkage level and grade level for the testlet being written. Testlets become more complex as linkage levels progress within grade band EEs and as grade bands go up.
- Writers kept the student population in mind. For example, when writing sentences, they used single syllable, decodable words when possible, and used simple sentences, avoiding commas, negation (using “not”), and pronouns.
- Technical vocabulary was used only when it was necessary for the linkage level. For example, for HS.LS2-2, the Precursor level reads, “Recognize the relationship between population size and available resources from a graphical representation.” A student with the most significant cognitive disabilities can grasp this concept without knowing the vocabulary word “population.” Therefore, for this particular linkage level, “number of deer” would be more accessible.

- Content developers ensured representative diversity of people in images and names and always used people-first language when writing about someone with a disability, making sure to avoid regional references.
- Science testlets were written in the present tense.
- Science stories were developed if they were useful and plausible.
- Finally, developers wrote science stories such that the student can use the science knowledge that they have been taught to answer questions about concepts that have been broken down into more manageable sections.

### *III.1.A.iii.a Testlet Engagement Activities*

Science testlets have different types of engagement activities, depending on the nature and/or complexity of the linkage level. One type of engagement activity has two or three sentences on a screen that leads into questions about the science concept. For example, an engagement activity for EE.5.PS3-1 Precursor level “use models to describe that plants capture energy from sunlight” could be as follows: “Jon plants a flower. Jon knows that plants need light to grow. Jon makes models to show how plants get light.” This would be followed by showing Jon’s models and asking questions about each one. The purpose of this type of engagement activity is to activate students’ prior knowledge of the science concept and engage student interest. Other times, the engagement activity will be an activity, experiment, or hypothetical situation involving a fictional student. This is called a science story and involves more information, similar to informational text in ELA. These types of engagement activities provide descriptive information about a situation that the student can use to respond to questions.

A science story tells about a fictional student and consists of multiple screens that set up the context for the upcoming items. The story involves walking a fictional student through an experiment or activity with items embedded throughout the process. Scaffolding, or breaking down a concept into smaller parts, is often used when teaching this population. A science story can help lead the student through a classroom activity or experiment that could have been done during instruction to break down a complex concept and make it more accessible.

Figure 10 below shows an example of a science story for EE.HS.PS3-4. The Target linkage level reads, “Investigate and predict the temperatures of two liquids before and after combining to show uniform energy distribution.” The purpose of this science story is to create a context for the student to use the scientific practice of Planning and Carrying Out Investigations. The screens of the science story walk the student through the process of mixing water with different temperatures to investigate the effects of variables such as temperature and amount of water, mirroring activities that take place during science instruction. The abstract concept of energy distribution is made more concrete through comparisons of temperature readings.

Jill experiments with water temperatures. Jill measures the temperature of water in two beakers.

Each beaker has one cup of water. The water in the first beaker is 40 degrees. The water in the second beaker is 80 degrees.

Figure 10. Example science story (EE.HS.PS3-4).

Many linkage levels, while complex, can be assessed without walking a student through a hypothetical science classroom activity or experiment. Such testlets may be more accessible without a science story. Science stories were written where they were needed to align the testlet to the content and science practice of the linkage level. The Initial level testlets do not use science stories; rather, they present an engagement activity within a set of directions for the educator that introduces the student to the pictures or objects that will be used in the testlet.

#### **III.1.A.iv. Selection of Accessible Graphics for Testlets**

Graphics for science testlets were selected using guidelines developed with input from state partners to ensure that they were accessible for students. For graphics in science testlets use colored line drawings. Graphic designers created images for science to employ high contrast and provide clear, simple graphic representations of content only in cases where required to assess the construct and for engagement activities. Graphic designers and item writers received training to avoid the creation of items that relied on students' perception of color. Image quality and accessibility were reviewed as a part of the external review process for items and testlets.

#### **III.1.A.v. Items**

Science testlets contain multiple-choice items. For many multiple-choice items, the stem is a question related to the text of the science story. For others, the stem includes a line from the engagement activity followed by a question. 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 where educators select the option that best describes the student's behavior in response to the item.

#### **III.1.A.vi. Alternate Testlets for Students Who Are Blind or Have Visual Impairments**

Alternate testlets, called BVI forms, were created when learning targets were difficult to assess online for students who had visual impairments, even with features such as read aloud or magnification. Computer-delivered BVI testlets begin with an instruction screen for the test administrator, then continue with content intended for the student to access. These testlets list materials that the educator may use to represent the onscreen content for the student. In

teacher-administered BVI testlets, test administrators receive recommendations for special materials to use with students who are blind or have visual impairments, but other familiar materials may be substituted. Details about needed materials for testlets delivered in both modes (computer-delivered and teacher-administered) are provided on the Testlet Information Page (see Chapter IV).

### **III.2. ESSENTIAL ELEMENT CONCEPT MAPS FOR TESTLET DEVELOPMENT**

ECD describes a conceptual framework for designing, developing, and administering educational assessments (Mislevy, Steinberg & Almond, 1999). The use of an ECD framework in developing large-scale assessments supports arguments for validity of the interpretations and uses of the assessment results. ECD requires test designers to make explicit the relationships between inferences that they want to make about student skills and understandings and the tasks that can elicit evidence of those skills and understandings in the assessment. The ECD approach is structured as a sequence of test development layers that include (a) domain analysis, (b) domain modeling, (c) conceptual assessment framework development, (d) assessment implementation, and (e) assessment delivery (Mislevy & Riconscente, 2005). Since the original introduction of ECD, the principles, patterns, examples, common language, and knowledge representations for designing, implementing, and delivering educational assessment using the processes of ECD have been further elaborated for alternate assessment (DeBarger, Seeratan, Cameto, Haertel, Knokey, & Morrison, 2011; Flowers, Turner, Herrera, Towles-Reeves, Thurlow, Davidson, & Hagge, 2015).

Item and testlet writing was based on Essential Element Concept Maps, a tool proven useful for item writers when first developed for ELA and mathematics (Bechard & Sheinker, 2012). Since the EECMs were shown to be valuable resources for ELA and mathematics item writing, they were subsequently adapted for science. These templates used principles of ECD to define science content specifications for assessment. Science content teams developed the content used within the EECM templates. Staff with student population expertise also reviewed EECMs. Item writers used the EECMs because they are content-driven guides on how to develop content-aligned and accessible items and testlets for the DLM student population. Each EECM defines the content and science practices framework of a Target EE with three levels of complexity and identifies key concepts and vocabulary at each level. They also describe and define common misconceptions, common questions to ask, and prerequisite and requisite skills. Finally, the EECMs identify accessibility issues related to particular concepts and tasks.

The EECM science templates were adopted by states in the DLM Science Alternate Assessment Consortium. After states approved the EECM structure, they were utilized for each EE in the development of assessments. The templates were specifically designed for clarity and ease of use as the project engaged non-professional item writers from participating consortium states who needed to create a large number of items in a constricted timeframe. Appendix C shows an example of an EECM.

The EECM has seven functions:

- Identify the targeted standard by domain, core idea, topic, science and engineering practices, and EE;
- Identify key vocabulary to use in testlet questions;
- Describe and define a range of skill development (three levels);
- Describe and define misconceptions;
- Identify prerequisite skills;
- Identify questions to ask; and
- Identify content through the use of accessibility flags that may require an alternate approach to assessment for some students.

Item writers were asked to look at each section of the EECM and do the following:

1. Review the content framework for the testlet set (the Domain, the Core Idea, the Topic, the NGSS Standard and the DLM EE);
2. Determine which level they were asked to write a testlet for (Target, Precursor, Initial), read the level description and look at the relationship of that level with the other levels within the EE;
3. Review the vocabulary and concepts at each level for ideas about how observable student behaviors will change from level to level to understand the distinctions between each level; and
4. Read the questions to ask and the misconceptions students may have about this construct.

More information on how the EECMs were used is provided in section III.3.E.

### **III.3. ITEM WRITING**

DLM items and testlets were developed in two sessions in 2015: one in January and one in July. Item writing occurred during item writing events where content and special education specialists worked on-site either in Lawrence or Kansas City, Kansas, to develop DLM assessments. In addition to item writers, DLM staff and graduate research assistants supported item writing efforts by developing supporting resources and EECMs, serving as internal reviewers, and in some cases, writing testlets.

#### ***III.3.A. RECRUITMENT AND SELECTION***

The item writer recruitment and selection process secured qualified and experienced individuals to write high-quality testlets, as shown in the following sections. Science content teams used several recruitment strategies to solicit applicants. An electronic recruitment survey

was sent to state partners to be distributed to science and special education educators in DLM member states. This recruitment survey included a brief description of the job and inquired about skills and availability. Additionally, the job description was sent to DLM science state partners for distribution. Content teams screened applicant materials, conducted interviews, and made hiring offers to selected candidates. Applicants were evaluated on the following required qualifications: experience with science academic content, ability to complete pre-workshop online training modules, and availability to attend the duration of the on-site workshop. The preferred qualifications included teaching experience in science, experience working with or instructing students with SCD, and experience with or knowledge of large-scale assessments, item development, state testing, and/or state standards. The hired applicants exhibited a balance of expertise in science and special education. All item writers signed security agreements and were trained on item security procedures.

### **III.3.B. ITEM WRITER CHARACTERISTICS**

The January 2015 item writing event had 42 item writers. There were 17 science item writers at the July 2015 item writing event, all of whom had previously attended in January.

An item writer survey was used to collect demographic information about the educators and other professionals who were hired to write DLM assessments during the 2015 item writing events. In total, 59 item writers responded to the item writer surveys across both events. Data gathered through this survey included years of teaching experience, grades taught, degree type, experience with the population, experience with alternate assessment based on alternate achievement standards (AA-AAS), and whether the item writer currently taught students eligible for AA-AAS. Each survey category is described below, with an accompanying table when applicable. Data were aggregated across both years. Table 9 shows the years of teaching experience for science item writers.

Table 9. Item Writers’ Years of Teaching Experience

	January 2015 Workshop (N = 42)		July 2015 Workshop (N = 17)	
	Median	Range	Median	Range
Science	9	1–34	9.5	1–30
Special Education	15	0–34	14	0–30
Students w/Significant Cognitive Disabilities	13	1–30	15	1–25

The January 2015 event had 13 item writers with high school teaching experience participate. There were 22 science item writers with experience at the elementary level, grades 3–5, and 25 with experience in middle school, grades 6–8.



The July 2015 event had eight science item writers with experience at the elementary level, grades 3–5; ten had experience with middle school, grades 6–8; and four had experience in high school. See Table 10 for a summary.

Table 10. Item Writers’ Grade-Level Teaching Experience

	January 2015 Workshop (N = 42)		July 2015 Workshop (N = 17)	
	<i>n</i>	%	<i>n</i>	%
Elementary	22	52.38	8	47.06
Middle School	25	59.52	10	58.82
High School	13	30.95	4	23.53

*Note:* Multiple grades could be selected on the survey. Percentages do not sum to 100%.

The 59 item writers represented a highly qualified group of professionals in the education and assessment field. Over 90% of the item writers held at least a bachelor’s degree. Master’s level degrees were held by 67% of the January item writers and 70% of the July item writers. Twelve item writers held a National Board certification. Table 11 shows the number and types of degrees held by item writers.

Table 11. Item Writers’ Level of Degree

	January 2015 workshop (N = 42)		July 2015 workshop (N = 17)	
	<i>n</i>	%	<i>n</i>	%
Bachelor's	10	23.81	4	20.53
Master's	28	66.67	12	70.59
Other	3	7.14	0	0.00

Most item writers had experience working with students with disabilities. In the January workshop, item writers had the highest levels of experience in the Emotional Disability, Mild Cognitive Disability, Severe Cognitive Disability, and Specific Learning Disability categories. In July, the highest levels of experience occurred in the Mild Cognitive Disability, Multiple Disabilities, and Specific Learning Disability categories. The disability categories of Blind/Low Vision and Deaf/Hard of Hearing had the fewest number of responses in both item writing groups. Traumatic Brain Injury also had the fewest number of responses in July. All disability categories reported on the survey are listed in Table 12.



Table 12. Item Writers’ Experience with Disability Categories

Content Area	January 2015 workshop (N = 42)		July 2015 workshop (N = 17)	
	<i>n</i>	%	<i>n</i>	%
Blind/Low Vision	13	30.95	4	23.53
Deaf/Hard of Hearing	12	28.57	5	29.41
Emotional Disability	28	66.67	10	58.82
Mild Cognitive Disability	31	73.81	11	64.71
Multiple Disabilities	25	59.52	11	64.71
Orthopedic Impairment	17	40.48	6	35.29
Other Health Impairment	26	61.90	9	52.94
Severe Cognitive Disability	29	69.05	10	58.82
Specific Learning Disability	29	69.05	11	64.71
Speech Impairment	27	64.29	9	52.94
Traumatic Brain Injury	16	38.10	4	23.53

*Note:* Multiple categories could be selected on the survey of item writers. Percentages do not sum to 100%.

Of the item writers, 64% had experience administering an alternate assessment based on alternate achievement standards (AA-AAS) prior to their work on the DLM project, with 75%, or 44 out of 59, reporting that at the time of the survey, they worked with students eligible for AA-AAS.

### ***III.3.C. ITEM WRITER TRAINING***

Training for item writers consisted of multi-day sessions at the beginning of the 2015 item writing events. Processes for test development were streamlined between the item writing events, which resulted in requiring less training for item writers in July.

Before beginning specific training on the writing process, item writers had training on confidentiality and signed security agreements (see Appendix C). After that, item writers were introduced to the DLM system and completed DLM professional development pre-workshop modules. Using the modules for training ensured that the item writers had a common level of knowledge about DLM and the student population before writing items. Modules focused on

assessment system design, population of students, and accessibility. There was a brief quiz at the end of each module that item writers were required to pass with 80% accuracy. Additional in-person training focused on science content.

Training was divided into sections that focused on accessibility, content development, use of images and graphics, bias and sensitivity, use of a cognitive process dimension taxonomy, and appropriate assignment of item metadata for the content management system in the Kansas Interactive Testing Engine (KITE) platform.

The science content teams, DLM test development staff internal reviewers, and editors were all involved in monitoring, mentoring, and retraining item writers to ensure the quality of the testlets produced. Editors evaluated the first testlet each item writer wrote and provided specific, individualized feedback during individual and group retraining sessions. Retraining opportunities were held, where content teams and editors identified patterns of errors or problems with content, accessibility, or bias and sensitivity.

The content teams led retraining sessions with item writers as needed, providing examples, visuals, and additional documentation. Internal reviewers also provided feedback (e.g., vocabulary too complex) for targeted retraining. Editors held periodic retraining sessions with item writers to review the most common errors and solutions for resolving them.

### ***III.3.D. ITEM WRITING RESOURCE MATERIALS***

Item writers used the EECMs to develop testlets at different linkage levels for each EE. In addition to the EECMs, item writers used materials developed by content teams to support the development of testlets. All item writers used the DLM Core Vocabulary list. Core vocabulary is made up of words used most commonly in expressive communication (Yorkston, et al., 1988). DLM Core Vocabulary is a comprehensive list of words, spanning grades K–12, that reflects the research on vocabulary in Augmentative and Alternative Communication (AAC) and includes words needed to successfully communicate in academic settings where the EEs are being taught (Dennis, Erickson & Hatch, 2013).

Additionally, all item writers used a guide to good practices in item writing, which included a checklist of common item writing challenges and errors. The content team prepared additional materials to support item writing, including materials prepared to support writing items for testlets designed for students who were blind or had visual impairments. Prototypes of testlets were used during training and available for item writer review. These prototypes went through multiple rounds of input from state partners and other stakeholders, internal content reviews, and editorial reviews. Prototypes were written at all three linkage levels and included examples of teacher-administered and computer-delivered testlets.

### ***III.3.E. ITEM WRITING PROCESS***

As noted above, item writers were given writing assignments for EEs, including all linkage levels outlined on the EECM. Because testlets were conceived as being a short set of coherent, instructionally relevant assessments, item writers produced entire testlets rather than stand-

alone items. Item writers frequently wrote testlets for the same EE at different linkage levels. Item writers were encouraged to use the DLM linkage level relationships in the EECM when thinking about the content of testlets at different linkage levels.

Item writers reviewed the vocabulary (concepts and words) on the EECM appropriate for each testlet level. Item writers were to assume that students would be expected to understand, but not necessarily use, these terms and concepts. Using the EECMs, item writers selected specific vocabulary for each testlet that matched the cognitive complexity of the learning target being assessed.

Item writers used the EECM information on “questions to ask” and “misconceptions” when writing testlets. The questions describe what evidence is needed to show that the student can move from one level to the next, more complex level. The information about possible misconceptions or errors in thinking provides examples that could be indicative of the level of understanding a student may have. These misconceptions can inform the selection of construct-relevant answer options for items. These EECM sections assisted the item writers in creating stems and answer options for items in testlets.

Item writers focused on all of the students who might receive each testlet and considered any accessibility issues. The goal for the item writer was to create testlets that were accessible to the greatest number of students possible, and to be specific about the conditions necessary to achieve that. Writers were prompted to ask questions such as, “Are there accessibility tools (online or offline) that may be necessary for some students?” They were also directed to consider barriers caused by sensitive nature of the content or bias that may occur, which could advantage or disadvantage a particular subgroup of students. Then, item writers focused on access to the testlet by asking, “Is this testlet designed for a particular group of students who will need a specific approach due to their disability?” Writers were asked specifically to think about students with sensory and mobility challenges.

During item development, item writers and DLM staff maintained the security of materials. Item writers all signed security agreements. Training about best practices to maintain test security was provided to item writers and staff. Materials were stored in locked facilities. Electronic transfers were made on secured network drives and within the secure content management system in KITE Client.

### ***III.3.F. ITEM WRITER EVALUATIONS***

An evaluation survey of the item writing experience was sent to all participating item writers after the 2015 item writing events. Item writers were asked to provide feedback on the perceived effectiveness of training and the overall experience in each item writing event, as well as narrative comments on their experience and suggestions for future DLM item writing events. Thirty-nine of the 42 (93%) item writers who participated in the January item writing event responded; 15 out of 17 (88%) item writers who participated in the July item writing event responded.

Of the 54 respondents, 22 felt training activities were very effective, 16 felt the first week of training was somewhat effective, and no one felt the training activities were not at all effective. In January, with the initial group, brainstorming with colleagues was seen as very effective by 37 out of 39 item writers. Contents of on-site training, feedback from DLM staff, and resource materials were perceived as very effective by 36 of the 39 item writers who responded. In July, all item writers saw peer review of their work as very effective for their writing. Table 13 and Table 14 show detailed responses to the perceived effectiveness questions from the item writer surveys from the January and July item writer groups, respectively.

Table 13. Perceived Effectiveness of Training for January 2015 Workshop (n = 39)

	Very Effective		Somewhat Effective		Not At All Effective	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Brainstorming with colleagues	38	97.44	1	2.56	0	0
Contents of final review checklist	24	61.54	15	38.46	0	0
Contents of on-site training	36	92.31	2	5.13	0	0
Feedback from DLM staff	36	92.31	3	7.69	0	0
Online training course activities	22	56.41	16	41.03	0	0
Online training course quizzes	19	48.72	18	46.15	1	2.56
Online training course supplementary materials	32	82.05	7	17.95	0	0
Online training course videos	23	58.97	15	38.46	1	2.56
Peer review process	26	66.67	13	33.33	0	0
Resource materials	36	92.31	3	7.69	0	0

Table 14. Perceived Effectiveness of Training for July 2015 Workshop (n =15)

	Very Effective		Somewhat Effective		Not At All Effective	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Brainstorming with colleagues	14	93.33	0	0	0	0
Contents of final review checklist	12	80.00	2	13.33	0	0
Contents of on-site training	14	93.33	1	6.67	0	0

	Very Effective		Somewhat Effective		Not At All Effective	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Feedback from DLM staff	13	86.67	2	13.33	0	0
Online training course activities	14	93.33	0	0	0	0
Online training course quizzes	11	73.33	3	20	0	0
Online training course supplementary materials	6	40.00	7	46.67	0	0
Online training course videos	5	33.33	10	66.67	0	0
Peer review process	15	100	0	0	0	0
Resource materials	13	86.67	0	0	0	0

Overwhelmingly, January item writers who responded agreed or strongly agreed that the overall goals and objectives for the item writing workshop were clear (38 out of 39, or 97%). In July, all 15 item writers agreed or strongly agreed that the overall goals and objectives for the item writing workshop were clear. Almost all January respondents (97%) and all July (100%) agreed or strongly agreed that the item writing process was a valuable professional development experience.

Table 15 and Table 16 show the responses to the overall experience questions from the survey from January and July science item writers, respectively.

Table 15. Overall Experience for January 2015 Workshop (n = 39)

	Strongly Agree		Agree		Disagree		Strongly Disagree	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
I am confident that the testlets I produced will be good assessments for students with significant cognitive disabilities.	28	71.79	9	23.08	1	2.56	1	2.56
I had enough time to complete my testlet assignments.	29	74.36	8	20.51	1	2.56	1	2.56
Other educators would find the testlets I wrote to	25	64.1	11	28.21	2	5.13	1	2.56

	Strongly Agree		Agree		Disagree		Strongly Disagree	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
be instructionally relevant.								
Overall, I valued the DLM item writing process as a professional development experience.	36	92.31	2	5.13	0	0	1	2.56
The <u>content</u> leaders were knowledgeable about academic content.	33	84.62	5	12.82	0	0	1	2.56
The content of the EECMs (questions, misconceptions) guided my decisions regarding testlet creation.	32	82.05	6	15.38	0	0	1	2.56
The overall goals and objectives for the item writing workshop were clear.	28	71.79	10	25.64	0	0	1	2.56
The <u>section</u> leaders were knowledgeable about testlet development procedures.	32	82.05	6	15.38	0	0	1	2.56

Table 16. Overall Experience for July 2015 Workshop (n = 15)

	Strongly Agree		Agree		Disagree		Strongly Disagree	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Discussing my testlets with colleagues helped me improve my testlets.	15	100	0	0	0	0	0	0
I am confident that the testlets I created will be good assessments for	9	60.00	5	33.33	0	0	0	0

	Strongly Agree		Agree		Disagree		Strongly Disagree	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
students with significant cognitive disabilities.								
I had enough time to complete my testlet assignments.	13	86.67	2	13.33	0	0	0	0
I would like to participate in future science item writing events (face-to-face and/or remote).	14	93.33	1	6.67	0	0	0	0
I would like to participate in other opportunities for DLM Science, such as standard setting.	14	93.33	1	6.67	0	0	0	0
Other educators would find the testlets I wrote to be instructionally relevant.	7	46.67	4	26.67	0	0	1	6.67
Overall, I valued the DLM item writing process as a professional development experience.	14	93.33	1	6.67	0	0	0	0
The content of the EECMs (questions, misconceptions, vocabulary, concepts, linkage level descriptions) guided my decisions regarding testlet creation.	12	80.00	3	20.00	0	0	0	0
The DLM leaders were knowledgeable about academic content.	11	73.33	2	13.33	0	0	0	0
The DLM leaders were knowledgeable about	9	60	5	33.33	0	0	0	0



	Strongly Agree		Agree		Disagree		Strongly Disagree	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
students taking the alternate assessment.								
The DLM leaders were knowledgeable about testlet development procedures.	14	93.33	1	6.67	0	0	0	0
The overall goals and objectives for the item writing workshop were clear.	13	86.67	2	13.33	0	0	0	0

### III.4. EXTERNAL REVIEWS

The purpose of external review is to evaluate items and testlets developed for the DLM alternate assessment in science. Using specific criteria established for DLM assessments, reviewers decided whether to recommend 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.

The external review process for science used the same procedures that were developed for ELA and mathematics. The DLM external review process for ELA and mathematics was piloted in a face-to-face meeting in Kansas City, Missouri in August 2013 before being implemented in the secure, online content management system in KITE Client.

Once the online external review capability was available in KITE Client, six educators tried out the online system. They used the online training and external review manual to guide their work as they evaluated testlets in the KITE system. DLM staff observed and provided assistance if the educator had difficulty with the platform or the rating process. The external review manual was revised to address those difficulties prior to finalizing the review materials. Since the face-to-face pilot process in 2013, DLM external reviews have been conducted with minor refinements online using the KITE Client. The external reviews of science testlets used the same general training, materials and procedures concurrent with ELA and mathematics.

### III.4.A. OVERVIEW OF REVIEW PROCESS

External reviews occurred after the initial internal reviews. Internal reviews involved a comprehensive editorial review and an internal content review by individuals with content expertise and/or experience with students with SCD. Figure 11 shows the order and relationship of reviews in the DLM test development process. Based on these initial reviews, DLM staff revised items as needed, performed a final editing review, and made final decisions. Each testlet was then sent for external review. External reviews were conducted online, independently, and asynchronously through an application in the secure content management system in KITE Client.

Resulting ratings were compiled with ratings from other reviewers and submitted to DLM staff, and DLM staff made final decisions regarding whether the testlet should be rejected, revised, or accepted as is before pilot/field-testing.

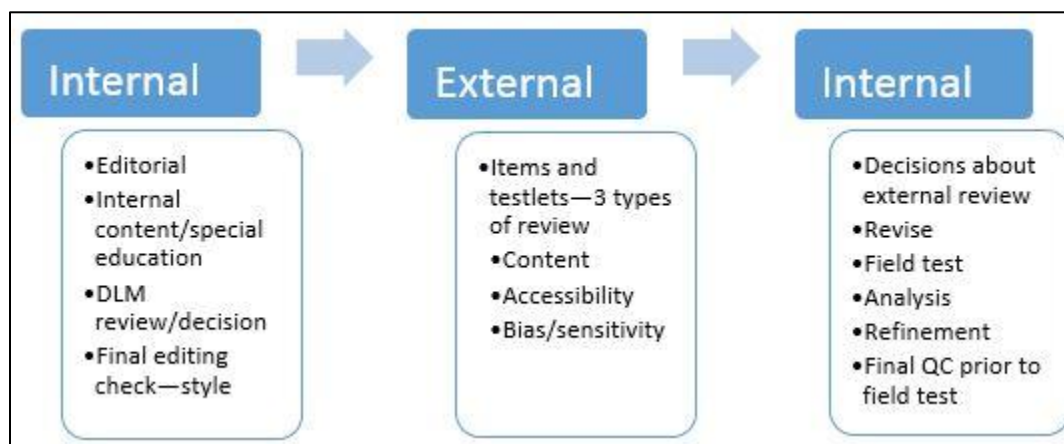


Figure 11. Overview of the item review processes prior to field testing for the DLM Alternate Assessment System.

External reviews were conducted by members of three distinct review panels: content, accessibility, and bias and sensitivity. Reviewers were assigned to one type of review panel and used the criteria established for that panel to conduct reviews. Reviewers evaluated items grouped together in testlets. For each item and each testlet, reviewers made one of three decisions: accept, requires critical revision, or reject. Reviewers made decisions independently and without discussion with other reviewers.

Reviews of testlets for students who are blind or have visual impairments were also conducted during the 2015–2016 academic year. These testlets were assigned to volunteers who had experience working with students with SCD or experience working with students who are blind or have low vision. The results of these reviews are included with the results of the other external reviews in the following sections.

### ***III.4.B. REVIEW ASSIGNMENTS AND TRAINING***

For external reviews in 2015-2016, 136 people responded to a volunteer survey used to recruit panelists. Volunteers for the external review process completed a Qualtrics survey to capture demographic information as well as information about their education and experience. This data are then used to identify panel types for which the volunteer would be eligible. Of the 136 respondents, 71 people were eligible and completed the required training, and nine of those were placed onto science external review panels. Each reviewer was assigned to one of the three panel types. Of the nine science reviewers, three were assigned to accessibility panels, three to content panels, and three to bias and sensitivity panels.

The current professional roles reported by reviewers indicated that eight were classroom educators and one was an instructional coach. Science reviewers had a median of 3 years of experience teaching students with SCD.

Review assignments were made throughout the year. Reviewers were notified by e-mail each time they were given an assignment of collections of testlets. Each review assignment took 1.5 to 2 hours. In most cases, reviewers had two weeks to complete an assignment.

Before reviewing testlets, participating reviewers were required to complete several online training modules. These modules included detailed instructions on the review process, security expectations, a quiz, and a practice activity. This training had to be completed successfully before reviewers began reviewing for the year. Training was completed in segments, taking 60 to 75 minutes total. Training information was made available online.

### ***III.4.C. REVIEWER RESPONSIBILITIES***

The primary responsibility for reviewers was to review testlets using established standards and guidelines. These standards and guidelines are found in the *Guide to External Review of Testlets* (Dynamic Learning Maps, 2014a). Reviewers completed a security agreement before reviewing and were responsible for maintaining the security of all materials at all times.

### ***III.4.D. DECISIONS AND CRITERIA***

External reviewers looked at testlets and made decisions about both the items in a testlet, and the testlet overall. The overview of the decision-making process is described below.

### III.4.D.i. General Review Decisions

For DLM assessments, “acceptability” at the external review phase was defined as meeting minimum standards to be ready for field testing. Reviewers made one of three general decisions: accept, revise, or reject. The definition of each decision is summarized in Table 17.

Table 17. General Review Decisions for External Reviews

<b>Decision</b>	<b>Definition</b>
Accept	Item/testlet is within acceptable limits. It may not be perfect, but it is worth putting through field tests and seeing how it goes.
Critical Revision Required (Revise)	Item/testlet violates one or more criteria. It has some potential merits and can be acceptable for field-testing after revisions to address the criteria.
Reject	Item/testlet is fatally flawed. No revision could bring this item/testlet to within acceptable limits.

Judgments about items were made separately from judgments about testlets because different criteria were used for items and testlets. Therefore, it was possible to recommend revisions or rejections to items without automatically having to recommend revision or rejection to the testlet as a whole. If a reviewer recommended revision or rejection, he or she was required to provide an explanation that identified the problem and, in the case of revision, proposed a solution.

### III.4.D.ii. Review Criteria

The criteria for each type of panel (i.e., content, accessibility, bias and sensitivity) were different. All three panel types had criteria to consider for items and other criteria for testlets as a whole. Training on the criteria was provided in the online training modules and in the practice activity. There were specific criteria for external reviewers of content, accessibility, and bias and sensitivity. Figure 12, Figure 13, and Figure 14 show the review criteria.

**CRITERIA FOR CONTENT PANELS**

*Items*

1. The item assesses the content of the targeted node.
2. The level of DOK required in the node matches the DOK identified for the item.
3. The content of the item is technically correct (wording and graphics).
4. Item answer options should contain only one correct answer (the key), distractors are incorrect and not misleading, and nothing in the item cues the correct response.
5. The item type is logical and appropriate for the content being assessed and the graphics contribute to the quality of the item.

*Testlets*

6. The testlet is instructionally relevant to students for whom it was written and is grade level appropriate.
7. If items are embedded within text, items are placed within the text at logical places and conclusion items are placed at the end.
8. If text is provided, the text's content provides an appropriate level of challenge. It is reduced in depth, breadth and complexity from grade level. The text is written to align the tasks in the testlet.
9. Elements of graphics or diagrams such as perspective or dimension do not conflict with information in the text or other graphics or diagrams used in the text.

Figure 12. Content review criteria.

***CRITERIA FOR ACCESSIBILITY PANELS***

*Items*

1. The text within the item provides an appropriate level of challenge and maintains a link to grade-level content without introducing unnecessary, confusing, or distracting verbiage. The text uses clear language and minimizes the need for inferences and prior knowledge to comprehend the content.
2. Graphics are clear and do not introduce confusion. Graphics can be presented in tactile form.

*Testlets*

3. The testlet is instructionally relevant to students for whom it was written and is grade level appropriate.
4. The testlet does not introduce barriers for students with (a) limited working memory, (b) communication disorders dependent on spoken English grammatical structures, or (c) limited implicit understandings of others' emotions and intentions.
5. If text is provided, it uses clear language and minimizes the need for inferences and prior knowledge to comprehend the content. The text does not introduce unnecessary, confusing, or distracting verbiage.

Figure 13. Accessibility review criteria.

**CRITERIA FOR BIAS & SENSITIVITY PANELS**

*Items*

1. Item does not require prior knowledge outside the bounds of the targeted content.
2. Where applicable, there is a fair representation of diversity in race, ethnicity, gender, disability, and family composition.
3. Stereotypes are avoided. Appropriate labels are used for groups of people. People-first language is used for individuals with disabilities.
4. Language used does not prevent nor advantage any group from demonstrating what they know about the measurement target.
5. Item does not focus on material that is likely to cause an extreme emotional response.

*Testlets (sensitivity criteria)*

6. Testlet is free of content that is controversial, disturbing, or likely to cause an extreme emotional response due to issues of culture, region, gender, religion, ethnicity, socio-economic status, occupation, or current events.
7. The language in the testlet neither prevents nor disadvantages any regional or cultural group from demonstrating what they know about the targeted content. People first language is used for individuals with disabilities. Populations are not depicted in a stereotypical manner.
8. The text represents the topic accurately without requiring prior knowledge.
9. Where applicable, there is a fair representation in the text of diversity in race, ethnicity, gender, disability and family composition.
10. The text does not contain sensitive topics. (See Appendix for the DLM list of topics not to be used in testlets).

Figure 14. Bias and sensitivity review criteria.

All three types of reviews focused on both items and testlets. Content reviews of items included alignment of the item and learning target, level of cognitive process dimension, quality and appropriateness of the content, accuracy of response options, and appropriateness of distractors. Testlet content reviews also focused on the instructional relevance to students and grade-appropriateness, as well as the logic of item placement within science story text.

Accessibility item reviews focused on appropriate challenge levels and the maintenance of links to grade band content. For accessibility reviews, testlets were checked for instructional relevance at grade level and minimizing of barriers to students with specific needs.

Finally, item-level bias and sensitivity reviews included identifying items that require prior knowledge outside the bounds of the targeted content, ensuring fair representation of diversity, avoiding stereotypes and negative naming, removing language that affects a student's demonstration of their knowledge on the measurement target, and removing any language that was likely to cause strong emotional response. For testlet bias and sensitivity reviews, criteria similar to item-level reviews were applied, with emphasis on reducing the chance of construct-



irrelevant variance due to inadvertent use of controversial, disturbing, stereotypic, or negative language or graphics.

### **III.4.E. RESULTS OF REVIEWS**

The percentage of items or testlets rated as “accept” ranged across grades, pools, and rounds of review from 80% to 93%. The rate at which content was recommended for rejection ranged from approximately 0% to 4% across grades, pools, and rounds of review. A summary of the content team decisions and outcomes is provided below. A more detailed report and outcomes from external reviews are included in the external review technical report (Clark, Beitling, Bell, & Karvonen, 2016).

#### **III.4.E.i. Content Team Decisions**

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

Following this process, content teams made a final decision to accept, revise, or reject each of the items and testlets. The science content team retained almost 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 option replaced). The science team made a total of 85 minor revisions to items and 52 minor revisions to testlets.

### **III.5. THE FIRST CONTACT SURVEY**

The linkage level for the student’s first testlet is determined based on responses to the First Contact survey. The First Contact survey is a survey of learner characteristics that covers a variety of areas, including communication, academic skills, attention, and sensory and motor characteristics. A completed First Contact survey is required for each student prior to the assignment of assessments. Supporting procedures and a complete list of First Contact survey questions are included in the *Test Administration Manual 2015-16* (Dynamic Learning Maps, 2016a). Test administrators are trained on the role of First Contact survey in testlet assignment as part of required test administrator training (see Chapter X).

For the 2015-2016 DLM science administration, one section of the First Contact survey was used to provide a match between student and testlet during the initial DLM science testing experience—the expressive communication section. Two other sections of the First Contact survey address students’ academic skills in English language arts and mathematics and are used, in conjunction with the expressive communication section, to assign assessments in the respective content areas. Responses to survey items in each category are used to calculate a

complexity band for the student, that is then matched to a linkage level (see Chapter IV for additional information).

While development of science academic skills survey questions and complexity band calculation method were underway,<sup>1</sup> only the expressive communication section was used for science testlet assignment. The expressive communication section results have a high rate of correspondence with the academic skills section results (91% to 95% across complexity bands; see Clark, Kingston, Templin, & Pardos, 2014, p. 6) and was therefore a reasonable choice for placing students into a new science assessment.

The student's assigned complexity band is calculated automatically and stored in the KITE system. The goal is to present a testlet that is approximately matched to a student's knowledge, skills, and understandings. That is, within reason, the system should present a testlet that is neither too easy nor too difficult and should provide a positive experience for the student entering the assessment. Based on the student's assigned complexity band, the student's first testlet could be delivered at one of three levels. The Foundational band, or Band 1, will deliver a testlet written at the Initial level, which is appropriate for students who either do not use speech, sign, or AAC or use one word, sign, or symbol to communicate. Band 2 will deliver a testlet at the Precursor level for students who use two words, signs, or symbols to communicate. Band 3 will deliver a testlet at the Target level for students who regularly combine three or more spoken words to communicate for a variety of purposes.

### **III.6. PILOT ADMINISTRATION**

The spring 2015 DLM science pilot testing window was from April 22 through June 5, 2015 and included Iowa, Kansas, Missouri, and Oklahoma. States were able to select their own windows within the consortium-wide window if needed. Results from the spring pilot tests were used for research and development purposes only and were not reported this year.

The purpose of the pilot test was to evaluate the new science testlet content. To be eligible for the DLM science pilot test, students needed to be in grades 3-12, have the most significant cognitive disabilities, and be eligible for their state's current alternate assessment based on alternate achievement standards. Students were enrolled based on their current grade level within one of the three science grade bands or within the End-of-Instruction (EOI) biology assessment if the student was in a state participating in EOI. States were encouraged to implement the same eligibility guidelines for alternate assessment participation in English language arts and mathematics for the science pilot test. All computer-delivered testlets included read aloud capability; however, the pilot test was not specifically designed for students who are blind or have visual impairments.

---

<sup>1</sup> During the fall 2015 science field test educators were surveyed about their students' science academic skills using possible science-related questions. After analyses are completed, these questions will be used to calculate a science-specific complexity band for the 2016-2017 administration.

The linkage level was chosen for each student based on information from the students’ First Contact survey (described above). For the spring pilot test, only the expressive communication questions were used for testlet linkage level assignment. This assignment was the same for all administered testlets. That is, the testlets a student received were all at the same linkage level (in operational administrations, student’s testlet linkage level may vary based on how the student performed on the previous testlet.).

All students were assigned testlets that covered the entire blueprint. During the spring pilot test, students received a fixed form test that contained either nine or 10 testlets at the same linkage level (i.e., 10 for biology and nine for all other grade bands) depending on the blueprint. One fixed form test was available at each grade band and biology. Each testlet included three to four items related to one EE in the blueprint.

A total of 1,605 students from Iowa, Kansas, Missouri, and Oklahoma completed assessments. The total number of participants by grade band is presented in Table 18. The table indicates that 36% of students were in elementary (grades 3–5), 35% were in middle school (grades 6–8), and 29% were in high school (grades 9–12).

Table 18. Number of Participants in the Spring 2015 Science Pilot Test by Grade Band

Grade Band	Students
Elementary	577
Middle School	562
High School	448
Biology	20
<b>Total</b>	<b>1,607</b>

*Note.* Oklahoma administers an end-of-course biology test at the high school grade band.

Following the pilot test, item statistics were computed for all items and testlets. Specifically, a percent correct ( $p$ -value) was calculated for every item and a  $z$ -score was calculated for every item to reflect the standardized difference between the item’s  $p$ -value and the weighted average  $p$ -value for items within the testlet. Given the intended population and purpose of DLM assessments, it was determined that a discrimination index would not be included as an item statistic since the intent was not to differentiate between generally high and low performers. Using these item statistics, items were flagged for further review.

Items were flagged for review if they met either of the following statistical criteria:

- Too challenging: percent correct ( $p$ -value) less than 35%.
- Significantly easier or harder than other items within the same testlet (standardized difference): any  $p$ -value greater than two standard errors from the mean  $p$ -value.

Data review was conducted by the content team and included a review of the item statistics (including proportion of students selecting each answer option) alongside the item content. Flagging criteria served as one source of evidence for the content teams in evaluating item quality. Final judgments were content based. The team reviewed items that had a sample size of at least 20 cases. Due to low participation ( $n < 20$ ) in biology, item statistics were not calculated; rather all biology items were examined using insights gained from the review of other items.

Flagged items were discussed and possible causes for the flag were considered. Group consensus was used to make item-level decisions. Options included (1) no change to item; (2) identify concerns that require item modifications, are clearly identifiable, and can improve the item; (3) identify concerns that require item modification, are not clearly identifiable, but the content is worth preserving; or (4) reject item because it is not worth revising. After item-level decisions were made, testlets for items assigned to options three or four were evaluated to determine if the testlet would be retained or rejected.

Table 19 reports the percentage of flagged items from the total number of eligible items for each grade band. Table 20 displays the decisions that were made by the content team as a result of the data review and additional review of biology items. Across grade bands, approximately 15% of items were flagged and the overall rejection rate was 31%.

Table 19. Item Flags for Content Administered During the 2015 Science Spring Pilot Test

<b>Grade Band</b>	<b># Flagged Items</b>	<b># Eligible Items</b>	<b>% Flagged</b>
Elementary	13	83	15.7
Middle School	12	83	14.5
High School	13	85	15.3
Biology*	NA	NA	NA
<b>Total</b>	<b>38</b>	<b>251</b>	<b>15.1</b>

Note. \*Sample sizes < 20 for all biology testlets

Table 20. Content Team Response to Item Flags for the 2015 Science Spring Pilot Test

Grade Band	# Reviewed Items	Accept	Revise	Reject
Elementary	13	0	6	7
Middle School	12	1	5	6
High School	13	1	6	6
Biology*	27	20	6	1
Total	65	22	23	20
<b>Percentage of Total</b>		<b>33.8%</b>	<b>35.4%</b>	<b>30.8%</b>

Note. \*Sample sizes < 20 for all biology testlets—all items were included in the content review.

Of the 38 flagged items, 27 (71%) were at the Precursor level. This finding led the content team to examine the Precursor testlets to determine possible causes for higher difficulty of Precursor testlets. Linkage level descriptors at the Precursor level ask students to use more complex skills than the Initial level, such as developing models and making claims that are supported by evidence. The content team decided that the difficulty of Precursor level testlets could be reduced while still assessing the skills that are described by the linkage level if more context was provided to students. Science stories were used to provide this context and activate students' prior knowledge in revised testlets. Revisions to biology items generally involved accessibility of tables and graphs, as well as consistency of format and presentation. All items and testlets that were revised were included in the fall 2015 field test.

### III.7. 2015 FALL FIELD TEST

The fall 2015 DLM science field testing window was from November 9 through December 9, 2015. Participating states included Iowa, Kansas, Missouri, Oklahoma, West Virginia, and Mississippi. Results from the fall field tests were used both for research and development purposes as well as to contribute to the data for the spring 2016 model parameter calibrations. A science survey was also administered to a sample of field test participants, and results were used for research and development purposes.

The purposes of the fall field test were to

- Evaluate new and edited science testlet content;

- Pilot new science academic skills questions for the First Contact survey and use data to inform the development of a method for assigning an appropriate first testlet based on students’ science academic skills;<sup>2</sup>
- Gather cross-linkage level data to evaluate relationships and support modeling research; and
- Evaluate students’ opportunity to learn science content and practices, science academic skills, and experience using the DLM science assessment system.

The eligibility criteria for the fall field test were the same as the pilot test with one exception: the 2015 fall field test was also designed for students who are blind or have visual impairments.

The linkage level was chosen for each student based on information from the student's First Contact survey. For the fall field test, only the expressive communication questions were used for testlet linkage level assignment. This assignment placed students into one of three science linkage levels. For each linkage level, several fixed-test forms were available for administration, and each form contained two testlets at the assigned linkage level and one testlet at an adjacent linkage level, similar to field test procedures in ELA and math. That is, all students received two testlets at their assigned linkage level and one testlet at a higher or lower linkage level. For biology, a fixed form had seven testlets, four of which were at the assigned linkage level and three at an adjacent (higher or lower) level. Testlets did not cover the entire blueprint. Each testlet included three to four items related to one EE in the blueprint.

A combination of new content developed at the July item writing workshop and revised content from the spring pilot served as the content field-tested in the fall. One new testlet at each EE and linkage level was field tested, with the goal of expanding the operational item bank to have two testlets for every EE and linkage level.

Table 21 presents an example of the matrix design for one grade band and the Life Science (LS) domain employed for the 2015 science fall field test.

Table 21. 2015 Science Fall Field Test Sampling Design Example – Life Science

Complexity Band	Form	Initial			Precursor			Target		
		LS1	LS2	LS3	LS1	LS2	LS3	LS1	LS2	LS3
Bands 1 & 2	1	X	X		X					
	2		X	X		X				
	3	X		X			X			

<sup>2</sup> A research report is planned for after the 2016-2017 science administration which will implement the new science-specific complexity band.

Complexity Band	Form	Initial			Precursor			Target		
		LS1	LS2	LS3	LS1	LS2	LS3	LS1	LS2	LS3
Band 3	4	X			X	X				
	5		X			X	X			
	6			X	X		X			
	7				X	X		X		
	8					X	X		X	
	9				X		X			X
Band 4	10				X			X	X	
	11					X			X	X
	12						X	X		X

In the 2015 fall field test, each EE and linkage level was assessed by one testlet. In total, 111 testlets were tested, each consisting of three or four items. The number of testlets by grade band is presented in Table 22.

Table 22. Number of Testlets by Grade Band for Fall 2015 Field Test

Linkage Level	Elementary	Middle School	High School	Biology*	Total
Initial	9	9	9	10	37
Precursor	9	9	9	10	37
Target	9	9	9	10	37
<b>Total</b>	<b>27</b>	<b>27</b>	<b>27</b>	<b>30</b>	<b>111</b>

*Note.* Three of the EEs overlap with the high school blueprint; there are seven unique EEs on the biology blueprint.

A total of 5,613 students participated in the 2015 field test. The total number of participants by grade band is presented in Table 23.



Table 23. Number of Participants in the Fall 2015 Science Field Test by Grade Band

Grade Band	Students	Percent
Elementary (Grades 3–5)	1,718	30.6
Middle School (Grades 6–8)	1,869	33.3
High School (Grades 9–12)	1,958	34.9
Biology <sup>a</sup>	68	1.2
<b>Total</b>	<b>5,613</b>	<b>100</b>

<sup>a</sup> Only Oklahoma participated in biology.

Table 24 displays the demographic summary for the field test participants by gender, primary disability label, comprehensive race, Hispanic ethnicity, and ESOL. Approximately 65% of students were male, 69% did not indicate a primary disability, 74% were white, 94% were not of Hispanic ethnicity, and 98% of students were not eligible or monitored for ESOL.

Please note that the primary disability field is not currently a required field for educators to complete. Also note that braille and large print were not available for the field test. However, students who indicated visual impairment as an accessibility flag in their Personal Needs and Preferences (PNP) Profile were assigned to testlets that were specifically designed to remove any visual barriers.

Table 24. Demographic Summary of Students Participating in the Fall 2015 Science Field Test

Demographic	Number	Percent
<b>Gender</b>		
Female	1,978	35.24
Male	3,635	64.76
<b>Primary Disability</b>		
Autism	372	6.63
Deaf-Blindness	3	0.05
Developmental Delay	3	0.05
Documented Disability	165	2.94
Emotional Disturbance	21	0.37
Hearing Impairment	1	0.02
Intellectual Disability	615	10.96

<b>Demographic</b>	<b>Number</b>	<b>Percent</b>
Multiple Disabilities	156	2.78
No Disability	2	0.04
Orthopedic Impairment	16	0.29
Other Health Impairment	86	1.53
Specific Learning Disability	20	0.36
Speech or Language Impairment	8	0.14
Traumatic Brain Injury	13	0.23
Visual Impairment	3	0.05
Missing	4,129	73.56
<b>Race</b>		
White	4,176	74.40
African American	1,056	18.81
Asian	114	2.03
American Indian	95	1.69
Alaska Native	19	0.34
Two or More Races	126	2.24
Native Hawaiian or Pacific Islander	16	0.29
Missing	11	0.20
<b>Hispanic Ethnicity</b>		
No	5,288	94.21
Yes	322	5.74
<b>ESOL Participation</b>		
Not ESOL eligible/monitored student	5,508	98.13
ESOL eligible/monitored student	105	1.87

Following the field test, item statistics were re-computed for all items and testlets, and the same process and criteria that were used for data review of the pilot test were followed.

Table 25 and Table 26 report the results of the data review. The number of items flagged out of the number eligible indicates that approximately 26% of eligible items were flagged for further review based on item performance. The content team reviewed all flagged items and made

decisions accordingly. Of those flagged items, 20% were not revised, 68% were revised, and almost 11% were rejected from the item pool.

Table 25. Item Flags for Content Administered During the 2015 Science Fall Field Test

Grade Band	# Flagged Items	# Eligible Items	% Flagged
Elementary	19	81	23.5
Middle School	26	85	31.0
High School	29	90	28.9
Biology <sup>a</sup>	0	23	0.0
<b>Total</b>	<b>74</b>	<b>279</b>	<b>26.5</b>

<sup>a</sup> Sample sizes < 20 for all Initial and Precursor level biology testlets.

Table 26. Content Team Response to Item Flags for the 2015 Science Fall Field Test

Grade Band	Accept	Revise	Reject
Elementary	5	14	0
Middle School	2	19	5
High School	8	17	3
Biology	NA	NA	NA
Total	15	50	8
<b>Percentage of Total (%)</b>	<b>20.30</b>	<b>68.0</b>	<b>10.80</b>

Based on the pattern of findings from the data review, the content team determined that the decision from the pilot test results to add context through science stories, particularly at the Precursor linkage level, was effective at improving student performance. In some cases text was revised in testlets with flagged items to be more concise and clear. Unnecessarily difficult vocabulary was removed. Recommendations were also made for future development to reduce the text complexity in the Initial level testlets, particularly in the test administrator directions to the student (e.g., “Show me the one that changes from a solid to a liquid”).

As students were administered testlets at two adjacent linkage levels for the same EE, evaluations could be made regarding the ordering of the levels in terms of difficulty. To accomplish this a weighted average testlet difficulty across linkage levels for students assigned to the same complexity band was calculated for each linkage level. Table 27 show the average proportion of correct responses to items weighted across all testlets within a linkage level organized by complexity band.

Table 27. Average Proportion Correct on Testlets by Complexity Band for each Grade Band

Grade Band	Initial	Precursor	Target
<b>Elementary</b>			
Foundational & Band 1	0.45	0.32	N/A
Band 2	0.67	0.53	0.48
Band 3	N/A	0.67	0.57
<b>Middle</b>			
Foundational & Band 1	0.44	0.34	N/A
Band 2	0.71	0.47	0.45
Band 3	N/A	0.57	0.55
<b>High School</b>			
Foundational & Band 1	0.37	0.30	NA
Band 2	0.66	0.51	0.44
Band 3	NA	0.67	0.56

*Note.* As described in Chapter IV, Foundational and Band 1 are both assigned to the initial linkage level in science. Biology testlets were not included due to low sample sizes (< 20).

These results provide preliminary evidence to support the ordering of linkage levels and the increasing level of difficulty from the initial to the target level. There is a general trend in decreasing p-values across higher linkage levels (i.e., increasing average item difficulty), albeit slight in some cases (e.g., between middle school precursor and target levels). It is important to note that placement into complexity band for the field test was based on students' expressive communication skills only (see Chapter IV for description of complexity band assignment method).

### **III.7.A. FIELD TEST SURVEY**

As part of the field test administration, a survey was also administered to educators in order to obtain feedback on their students' science academic skills, opportunity to learn science content, and overall experience with the science field test. Students were randomly selected and enrolled to participate in the survey. If a student was enrolled in the survey, the rostered educator would complete the survey questions about that student. Of the 2,037 students that were enrolled in the survey, 837 had completed surveys, for a response rate of approximately 41%.

Table 28 displays the demographic data for the students whose educators responded to the fall field test survey. Students were primarily male, White, and non-Hispanic.

Table 28. Demographic Summary of Students Whose Educators Participated in the Science Field Test Survey

<b>Demographic</b>	<b><i>n</i></b>	<b>%</b>
<b>Gender</b>		
Female	281	33.57
Male	556	66.43
<b>Primary Disability</b>		
Autism	26	3.11
Deaf-Blindness	0	0.00
Developmental Delay	1	0.12
Documented Disability	35	4.18
Emotional Disturbance	2	0.24
Hearing Impairment	0	0.00
Intellectual Disability	47	5.62
Multiple Disabilities	14	1.67
No Disability	0	0.00
Orthopedic Impairment	1	0.12
Other Health Impairment	3	0.36
Specific Learning Disability	1	0.12
Speech or Language Impairment	0	0.00
Traumatic Brain Injury	1	0.12
Visual Impairment	0	0.00
Missing	706	84.35
<b>Race</b>		
White	650	77.66
African American	121	14.46
Asian	18	2.15

<b>Demographic</b>	<b><i>n</i></b>	<b>%</b>
American Indian	28	3.35
Two or More Races	15	1.79
Native Hawaiian or Pacific Islander	2	0.24
Missing	3	0.36
<b>Hispanic Ethnicity</b>		
No	777	92.83
Yes	60	7.17
<b>ESOL Participation</b>		
Not ESOL eligible/monitored student	812	97.01
ESOL eligible/monitored student	25	2.99

There were three sections in the survey. The first section asked educators to indicate how consistently each of their students used specific science academic skills during science instruction. Table 29 shows the number and percentage of students that educators perceived who demonstrated each skill on a scale ranging from "never" to "consistently." In general, most students could sort objects by common properties, identify similarities and differences, and recognize patterns at least 21-50% of the time or more. Conversely, most students never or almost never compared initial and final conditions to determine change, use data to answer questions, identify cause and effect relationships, identify evidence to support a claim, or use diagrams to explain phenomena.

Table 29. Perceived Consistency of Student Skill during Science Instruction

<b>Skill</b>	<b>Never or almost never (0-20%)</b>		<b>Occasionally (21-50%)</b>		<b>Frequently (51-80%)</b>		<b>Consistently (81-100%)</b>		<b>Missing</b>	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Sort objects or materials by common properties	226	27.5	240	29.2	232	28.2	124	15.1	15	1.8
Identify similarities and differences	310	37.9	306	37.4	162	19.8	41	5.0	18	2.2
Recognize patterns	319	38.9	295	35.9	154	18.8	53	6.6	16	1.9
Compare initial and final conditions to	462	56.1	245	29.8	99	12.0	17	2.1	14	1.7

Skill	Never or almost never (0-20%)		Occasionally (21-50%)		Frequently (51-80%)		Consistently (81-100%)		Missing	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
determine if something changed										
Use data to answer questions	482	58.6	239	29.0	89	10.8	12	1.6	14	1.7
Identify cause and effect relationships	489	59.6	245	29.9	72	8.8	14	1.7	17	2.0
Identify evidence that supports a claim	564	68.8	198	24.2	51	6.2	7	0.9	17	2.0
Use diagrams to explain phenomena	583	71.3	175	21.4	51	6.2	9	1.1	19	2.3

The second section of the survey asked educators to indicate the average number of hours they either spent on instruction or planned for instruction on science content during the 2015-16 school year and is provided in Chapter XI. Overall, the majority of educators spent on average 10 or fewer hours of instruction on most science topics during the 2015-16 school year.

The third section of the survey asked educators to respond to questions regarding their students' experiences using the DLM science assessment system. Specifically, educators were asked about PNP features that met their students' accessibility needs, as well as factors that negatively and positively impacted their students' experiences using the system. Table 30, Table 31 and Table 32 below summarize responses to these questions.

Of the 837 students who used at least one PNP feature listed in Table 30, almost 60% had their accessibility needs met by using synthetic read aloud with sentence highlighting while about 8% had their needs met by using a single or two-switch system; both of which were either in conjunction with other features or as the only features used.

Table 30. Personal Needs and Preferences Profile (PNP) Features That Met Students' Accessibility Needs (N=837)

PNP Features	<b>n</b>	<b>%</b>
Synthetic read aloud with sentence highlighting (Text to Speech)	495	59.1
Magnification	99	11.8



Other display changes (color contrast, reverse contrast)	97	11.6
Switch (single switch or two-switch system)	66	7.9

*Note.* Educators were allowed to select multiple responses.

With respect to factors that impacted students’ assessment experiences, most educators felt that their students had not yet learned the topics covered by the assessments, the items did not correspond to their students’ true knowledge and skills, and that the engagement activities and vocabulary were too complex, which negatively impacted the experience (Table 31). Conversely, the majority of educators believed that the instructions for the test administrator were clear, and that this positively impacted students’ experiences (Table 32).

Table 31. Factors That Negatively Impacted Students’ Assessment Experience (N=837)

<b>Factors</b>	<i>n</i>	%
Student has not yet learned the topics covered by the assessments	523	62.5
The items did not correspond to the student's true knowledge, skills, and understandings	447	53.4
Complexity of the engagement activity	437	52.2
The vocabulary used in the testlets was too complex	418	49.9
Student has had limited experience with a computer	141	16.9
Too many testlets	126	15.1
Use of video as the engagement activity	65	7.8
Instructions to the test administrator were not clear	54	6.5

*Note.* Educators were allowed to select multiple responses.

Table 32. Factors That Positively Impacted Students' Assessment Experience (N=837)

<b>Factors</b>	<i>n</i>	%
Clear instructions to the test administrator	440	52.6
Use of video as the engagement activity	273	32.6
Quality of the engagement activity	269	32.1
This student was instructed in the areas covered by the assessments	190	22.7
The student was familiar with the vocabulary used in the testlets	189	22.6
The items corresponded to the student's true knowledge, skills, and understandings	148	17.7
Intuitiveness of the assessment system	109	13.0

*Note.* Educators were allowed to select multiple responses.

### III.8. OPERATIONAL ASSESSMENT ITEMS FOR 2015-2016

Operational assessments were administered during the spring window. Table 33 presents the participation numbers. One test session is one testlet taken by one student. Only test sessions that were complete or in progress at the close of the window counted towards the total test sessions by model.

Table 33. Operational Window Participation

<b>Participation</b>	<i>N</i>
Test Sessions	173,656
Students	21,470
Educators	8,190
Schools	5,805
Districts	2,068

Testlets were made available for operational testing following promotion from pilot or field test item review. Table 34 summarizes the total number of operational testlets by grade band for 2015-2016. There were a total of 103 operational testlets available. This included nine testlets shared between the high school and biology pools and also one EE/linkage level combination that had more than one testlet available during an operational window due to having both a BVI and general version of the testlet available.

Table 34. 2015–16 Science Operational Testlets

<b>Grade Band</b>	<b><i>n</i></b>
Elementary	27
Middle school	28
High school	27
Biology <sup>a</sup>	21
<b>Grand Total</b>	<b>103</b>

<sup>a</sup> Biology consisted of 30 testlets; however, nine of those testlets were also in the high school pool and therefore were removed from the biology counts.

Similar to the field test item review, *p*-values were calculated for all operational items to provide information about item difficulty. Figure 15 includes the *p*-values for each operational item for science. The sample size cutoff for inclusion in the *p*-values plot was 20, to prevent items with small sample size from potentially skewing the results. In general, most items had *p*-values that ranged from 0.50 to 0.59.

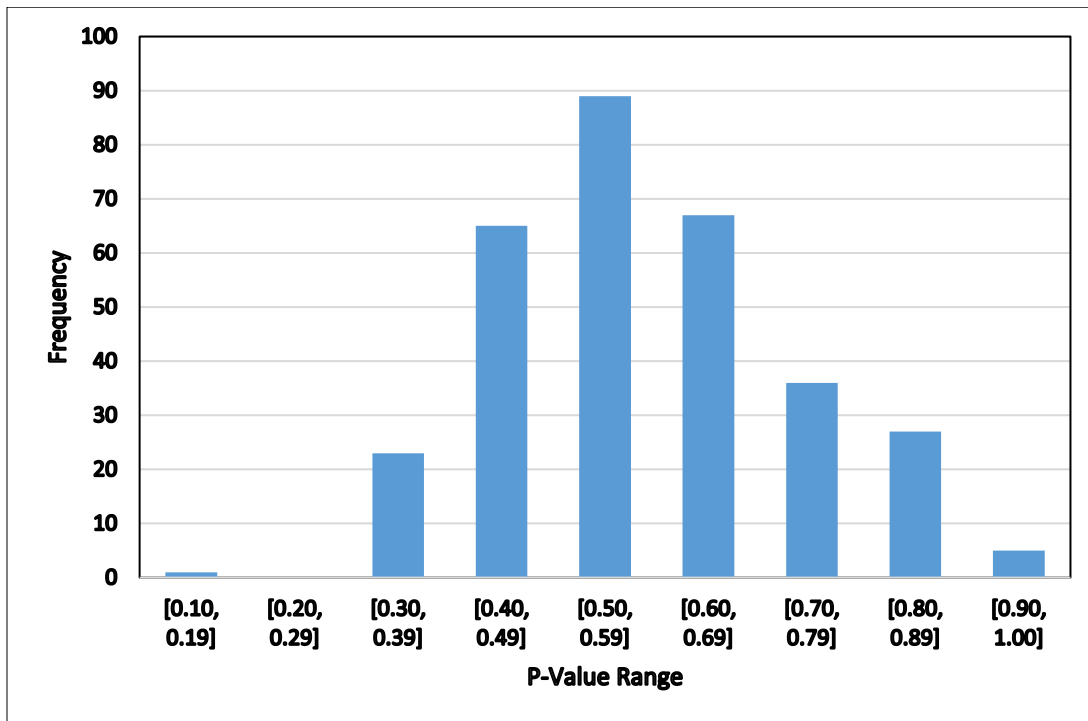


Figure 15. P-value for science operational items.

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

Standardized difference values were also calculated for all operational items with a sample size of at least 20. The standardized difference values were calculated to compare the  $p$ -value for the item to all other items measuring the EE and linkage level. Figure 16 summarizes the standardized difference values for operational items. All items fell within two standard deviations from the mean for the EE and linkage level.

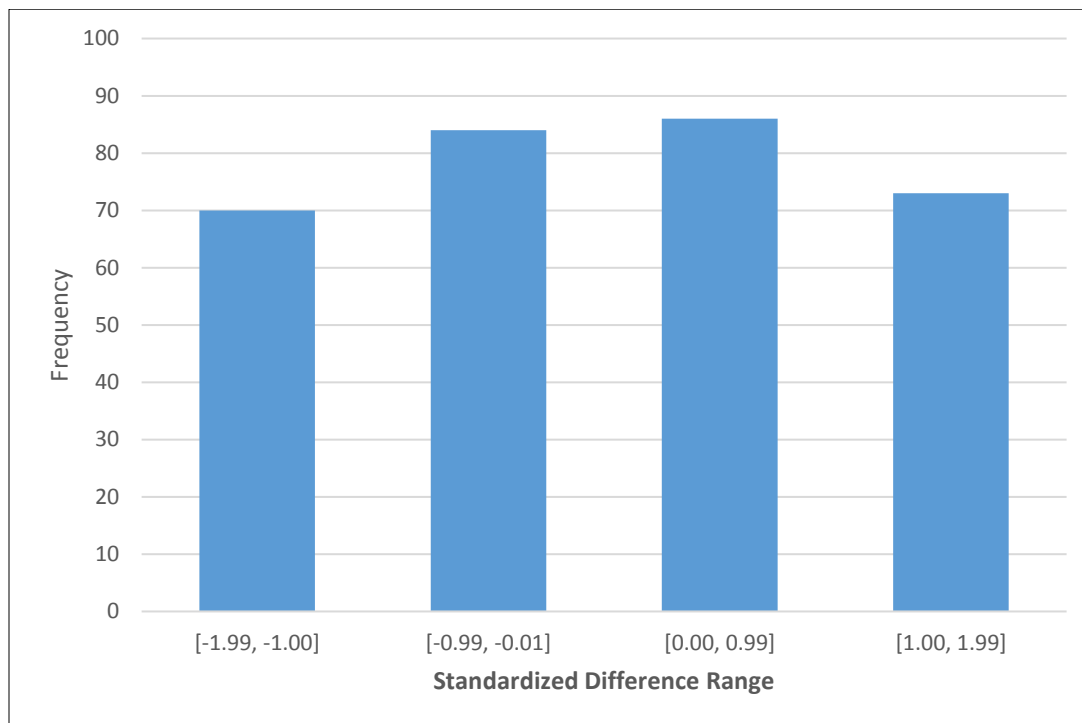


Figure 16. Standardized difference z scores for science operational items.

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

For information on a summary of the total linkage levels mastered during operational testing and the distribution of students by performance level, see Chapter VI.

### III.9. CONCLUSION

The development process for the DLM science assessment was intentionally ambitious to meet the needs of the state partners, and the result is a science assessment that is accessible to students with SCD based on content and standards that are intended to improve teaching and learning science curriculum within this population. Science was able to leverage much of what was already built and learned from the DLM English language arts and Mathematics assessment programs in terms of accessibility features, content development and review processes, and test and item design.

Overall, the Science pilot and field test data provided useful information for nuances specific to assessing science content such as, providing additional context within testlets to reduce cognitive load and to reduce text complexity at the lowest linkage level. Finally, engagement activities for science evolved throughout the development process into more instructionally relevant science stories that guide students through familiar science activities and experiments. These science stories are intended to draw upon students' prior experiences and knowledge and provide context for assessing relevant science skills. These improvements were incorporated into the 2015-2016 operational assessment.

## IV. TEST ADMINISTRATION

Chapter IV presents the processes and procedures used to administer the Dynamic Learning Maps (DLM) science alternate assessments in 2015–2016. As described in earlier chapters, the DLM Consortium developed adaptive computer-delivered alternate assessments that provide the opportunity for students with the most significant cognitive disabilities (SCD) to show what they know and are able to do in science in grades bands 3-5, 6-8, and high school.<sup>3</sup> Assessment blueprints are composed of Essential Elements (EEs), which are alternate content standards that describe what students with SCD should know and be able to do at each grade level. The DLM assessments are administered in small groups of items called testlets. The DLM assessment system incorporates accessibility by design and is guided by the core beliefs that all students should have access to challenging, grade-level content and that educators adhere to the highest levels of integrity in providing instruction and administering assessments based on this challenging content.

First, Chapter IV provides an overview of the key features of test administration. This overview explains how students are assigned their first testlet using the First Contact (FC) survey results. The chapter also describes testlet formats (computer-delivered and teacher-administered) and the assessment window. Sections that follow define test administration protocols, accessibility tools and features, test security, and evidence of educator and student experiences with test administration in 2015–2016.

### IV.1. OVERVIEW OF KEY ADMINISTRATION FEATURES

Consistent with the DLM theory of action described in Chapter I, the DLM test administration features reflect the multidimensional, non-linear, and diverse ways that students learn and demonstrate their learning. Therefore, test administration procedures use multiple sources of information to assign testlets, including student characteristics and prior performance. Based on students' support needs, DLM assessments are designed to be administered in a one-to-one, student/test administrator format. Most test administrators are the special education educators of the students, as they are best equipped to provide the most conducive conditions to elicit valid and reliable results. Test administration processes and procedures also reflect the priorities of fairness and validity through a broad array of accessibility tools and features that are designed to provide access to test content and materials and to limit construct-irrelevant variance.

This section describes the key overarching features of the DLM test administration. The first portion explains the year-end assessment model. The year-end model is currently the only model available to science, which yields summative results based on spring assessments that cover the test blueprints. English language arts (ELA) and mathematics have a second option—an integrated model in which educators collect scores in testing sessions held throughout the year. The next part describes the two assessment delivery modes and the online testing

---

<sup>3</sup> Specific grades required are determined by each state.

platform, the Kansas Interactive Testing Engine (KITE). The final portion details the system-driven adaptive delivery that determines the linkage levels of testlets assigned during the spring assessment window.

#### ***IV.1.A. THE YEAR-END ASSESSMENT MODEL***

While two testing models are used within the DLM ELA and mathematics consortium, the science states chose to follow one of these models: the year-end assessment model. In the year-end assessment model, the DLM system is designed to assess a student's learning consistent with the theory of action (see Chapter I). The year-end model uses testlets that each assess one EE delivered in the spring of each year. In the year-end model, all students are assessed during the spring window on the entire breadth of the blueprints in science.

##### **IV.1.A.i. Assessments**

The DLM alternate assessments are delivered in testlets. In science, testlets are based on either the Target, Precursor, or Initial linkage levels for one EE. Each testlet contains an engagement activity and three to four items. During the testing window, students received either nine or ten testlets, depending on their grade level. The system delivers only one testlet at a time in each subject. The system uses the FC information to initiate the first testlet assigned in science. After the student takes the first testlet, the system delivers the next testlet. The student's performance on the first testlet determines how the system selects and delivers the second testlet. An explanation of the selection procedures that assign the first and subsequent testlets is described in the Adaptive Delivery section in this chapter.

##### **IV.1.A.ii. Calculation of Summative Results**

Summative results are based on student responses on testlets and information about the relationships between the linkage levels. Together, this information is used to determine which linkage levels the student has likely mastered. Results for each linkage level are determined based on the probability that the student has mastered the skills at that linkage level (see Chapter V for a full discussion of modeling).

Linkage level mastery data determine summative results. The information about each linkage level leads to a summary of the student's mastery of skills in each domain and for the subject overall. See Chapter VII for a full description of how summative results are calculated.

#### ***IV.1.B. ASSESSMENT DELIVERY MODES***

The DLM system includes testlets designed to be delivered via computer directly to the student and testlets designed to be delivered via the teacher outside the system, with the teacher recording responses in the system. The majority of testlets were developed for the computer-delivered mode because evidence suggests that the majority of students with SCD are able to interact directly with the computer or are able to access the content of the test on the computer with navigation assistance from a test administrator (Nash, Clark, & Karvonen, 2015). Teacher-administered testlets, designed for educator delivery, included all testlets at the Initial linkage



level and some alternate forms for students who are blind or who have visual impairments. The 2015–2016 operational testlet pool was comprised of 75% computer-delivered testlets and 25% teacher-administered testlets.

#### **IV.1.B.i. Computer-Delivered Assessments**

The DLM science alternate assessment’s Target and Precursor testlets are delivered directly to students by computer through the KITE system. Computer-delivered assessments were designed so students can interact independently with the computer, using special assistive technology devices such as alternate keyboards, touch screens, or switches as necessary.<sup>4</sup>

The computer-delivered testlets use single-select multiple choice items with three response options and text or images as answer choices. See Chapter III for more information about item types.

To illustrate, a released testlet is presented here as an example of the computer-delivered model (Figures 17-24). This high school testlet at the Precursor linkage level assesses students’ knowledge of the EE: “Use data to compare the effectiveness of safety devices to determine which best minimizes the force of a collision” and addresses the science and engineering practice of “Constructing explanations and designing solutions.” The students use the safe drop height data to compare the effectiveness of the devices. The device with the largest safe drop height is the most effective (e.g., can protect the egg from the greatest impact speed). The safe drop height measurement is a proxy for force of impact data. This is instructionally relevant because egg drops are a typical classroom activity for this concept. Students drop their devices from higher and higher heights until they reach the height that the results in the egg breaking, known as the maximum safe drop height. More effective devices allow the egg to stay unbroken from higher drop heights.

---

<sup>4</sup> For students who cannot interact independently with the computer, test administration procedures allow for the student to indicate a response through any mode of expressive communication and for the test administrator to enter the response on the student’s behalf. See the Accessibility section in this chapter for details.

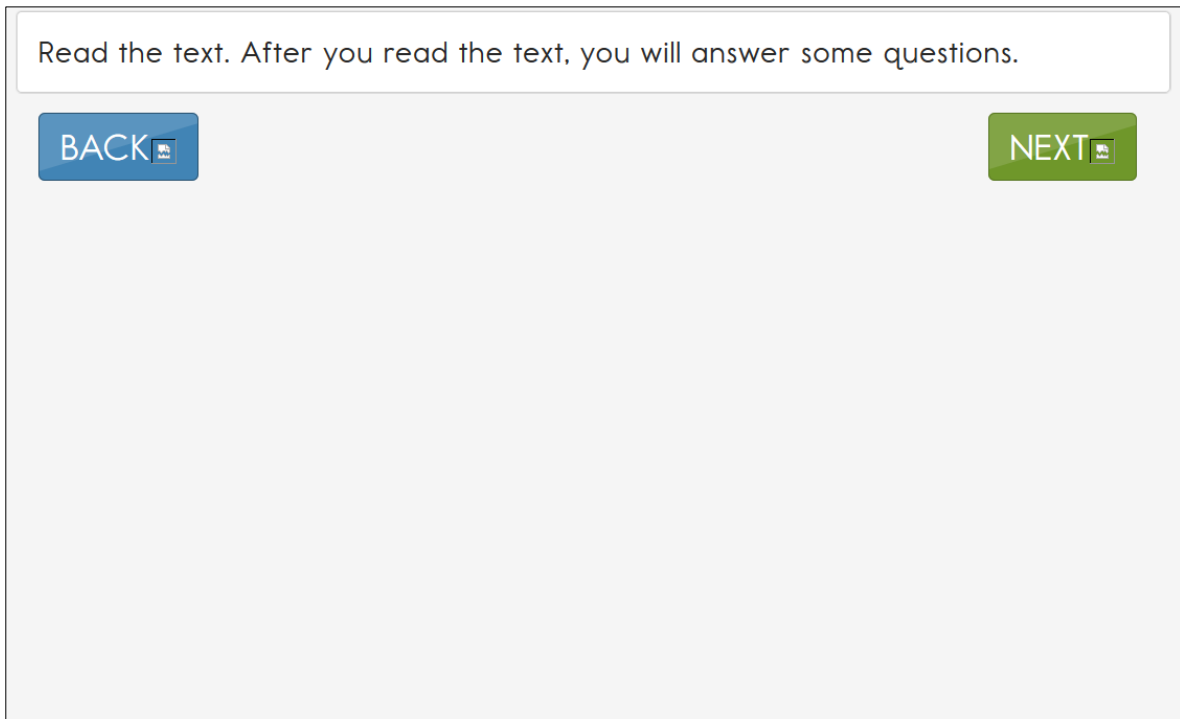


Figure 17. Computer-delivered released testlet – Opening screen with test directions and navigation buttons.


Figure 18 and Figure 19 demonstrate an engagement activity in the form of a science story. It describes a student carrying out an investigation. Alternate text is provided to describe pictures for students with visual impairment.

Tomas learns about safety devices. Tomas knows that safety devices lower forces.



BACK 

EXIT  
DOES NOT SAVE

NEXT 

Tomas compares safety devices. Tomas wants to protect an egg from breaking.



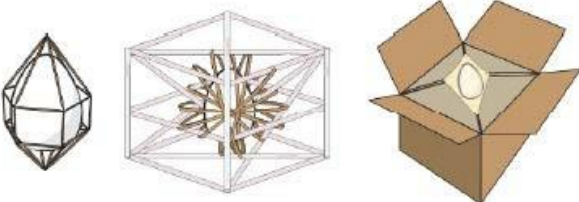
BACK 

EXIT  
DOES NOT SAVE




NEXT 

Figure 18. Computer-delivered released testlet – Science story 1.

Tomas makes 3 egg safety devices.



Device 1      Device 2      Device 3

BACK             NEXT 

Tomas drops the egg safety devices from different heights. Tomas compares the 3 safety devices. Tomas makes a table.

**Safety Devices**

Device	Safe Drop Height
1	10 feet
2	15 feet
3	18 feet




BACK             NEXT 

Figure 19. Computer-delivered released testlet – Science story 1 (continued).

The items illustrated in Figure 20 and Figure 21 use information from the science story to ask questions. The information is repeated so that students do not have to navigate back to the preceding science story screen.


Tomas compares the 3 safety devices. Tomas makes a table.


**Safety Devices**

Device	Safe Drop Height
1	10 feet
2	15 feet
3	18 feet

Which device has the biggest safe drop height?

Device 1  
Device 2  
Device 3

BACK 




NEXT 

Figure 20. Computer-delivered released testlet – Item 1.

Tomas compares the 3 safety devices. Tomas makes a table.

Device	Safe Drop Height
1	10 feet
2	15 feet
3	18 feet

Which device lowers the force on the egg the most?

Device 1  
Device 2  
Device 3








BACK   NEXT 

Figure 21. Computer-delivered released testlet – Item 2.


The science investigation continues with additional story presentation (Figure 22 and Figure 23).

Tomas drops the egg onto 3 different materials. Tomas wants to keep the egg safe.



BACK   NEXT 

Tomas drops the egg onto 3 different materials. Tomas has foam. Tomas has tissues. Tomas has cardboard. Tomas keeps the thickness of the materials the same.



foam      tissue      cardboard




BACK   NEXT 

Figure 22. Computer-delivered released testlet – Science story 2.



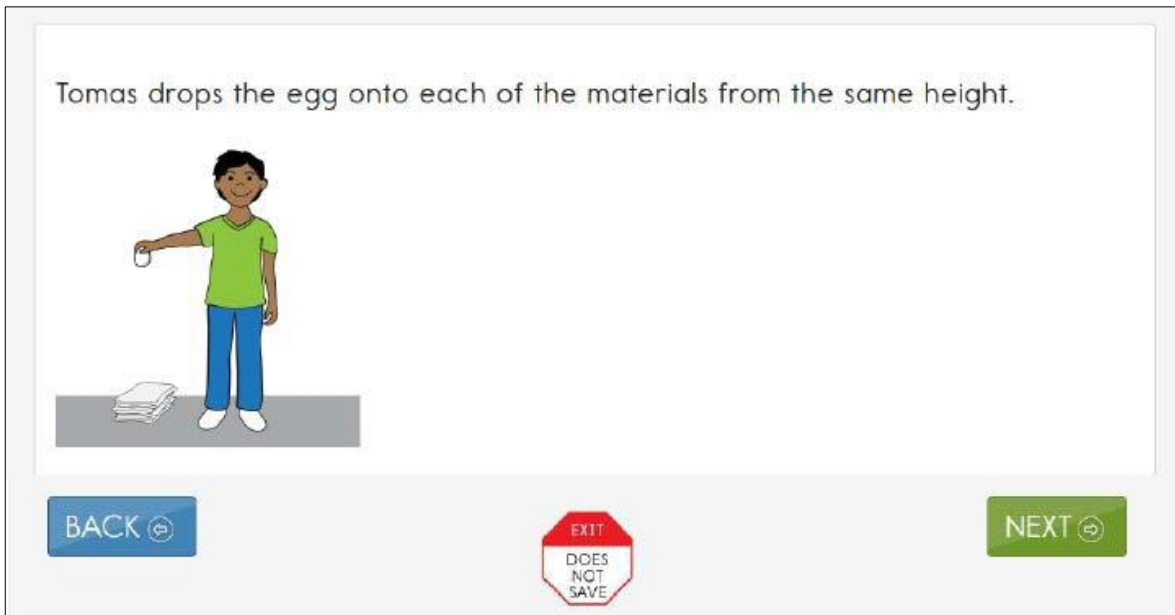


Figure 23. Computer-delivered released testlet – Science story 2 (continued).

Figure 24 displays the final item in the testlet.

Tomas drops the egg onto each of the materials. Tomas checks to see if the egg is safe. Tomas makes a table.


Safety Materials	
Material	Safety
Foam	Safe
Tissue	Not Safe
Cardboard	Not Safe


Which material protects the egg the best?

foam

tissue

cardboard

BACK 




NEXT 

Figure 24. Computer-delivered released testlet – Item 3.

#### IV.1.B.ii. Teacher-Administered Assessments

Some testlets were designed to be administered directly by the test administrator outside of the KITE system. In teacher-administered testlets, the test administrator was responsible for setting up the assessment, delivering it to the student, and recording responses in the DLM system. The KITE system delivered the test, but the test administrator played a more direct role in test administration than in computer-delivered testlets. In science, teacher-administered testlets were designed for students at the Initial level, that is, students who are typically developing symbolic understanding or who may not yet demonstrate symbolic understanding.<sup>5</sup> See Chapter III for a description of the structure of teacher-administered testlets.

To illustrate the format of teacher-administered testlets, figures 25-31 represent a released testlet. This middle school testlet at the Initial linkage level assesses students' knowledge of the EE "Observe and identify examples of change (e.g. state of matter, color, temperature, and

---

<sup>5</sup> These testlets only occurred at lowest linkage level, and the test administrator must be very familiar with the student's typical modes of expressive communication.

odor)” and addresses the science and engineering practice of “Analyzing and interpreting data.”

Figure 25 illustrates directions to the test administrator, including discussion of the Testlet Information Page (TIP). The TIP is available to the educator ahead of time. Generally, the TIP for initial level testlets includes a set of picture response cards which can be printed locally and be used as a stimulus for students to indicate their response. These pictures match the graphics used in the testlet. Teachers may substitute objects or use alternate text for picture response cards as appropriate. The TIP also identifies alternatives to pictures for students who are blind or have visual impairment on a blind/visual impairment directions page. Consistent with the computer-delivered testlets, navigation buttons are standard on each page. Figure 26 shows directions to the test administrator for administrating the first item using a set of picture response cards.

The image shows two screenshots of a testlet interface. The top screenshot displays the following text:

Educator Directions:

In this testlet, the student will identify examples of changes in the state of matter.

Before you begin working with the student, print the following pictures included in the Testlet Information Page:

- a solid stick of butter
- a melting stick of butter
- a cup of melting ice
- a cup of steaming water
- a pot of water
- a pot of boiling water

At the bottom of the first screenshot are three buttons: a blue 'BACK' button with a left arrow, a red octagonal 'EXIT DOES NOT SAVE' button, and a green 'NEXT' button with a right arrow.

The second screenshot displays the following text:

Educator Directions:

Read the text with the student. Maximize your interaction with the student. Lead with comments, and direct the student's attention to the text, images, or objects. Make sounds and perform actions when appropriate. After you read the text, the student will complete some tasks.

At the bottom of the second screenshot are three buttons: a blue 'BACK' button with a left arrow, a red octagonal 'EXIT DOES NOT SAVE' button, and a green 'NEXT' button with a right arrow.

Figure 25. Teacher-administered released testlet – General educator directions.

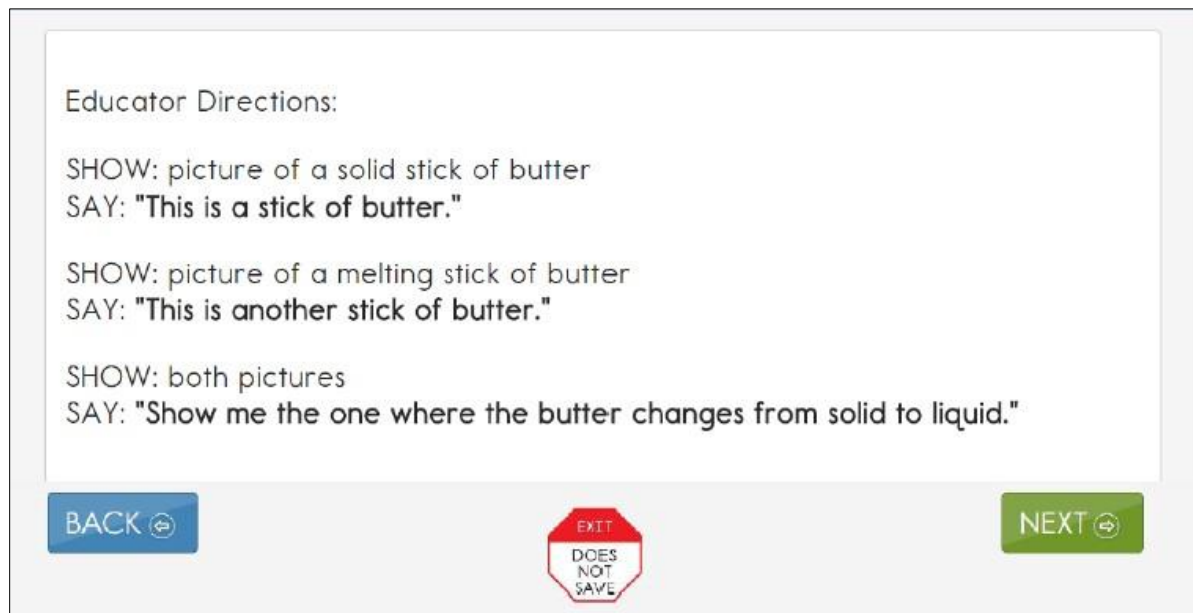




Figure 26. Teacher-administered released testlet – Educator directions for Item 1.

Figure 27 depicts where the educator records student responses to the item. The graphics of butter representing un-melted and melted states are the same as are found on the picture response cards on the TIP. The student indicates his or her response using the picture response cards. Using the options on the screen, the educator identifies the response that best matches the student's behavior. This format, with five options for responses, is typical. The last two response options are consistent across items. Figure 28, Figure 29, Figure 30, and Figure 31 illustrate the directions for administering the remaining two items in the testlet which follow the same format and structure as the first item.

Record student response:

- Indicates melting stick of butter:  

- Indicates solid stick of butter:  

- Indicates both pictures
- Indicates or attends to other stimuli
- No response




[BACK](#)   [NEXT](#) 


Figure 27. Teacher-administered released testlet – Student response record for Item 1.


Educator Directions:

SHOW: picture of a cup of melting ice cubes  
SAY: "This is a cup of water."

SHOW: picture of a cup of steaming water  
SAY: "This is another cup of water."

SHOW: both pictures  
SAY: "Show me the one where water changes from solid to liquid."

BACK 





NEXT 

Figure 28. Teacher-administered released testlet – Educator directions for Item 2.




Record student response:

Indicates cup of melting ice cubes:




Indicates cup of steaming water:



Indicates both pictures

Indicates or attends to other stimuli

No response

BACK 

EXIT  
DOES  
NOT  
SAVE


NEXT 


Figure 29. Teacher-administered released testlet – Student response record for Item 2.


Educator Directions:

SHOW: picture of pot of water  
SAY: "This is a pot of water."

SHOW: picture of pot of boiling water  
SAY: "This is another pot of water."

SHOW: both pictures  
SAY: "Show me the one where water changes from liquid to gas."

BACK 





NEXT 


Figure 30. Teacher-administered released testlet – Educator directions for Item 3.

Record student response:

Indicates pot of boiling water:




Indicates pot of water that is not boiling:



Indicates both pictures

Indicates or attends to other stimuli

No response

BACK 

EXIT  
DOES NOT SAVE


NEXT 

Figure 31. Teacher-administered released testlet – Student response record for Item 3.

#### ***IV.1.C. THE KITE SYSTEM***

The DLM alternate assessments are managed and delivered using the KITE platform, which was designed and developed to meet the needs of the next generation of large-scale assessments. The KITE system consists of four applications. Educators and students see two of these applications: Educator Portal and KITE Client (Test Delivery Engine). The KITE system has been developed with IMS Global Question and Test Interoperability item structures and Accessible Portable Item Protocol tagging on assessment content to support students' Personal Needs and Preferences (PNP) Profile and World Wide Web Consortium Web Content Accessibility Guidelines in KITE Client. Minimum hardware and operating system requirements for KITE Client and supported browsers for Educator Portal are published on the DLM website and in the DLM *Technical Liaison Manual* linked on each state's DLM webpage.

### IV.1.C.i. Educator Portal

Educator Portal is the administrative application where staff and educators manage student data, complete required test administration training, retrieve resources needed for each assigned testlet, and retrieve reports.

- Test administrators, usually educators, use Educator Portal to manage all student data. They are responsible for checking class rosters of the students who are assigned to take DLM alternate assessment testlets and for completing the PNP and FC for each student.
- Educator Portal hosts the required test administrator training modules. Test administrators complete facilitated or self-directed training and take post-tests to demonstrate their understanding of the material (see Chapter X for more information).
- After each testlet is assigned to a student, the system delivers a TIP through Educator Portal. The TIP, which is unique to the assigned testlet, is a PDF that contains any instructions necessary to prepare for testlet administration (see the Resources and Materials section of this chapter for more information).

### IV.1.C.ii. KITE Client (Test Delivery Engine)

The KITE Test Delivery Engine (TDE) is the portal that allows students to log in and complete assigned testlets. Practice activities and released testlets are also available to students through TDE. Students access TDE via KITE Client, a customized version of Firefox, which launches in kiosk mode and prevents students from accessing unauthorized content or software while taking secure, high-stakes assessments. The TDE interface is supported on desktops and laptops running Windows or OS X, on Chromebooks, and on iPad tablets.

The KITE system provides students with a simple, web-based interface with student-friendly and intuitive graphics. The student interface used to administer the DLM assessments was designed specifically for students with SCD. It maximizes space available to display content, decreases space devoted to tool-activation buttons within a testing session, and minimizes the cognitive load related to test navigation and response entry. An example of a screen used in a science testlet is shown in Figure 32. The blue **BACK** button and green **NEXT** button are used to navigate between screens. The octagonal **EXIT DOES NOT SAVE** button allows the user to exit the testlet without recording any responses. The **READ** button plays an audio file of synthetic speech for the content on screen. Synthetic read aloud is the only accessibility feature with a tool directly enabled through each screen in the testlet. Further information about accessibility features is provided in the Accessibility section in this chapter.

Tomas drops the egg safety devices from different heights. Tomas compares the 3 safety devices. Tomas makes a table.

**Safety Devices**

Device	Safe Drop Height
1	10 feet
2	15 feet
3	18 feet





BACK    READ 

Figure 32. An example screen from the student interface in KITE Client.

#### IV.1.C.iii. Local Caching Server

During DLM assessment administration, schools with unreliable network connections have the option to use the Local Caching Server (LCS). The LCS is a specially configured machine that resides on the local network and communicates between the testing machines at the testing location and the main testing servers for the DLM system. The LCS stores testing data from KITE Client in an internal database; therefore, if the upstream network connection becomes unreliable or variable during testing, students can still continue testing, and their responses will be transmitted to the KITE servers as bandwidth allows. The LCS submits and receives data to and from the DLM servers while the students are taking tests. The LCS must be connected to the Internet between testlets in order to ensure the next testlet is delivered correctly.

#### IV.1.D. ADAPTIVE DELIVERY

The DLM assessments are delivered in testlets. While blueprints determine the EEs that are selected for assessment, the adaptive delivery mechanism determines the linkage level for each testlet assigned to students. The linkage level of the first assigned testlet in science is determined based on educator responses to the FC, which has an inventory of learner characteristics in a variety of areas including communication and academic skills. For the spring 2016 science administration, one section of the FC was used to provide a match between student and testlet during the initial DLM testing experience—the Expressive Communication section. The process of assigning a student to a linkage level for the first testlet is known as initialization. FC items on expressive communication used for initialization purposes are included in Appendix D and are consistent with the ELA and mathematics assessment

(Dynamic Learning Maps, 2016). Based on the educator's responses, the student's assigned complexity band was automatically calculated and stored in the system. Subsequent testlets could be at higher or lower linkage levels, based on student performance on the prior testlet. Consistent with ELA and mathematics, no testlets written at the Successor level are delivered as the first testlet. However, a student is able to route to the Successor level by providing correct responses to items on a Target level testlet.

The correspondence among students' expressive communication skills, indicated on the FC, the corresponding FC complexity bands, and the recommended linkage levels are shown in Table 35.

Table 35. Correspondence Between Complexity Band and Assigned Linkage Level

<b>Common First Contact Survey Responses</b>	<b>First Contact Complexity Band</b>	<b>Linkage Level</b>
Does not use speech, sign, or AAC	Foundational	Initial
Uses one word, sign, or symbol to communicate;	Band 1	Initial
Uses 2 words, signs, or symbols to communicate	Band 2	Precursor
Regularly combines three or more spoken words to communicate for a variety of purposes	Band 3	Target

*Note:* AAC = augmentative or alternative communication device

The educator must complete the student's FC before assessments are delivered. Supporting procedures and a complete list of FC questions are included in the *Test Administration Manual 2015-2016* (Dynamic Learning Maps, 2016). Test administrators are trained on the role of the FC in testlet assignment as part of required test administrator training (see Chapter X).

Each student was assigned testlets per subject during the spring window. The system determined the linkage level for each testlet. The assignment was adaptive between testlets. Each spring testlet was packaged and delivered separately, and the test administrator determined when to schedule each testlet within the larger window. See *Spring Assessments* (Dynamic Learning Maps, 2016, p. 83) for more detail.

The second and subsequent testlets were assigned based on previous performance. That is, the linkage level associated with the next testlet a student received was based on the student's performance on the previously administered testlet. The goal was to maximize the match of student knowledge, skills, and understandings to the appropriate linkage level content with the following decision rules:

- The system adapted up one linkage level if students responded correctly to 80% or more of the items measuring the previously tested EE. If testlets are already at the highest level (i.e., Target), they remain there.
- The system adapted down one linkage level if students responded correctly to less than 35% of the items measuring the previously tested EE. If testlets are already at the lowest level (i.e., Initial), they remain there.
- Testlets remain at the same linkage level if students responded correctly to between 35% and 80% of the items measuring the previously tested EE.

Threshold values for routing were selected with the number of items included in a testlet (three to four items) in mind. In a testlet that contained three items measuring the EE, if a student responded incorrectly to all items or correctly answered only one item (proportion correct  $< 0.35$ ), the linkage level of the testlet was likely too challenging. To provide a better match to the student's knowledge, skills, and ability, the student was routed to a lower linkage level. A single correct answer could be attributed to either a correct guess or true knowledge that did not translate to the other items measuring the EE. However, if the student responded to two of the three items correctly or three of four items correctly (proportion correct, between 0.35 and 0.80), it could not be assumed the student had completely mastered the knowledge, skills, or understandings being assessed at that linkage level. Therefore, the student was neither routed up nor down for the subsequent testlet. If the student responded correctly to all of the items, then it was assumed the student had completely mastered the skill being assessed at that linkage level. The student was routed to a higher linkage level to allow the student the opportunity to demonstrate more advanced knowledge or skills.

Figure 33 provides an example of testlet adaptations for a student who completed five testlets. In the example, on the first assigned testlet at the Initial level, the student answered all of the items correctly, so the next testlet was assigned at the Precursor level. The next two testlets adapted up and down a level, respectively, whereas the fifth testlet remained at the same linkage level as the previous testlet.



EE 1	EE 2	EE 3	EE 4	EE 5
I	I	I	I	I
P	<u>P</u>	P	<u>P</u>	<u>P</u>
T	T	<u>T</u>	T	T

Figure 33. Linkage level adaptations for a student who completed five testlets.

Note: I = Initial; P = Precursor; T = Target.

#### IV.1.E. SPECIAL CIRCUMSTANCE CODES

In 2015-2016, state partners were given the option to allow entry of special circumstance codes in Educator Portal. For states implementing the use of special circumstance codes, state partners defined the list of allowable codes, including correspondence of the Common Education Data Standards codes to state-specific codes and definitions.

Special circumstance codes were made available for entry in the event that a student could not participate in a testlet, and could be entered in Educator Portal to provide explanation for the reason the student was not tested. These codes could later be used by the state when applying rules about counting student participation and performance in federal and state accountability systems.

The special circumstance fields were located in Educator Portal on the same screen where the TIP was accessed, and included descriptive words, e.g. medical waiver or parental refusal. Only educators with the role of District Test Coordinator, Building Test Coordinator, and State Assessment Administrator had the permissions to choose the code. DLM staff recommended that the special circumstance code not be entered until late in the state's testing window, to allow adequate time for testing to occur, but before the window closed. Codes needed to be entered once per content area associated with the first testlet to be delivered, or on an as needed basis when later testlets could not be administered due to a special circumstance. Data files delivered to state partners summarizing special circumstance codes are described in Chapter VII.

#### IV.2. TEST ADMINISTRATION

This section gives an overview of general test administration processes and procedures. For more detail, see the *Test Administration Manual 2015-2016* (Dynamic Learning Maps, 2016) and the *Science Supplement to Test Administration Manual* (Dynamic Learning Maps, 2016). Test

administration guidelines provide educators with the information necessary to administer the assessments with fidelity and for students to demonstrate their knowledge and skills at appropriate breadth, depth, and complexity of the content.

#### **IV.2.A. TEST WINDOWS**

During the consortium-wide spring testing window, which occurred between March 16 and June 10, 2016, all students were assessed on the EEs on the blueprint in science. Each state set its own testing window within the larger consortium window.

#### **IV.2.B. ADMINISTRATION TIME**

During the spring testing window, estimated total testing time was between 45-135 minutes per student (60-180 minutes for end-of-course biology), with each testlet taking approximately 5-15 minutes. Actual testing time per testlet varied depending on each student's unique characteristics.

The KITE system captured start and end dates and time stamps for every testlet. Actual testing time per testlet was calculated as the difference between start and end times. As the KITE system was still in development, the spring 2016 operational time-stamp data included some impossible values (i.e., negative times, values greater than 24 hours). Implausible values comprised 5% of the data.

Table 36 and Table 37 show the distribution of test times per testlet after removing negative values and test times greater than 8 hours (i.e., approximate maximum length of a school day) for Initial level testlets and Precursor and Target level testlets, respectively. Given the wide range of testlet response times (up to 8 hours), the interquartile range values most likely describe the typical range of testing time per testlet. Most Initial level testlets took approximately 4-5 minutes to complete, while most Precursor and Target level testlets took approximately 2-3 minutes to complete.

Table 36. Distribution of Response Times in Minutes for Initial Level Testlets

<b>Grade</b>	<b>Min</b>	<b>Max</b>	<b>Mean</b>	<b>Median</b>	<b>25%Q</b>	<b>75%Q</b>	<b>IQR</b>
3-5	0.15	464.18	4.65	2.48	1.45	4.30	2.85
6-8	0.17	412.50	4.30	2.38	1.37	4.23	2.87
9-12	0.13	431.87	4.73	2.38	1.22	4.33	3.12
BIO	0.20	278.70	4.82	2.75	1.53	4.99	3.45

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

Table 37. Distribution of Response Times in Minutes for Precursor and Target Level Testlets

Grade	Min	Max	Mean	Median	25%Q	75%Q	IQR
3-5	0.12	328.70	2.94	2.13	1.50	3.12	1.62
6-8	0.15	387.52	2.60	1.80	1.23	2.75	1.52
9-12	0.12	427.15	3.00	2.12	1.50	3.05	1.55
BIO	0.18	365.77	3.45	2.50	1.68	3.72	2.03

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

In the *DLM Test Administration Manual 2015-16*, test administrators were encouraged to allow students to take breaks in the case of fatigue, disengagement, or behavioral problems that were likely to interfere with a valid assessment of what the student knows and can do. The KITE system allowed for up to 90 minutes of inactivity without timing out to allow test administrators and students to pause for breaks during administration of a testlet. In cases where administration had begun but the student was unable to engage and respond for any reason and a short break was not sufficient, the **EXIT DOES NOT SAVE** button could be used to exit the testlet, allowing the test administrator and student to return to it at another time.

### IV.2.C. RESOURCES AND MATERIALS

Test administrators, school staff, and Individualized Education Program (IEP) teams had multiple resources throughout the test administration process, some of which were provided consortium wide. Some states provided additional, state-specific materials on their websites. The DLM website provided resources that covered DLM background and assessment administration training information; student and roster data management; test delivery protocols and setup; accessibility features, protocols, and documentation; and practice activities. This section provides an overview of all resources and materials for test administrators as well as more detail regarding the critical resources of TIPs and Practice Activities and Released Testlets.

#### IV.2.C.i. The DLM Website

The DLM website served as a way to communicate assessment information to educators. Pages such as EEs, Accessibility, and Test Development were included and cover topics related to the DLM system as a whole and those that may be of interest to a variety of audiences. To support assessment administration, each state also had its own customized landing page with an easy-to-remember URL (i.e., [dynamiclearningmaps.org/statename](http://dynamiclearningmaps.org/statename)). Through training, manuals, webinars, and replies from Service Desk inquiries, educators were made aware of their state-specific webpage to locate consortium-level resources and state-customized resources.

To provide consortium-wide updates and reminders, the DLM website also featured a Test Updates webpage. This was a newsfeed-style page that addressed timely topics such as

assessment deadlines, resource updates, or system status. Additionally, the Test Updates page offered educators the option to subscribe to an electronic mailing list to automatically receive the same message via email without visiting the website.

#### IV.2.C.ii. Test Administration Resources

The DLM website provided specific resources designed for test administrators. These resources (Table 38) were available to all states to ensure consistent test administration practices.

Table 38. DLM Resources for Test Administrators and States

Resource Title	Description
Test Administration Manual (PDF)	Supports for the test administrator in preparing themselves and students for testing.
About Testlet Information Pages (TIPs)	Provides guidance for test administrators on the types and uses of information in the TIPs provided for each testlet.
Accessibility Manual (PDF)	Provides guidance to state leaders, districts, educators, and IEP teams on the selection and use of accessibility supports available in the DLM system.
Educator Resource Page (Webpage)	Includes additional resources for test administrators, such as familiar texts, materials collection, testlet overview videos, and tested EEs and their associated mini-maps.
Guide to DLM Required Training and Professional Development 2015-2016 (PDF)	Helps users access DLM Required Test Administration Training and instructional professional development in Educator Portal.
Guide to Practice Activities & Released Testlets	Supports the test administrator in accessing practice activities in KITE Client.
Test Updates Page (Webpage)	Breaking news on test administration activities. Users can sign up to receive alerts when new resources become available.

Resource Title	Description
Training Video Transcripts (PDF)	Links to transcripts (narrator notes) for the Required Test Administration Training modules.

#### IV.2.C.iii. District-Level Staff Resources

Resources were available for three district-level supporting roles: Assessment Coordinator, Data Steward, and Technical Liaison. The Assessment Coordinator oversaw the assessment process, which included managing staff roles and responsibilities, developing and implementing a comprehensive training plan, developing a schedule for test implementation, monitoring and supporting test preparations and administration, and developing a plan to facilitate communication with parents or guardians and staff. The Data Steward managed educator, student, and roster data. The Technical Liaison verified that the network and testing devices were prepared for test administration.

Resources for each of these roles were made available on the state's customized DLM webpage. Each role had its own manual, a webinar, and a FAQ compiled from webinar questions. Each role was also guided to the supporting resources for other roles anywhere the responsibilities overlapped. For example, Data Stewards were also guided to the *Test Administration Manual 2015-2016* to support data-related activities assigned to the test administrator and connected to troubleshooting data issues experienced by the test administrator. Technical Liaisons were also guided to the KITE Client and Educator Portal webpage for information and documents connected to KITE Client, Local Caching Server use, supported browsers, and bandwidth requirements. Assessment Coordinators were also guided to resources developed for the Data Steward, Technical Liaison, and test administrator for specific information and supplemental knowledge of the responsibilities of each of those roles. Some of those resources include:

- *Guide To DLM Required Training & Professional Development 2015-2016*
- *Test Administration Manual 2015-2016*
- *Science Supplement to Test Administration Manual 2015-2016*
- Field Test webpage
- Test Updates webpage and electronic mailing list

Descriptions of the district-level role webinars are provided in Chapter X.

#### IV.2.C.iv. Testlet Information Pages (TIPs)

TIPs provided test administrators with information specific to each testlet. Test administrators received a TIP page after each testlet was assigned to a student, and they were instructed to review the TIP before beginning the student's assessment (see the sample TIP in Appendix D.)

Each TIP stated whether a testlet was computer-delivered or teacher-administered and indicated the number of items on the testlet. The TIP also provided information for each testlet regarding the materials needed or any substitute materials allowed.

The TIP provided information on the exceptions to allowable supports. While a test administrator typically used all appropriate PNP features and other flexibility tools described in the Allowable Practices section of the *Test Administration Manual 2015-16*, the TIP indicated when it was not appropriate to use a support on a specific testlet. This may have included limits on the use of definitions, translation, read aloud, or other supports.

If there were further unique instructions for a given testlet, they were provided in the TIP. For test administrators who delivered human read-aloud, this included descriptions of graphics, and alternate text descriptions of images were provided as additional pages after the main TIP.

TIPS for science testlets also provided picture response cards for teacher-administered testlets. Most teacher-administered testlets required the use of picture response cards in testlet items.

#### **IV.2.C.v. Practice Activities and Released Testlets**

Two practice activities and many released testlets were made available to support educators and students preparing for testing.

- The practice activities were designed to familiarize users with the way testlets and item features look in the KITE system. One activity was for educators and the other was for students.
- The released testlets were similar to real DLM testlets in content and format.

Practice activities and released testlets were accessed through KITE Client in the practice section. Using login information provided by the system, both types of activities could be completed as many times as desired.

The educator practice activity was a tutorial about testlets administered by the educator. In this tutorial, educators were instructed on how to read and follow the instructions on the screens and how to enter the student's responses to activities or exchanges that occurred outside the system. Most of these testlets required educators to gather materials to be used in the assessment. Directions for how to prepare for the testlet were provided as Educator Directions on the first screen.

The student practice activity was a tutorial about the testlets administered directly to the student. The student practice activity provided an opportunity for students to become familiar with navigation in the KITE system, the types of items used in DLM assessments, and the method for indicating responses to different item types.

Released testlets were similar to operational testlets. They were selected from a variety of EEs and linkage levels across grade bands. New released testlets are added periodically and include teacher-administered testlets and computer-delivered testlets.



#### ***IV.2.D. TEST ADMINISTRATOR RESPONSIBILITIES AND PROCEDURES***

Procedures for test administrators were organized into three sets of tasks for different parts of the school year: (1) before beginning assessments, (2) spring window assessment, and (3) preparing for next year. The *Test Administration Manual 2015-2016* (Dynamic Learning Maps, 2016) provided detailed description of each set of tasks with specific resources to support the work.

##### **IV.2.D.i. Before Beginning Assessments**

Test administrators performed multiple steps to prepare for student testing. They confirmed student eligibility to participate in the DLM alternate assessments and shared information about the assessments with parents to prepare them for their child's testing experience. Test administrators reviewed the entire *Test Administration Manual 2015-2016* and became familiar with all available resources, including state webpages, practice testlets, available content to be assessed, and procedures for preparing to give the assessment.

They also prepared for the computer-delivered aspects of the assessment system. Test administrators had to gain access to Educator Portal, activate their KITE accounts, complete the security agreements in their Educator Portal profile, and complete their required test administration training (see Chapter X). Test administrators also reviewed their state's guidance on required and recommended professional development modules.

Preparation also involved reviewing the *Accessibility Manual* (Wells-Moreaux, Bechard, & Karvonen, 2015) and working with the IEP team to determine what accessibility supports should be provided for each student taking the DLM assessments. Test administrators recorded the chosen supports in the PNP in Educator Portal and reviewed their state's requirement for documentation of the DLM accessibility supports as testing accommodations, adjusting the testing accommodations in the IEP as necessary.

Additional preparations included a review of student demographic information and roster data in Educator Portal for accuracy. Test administrators ensured that the PNP and the FC were updated and complete in Educator Portal. School staff installed KITE Client on testing devices and familiarized both educators and students with DLM testlets through practice activities and released testlets. Finally, student devices were checked for compatibility with KITE Client.

##### **IV.2.D.ii. During Spring Window Assessment**

The spring assessment procedures included checking student demographic information, PNP settings, and FC responses. School staff members considered the district and school assessment schedules to ensure students could complete all DLM testlets during the spring window, and then they scheduled assessment session locations and times.

Test administrators retrieved TIPs for the assigned first testlet and gathered the materials needed before beginning the assessment. After retrieving student usernames and passwords from Educator Portal, test administrators assessed each student with the first testlet. As each



remaining testlet became available, they retrieved TIPs, gathered materials as needed, and assessed the student.

#### **IV.2.D.iii. Preparing for Next Year**

With IEP teams, educators evaluated students' accessibility supports (PNP settings) and made decisions about supports and tools for the next year. With IEP teams, they reviewed the blueprint for the next grade as one source of information to plan academic IEP goals.

#### **IV.2.E. MONITORING ASSESSMENT ADMINISTRATION**

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

##### **IV.2.E.i. Consortium Test Administration Observation Protocol**

The DLM Consortium developed a test administration observation protocol (see Appendix D) in an effort to standardize test administration data collection across observers and locations. The majority of items in the protocol are based on direct recording of what is observed and require little inference or background knowledge. Information collected from this protocol is annually used to evaluate several assumptions in the validity argument.

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

The protocol was available as a PDF to be printed for handwritten observations and as an online survey (optimized for mobile devices and with branching logic) to support electronic data collection.

Training resources were provided to SEA staff to support fidelity of use of the test administration protocol and to increase the reliability of data collected (Table 39). SEA staff had access to the *Test Administration Observation Training* video (see Appendix D) on the use of the *Test Administration Observation Protocol*. The links to this video, the *Guidance for Local Observers* (see Appendix D), and the *Test Administrator Observation Protocol* are provided on the state side of the DLM website. State education agencies are encouraged to use this information in their state monitoring efforts. State education agencies were able to use these training resources to encourage use of the protocol among local education agency staff. States were also cautioned that the protocol was only to be used to document observations for the purpose of describing the administration process. It was not to be used for evaluating or coaching educators or gauging student academic performance. This caution, as well as general instructions for completing and submitting the protocol, are provided in the form itself.

Table 39. DLM Resources for Test Administration Monitoring Efforts

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

#### IV.2.E.ii. Formative Monitoring Techniques

The consortium used several techniques to formatively monitor the status of test administrations within and across states in 2015–2016. First, because DLM assessments are delivered as a series of testlets, a test administration monitoring extract was available on demand in Educator Portal. This extract allowed state and local staff to check each student's progress toward completion of all required testlets. For each student, the extract listed the number of testlets completed and expected for each subject. To support local capacity for monitoring, webinars were delivered in February and March 2016 before the spring testing window opened. These webinars targeted district and school personnel who monitored assessments and had not yet been involved in DLM assessments (see Appendix D).

Formative monitoring also occurred through regular consortium calls including DLM staff and state partners. Throughout most of the year, these calls were scheduled twice per month. Topics related to monitoring that regularly appeared on agendas for partner calls included assessment window preparation, anticipated high-frequency questions from the field, and opportunities for SEA–driven discussion. Particular attention was paid to questions from the field concerning

sources of confusion among test administrators that could compromise assessment results. During the spring window, check-in calls were hosted on the weeks between the regularly scheduled partner calls. The purpose of the check-in calls was to keep state partners apprised of any issues or concerns that arose during the testing window, allowing them to provide timely information to districts. States were provided with a description of the issue as well as actions that were in place to remedy the situation. During these meetings, partner states were encouraged to share any concerns that had arisen during the week from the field and to provide feedback on implemented fixes.

#### **IV.2.E.iii. Monitoring Testlet Delivery**

Prior to the opening of a testing window, Agile Technology Solutions (ATS) staff, the organization that develops and maintains the KITE system and provides DLM Service Desk support to educators in the field, initiated an automated enrollment process that assigned the first testlet. Students who had missing or incorrect information in Educator Portal, preventing testlet assignment, were included in error logs that detailed which information was missing (e.g., FC not submitted) or incorrect (e.g., student is enrolled in a grade that is not tested). These error logs were accessed by ATS staff. Once the student completed the first testlet, the adaptive delivery component of the KITE system drove the remaining testlet assignments. This process also generated error logs that could be accessed by ATS staff. When testlets could not be assigned for large numbers of students in a state due to missing or incorrect data, or when the adaptive delivery system did not work as intended, DLM staff worked with state partners to either communicate general reminders to the field or solve problems regarding specific students.

During each operational window, the DLM psychometric team monitored test delivery to ensure students received testlets according to auto-enrollment specifications. This included running basic frequency statistics to verify that counts appeared as expected by grade, state, and testing model and verifying correct assignment to initial testlet-based rules that govern that process. In addition, a script was run to verify that student routing through the system occurred as expected, whereby students routed to the correct linkage level for each subsequent testlet based on the algorithm described earlier in this chapter in the section called Test Administration.

### **IV.3. ACCESSIBILITY SUPPORTS**

The DLM system was designed to be optimally accessible to diverse learners through accessible content (see Chapter III), initialization, and routing driven by FC and prior performance (Chapters III and IV). Accessibility is also supported by a straightforward user interface in the KITE system (see Overview of Key Administration Features section, above). Consistent with the item and test development practices described in Chapters II and III, principles of Universal Design for Learning (UDL) were applied to design the test administration procedures and platforms. Decisions were largely guided by UDL principles of flexibility of use and equitability

of use through multiple means of engagement, multiple means of representation, and multiple means of action and expression.

In addition to these considerations, a variety of accessibility supports were made available for use in the DLM assessment system. The *Accessibility Manual* for 2015-2016 (Wells-Moreaux, Bechard, & Karvonen, 2015) outlined a six-step process for test administrators and IEP teams to use in making decisions about accessibility supports. This process began with confirming that the student meets the DLM participation guidelines (see Appendix D) and continued with the selection, administration, and evaluation of the effectiveness of the accessibility supports. Supports were selected for each student in the PNP in Educator Portal. The PNP could be completed any time before testing began. It could also be changed during testing as a student's needs changed. Once updated, the changes appeared the next time the student was logged in to the KITE system. All test administrators were trained in the use and management of these features (see Chapter X).

#### **IV.3.A. OVERVIEW OF ACCESSIBILITY SUPPORTS**

Appropriate accessibility supports to use during administration of computer-delivered or teacher-delivered testlets were listed in the *Accessibility Manual* (Wells-Moreaux, Bechard, & Karvonen, 2015). A brief description of the supports is provided here (see the *Accessibility Manual* for a full description of each support and its appropriate use). Supports were grouped into three categories: those accessed through the PNP, those requiring additional tools or materials, and those provided outside the system. Supports are listed in each of these categories in Table 40.

Table 40. Accessibility Supports in the DLM Assessment System

<b>Supports Provided via PNP</b>	<b>Supports Requiring Additional Tools/Materials</b>	<b>Supports Provided Outside the System</b>
<ul style="list-style-type: none"> <li>• Magnification</li> <li>• Invert color choice</li> <li>• Color contrast</li> <li>• Overlay color</li> <li>• Spoken audio               <ul style="list-style-type: none"> <li>• Text only</li> <li>• Text &amp; graphics</li> <li>• Nonvisual</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>• Uncontracted braille</li> <li>• Single-switch system/PNP enabled</li> <li>• Two-switch system</li> <li>• Individualized manipulatives</li> <li>• Alternate form – visual impairment</li> </ul>	<ul style="list-style-type: none"> <li>• Human read-aloud</li> <li>• Sign interpretation of text</li> <li>• Language translation of text</li> <li>• Test administrator enter responses for student</li> <li>• Partner-assisted scanning (PAS)</li> </ul>

*Note:* These supports are described for the DLM system as of spring 2016. PNP = Personal Needs and Preferences Profile.

Additional techniques that are traditionally thought of as accommodations are considered allowable practices in the DLM assessment system. These are described in a separate section below.

#### **IV.3.A.i. Category 1: Supports provided within the DLM system via the PNP**

Online supports include magnification, invert color choice, color contrast, and overlay color. Educators can test these options in advance to make sure they are compatible and provide the best access for students. Test administrators can adjust the PNP-driven accessibility during the assessment, and the selected options are then available the next time the student logs in to KITE Client.

- *Magnification* – Magnification allows educators to choose the amount of screen magnification during testing.
- *Invert color choice* – In invert color choice, the background is black and the font is white.
- *Color contrast* – The color contrast allows educators to choose from several background and lettering color schemes.
- *Overlay color* – The overlay color is the background color of the test.
- *Spoken audio* – Synthetic spoken audio (read aloud with highlighting) is read from left to right and top to bottom. There are three preferences for spoken audio: text only, text and graphics, and nonvisual (this preference also describes page layout for students who are blind).

#### **IV.3.A.ii. Category 2: Supports requiring additional tools or materials**

These supports include braille, switch system preferences, iPad administration, and use of special equipment and materials. These supports are all recorded in the PNP even though the one-switch system is the only option actually activated by PNP.

- *Single-switch system* – Single-switch scanning is activated using a switch set up to emulate the Enter key on the keyboard. Scan speed, cycles, and initial delay may be configured.
- *Two-switch system* – Two-switch scanning does not require any activation in the PNP. The system automatically supports two-switch step scanning.
- *Administration via iPad* – Students are able to take the assessment via iPad.
- *Adaptive equipment used by student* – Educators may use any familiar adaptive equipment needed for the student.
- *Individualized manipulatives* – Individualized manipulatives are suggested for use with students rather than requiring educators to have a standard materials kit. The TIP describes recommended materials and rules governing materials selection or substitution. Having a familiar concrete representation ensures that students are not



disadvantaged by objects that are unfamiliar or that present a barrier to accessing the content.

- *BVI forms* – Alternate forms for students who are blind or have visual impairments (BVI) but do not read braille were developed for certain EEs and linkage levels.<sup>6</sup> BVI testlets are mostly teacher-administered, requiring the test administrator to engage in an activity outside the system and enter responses into KITE Client. In science, one BVI testlet was not teacher-administered; instead, it translated a video-based testlet into an audio story of the scenario that could be listened to with alt text. The general procedures for administering the BVI forms are the same as with other teacher-administered testlets. Additional instructions include the use of several other supports (e.g., human read aloud, test administrator response entry, individualized manipulatives) as needed. When onscreen materials are being read aloud, test administrators are instructed to (1) present objects to the student to represent images shown on the screen and (2) change the object language in the testlet to match the objects being used. Objects are used instead of tactile graphics, which are too abstract for the majority of students with SCD who are also blind. However, educators have the option to use tactile graphics if their student can use them fluently.

#### **IV.3.A.iii. Category 3: Supports provided outside the DLM system**

These supports require actions by the test administrator, such as reading the test, signing or translating, and assisting the student with entering responses.

- *Human read-aloud* – The test administrator may read the assessment to the student. Test administrators were trained to follow guidance to ensure fidelity in the delivery of the assessment. This guidance included the typical tone and rate of speech, avoiding emphasizing the correct response or important information that would lead the student to the correct response. Educators are trained to avoid facial expressions and body language that may cue the correct response and to use exactly the words on screen, with limited exceptions to this guideline, such as the use of shared reading strategies on the first read of a text. Finally, guidance included ensuring that answer choices were always read in the correct order as presented on the screen, with comprehensive examples of all item types. For example, when answer choices are images presented in a triangle order, they are read in the order of top center, bottom left, and bottom right. In most cases, test administrators were allowed to describe graphics or images to students who need those described.

Typically, this additional support would be provided to students who are blind or have visual impairments. Alternate text for graphics and images in each testlet was included in the TIP as an attachment after the main TIP information. Test administrators who needed to read alternate text had the KITE system open and the TIPs in front of them

---

<sup>6</sup> See Chapter III of this manual for further explanation of BVI form availability and design.

while testing so they could accurately read the alternate text provided on the TIPs with the corresponding screen while the student was testing.

- *Sign interpretation of text* – If the student required sign language to understand the text, items, or instructions, the test administrator was allowed to use the words and images on the screen to guide while signing for the student using American Sign Language, Signed Exact English, or any individualized signs familiar to the student. The test administrator was also allowed to spell unfamiliar words when the student did not know a sign for that word and to accept responses in the student's sign language system. Sign is not provided via human or avatar video because of the unique sign systems used by students with SCD who are also deaf/hard of hearing.
- *Language translation of text* – The DLM assessment system does not provide translated forms of testlets because the cognitive and communication challenges for students taking DLM alternate assessments are unique and because students who are English language learners speak such a wide variety of languages; providing translated forms appropriate for all DLM-eligible students to cover the entire blueprint would be nearly impossible. Instead, test administrators are supplied with instructions regarding supports they can provide based on (a) each student's unique combination of language-related and disability-related needs and (b) the specific construct measured by a particular testlet. For students who are English language learners or who respond best to a language other than English, test administrators are allowed to translate the text for the student. The TIP includes information about exceptions to the general rule of allowable translation. For example, when an item assesses knowledge of vocabulary, the TIP includes a note that the test administrator may not define terms for the student on that testlet. Unless exceptions are noted, test administrators are allowed<sup>7</sup> to translate the text for the student, simplify test instructions, translate words on demand, provide synonyms or definitions, and accept responses in either English or the student's native language.
- *Test administrator enters responses for student* – During computer-delivered assessments, if students are unable to physically select their answer choices themselves due to a gap between their accessibility needs/supports and the KITE system, they are allowed to indicate their selected responses to the test administrator through their typical communication modes (e.g., eye gaze, verbal). The test administrator then enters the response. The *Test Administration Manual 2015-2016* provides guidance on the appropriate use of this support to avoid prompting or misadministration. For example, the test administrator is instructed not to change tone, inflection, or body language to cue the desired response or to repeat certain response options after an answer is

---

<sup>7</sup> Simplified instructions, definitions, and flexible response mode are supports also allowed for non-English language learner students.



provided. The test administrator is instructed to ensure the student continues to interact with the content on the screen.

- *Partner-assisted scanning* – Partner-assisted scanning is a commonly used strategy for students who do not have access to or familiarity with an augmentative or communication device or other communication system. These students do not have verbal expressive communication and are limited to response modes that allow them to indicate selections using responses such as eye gaze. In partner-assisted scanning, the communication partner, the test administrator in this case, "scans" or lists the choices that are available to the student, presenting them in a visual, auditory, tactual, or combined format. For test items, the test administrator might read the stem of an item to the student and then read the answer choices aloud in order. In this example, the student could use a variety of response modes to indicate a response. Test administrators may repeat the presentation of choices until the student indicates a response.

#### ***IV.3.B. ADDITIONAL ALLOWABLE PRACTICES***

The KITE Client user interface was specially designed for students with SCD. Testlets delivered directly to students via computer were designed to facilitate students' independent interaction with the computer through special devices such as alternate keyboards, touch screens, or switches as necessary. However, because computerized testing was new to many students using the DLM alternate assessment, the DLM Consortium recognized that students would need various levels of support to interact with the computer. Test administrators were provided general principles for the allowable practices when the supports built into the system did not support a student's completely independent interaction with the system.

When making decisions about additional supports for computer-delivered testlets, educators relied on training they received to follow two general principles. First, the student should be expected to respond to the content of the assessment independently. No matter what additional supports IEP teams and test administrators selected, all should have been chosen with the primary goal of student independence at the forefront. Even if more supports were needed to provide physical access to the computer-based system, the student should have been able to interact with the assessment content and use his or her normal response mode to indicate a selection for each item. Second, test administrators were to ensure that the student was familiar with the chosen supports. Ideally, any supports used during assessment were also used consistently during routine instruction. Students who had never received a support prior to the testing day would be unlikely to know how to make the best use of the support.

In order to select the most appropriate supports during testing, test administrators were encouraged to use their best professional judgment and to be flexible while administering the assessment. Test administrators were allowed to use additional supports beyond PNP options.

The supports detailed in Table 41 were allowed in all computer-delivered and teacher-administered testlets unless exceptions were noted in the TIP.

Table 41. Additional Allowable Practices

Practice	Explanation
Breaks as Needed	Students could take breaks during or between testlets. Test administrators were encouraged to use their best judgment about the use of breaks. The goal should have been to complete a testlet in a single session, but breaks were allowed when the student was fatigued, disengaged, or having behavioral problems that could interfere with the assessment.
Individualized Student Response Mode*	The linkage levels assessed in the teacher-administered testlets did not limit responses to certain types of expressive communication; therefore, all response modes were allowed. Test administrators could represent answer choices outside the system to maximize the student's ability to respond. For example, for students who use eye gaze to communicate, test administrators could represent the answer choices in an alternate format or layout to ensure the student could indicate a clear response.
Use of Special Equipment for Positioning	For students who needed special equipment to access the test material (i.e., a slant board for positioning, or Velcro objects on a communication board), test administrators were encouraged to use the equipment to maximize the student's ability to provide a clear response.
Navigation Across Screens	For students who had a limited experience with, motor skills for, and/or devices for interacting directly with the computer, the test administrator could assist students to navigate across screens or enter the responses.
Use of Interactive Whiteboard	If the student had a severe visual impairment and needed larger presentation of content than the ×5 magnification setting provided, the test administrator could use an interactive whiteboard or projector, or a magnification device that worked with the computer screen to enlarge the assessment to the needed size.

Practice	Explanation
Represent the Answer Options in an Alternate Format	Representing the answer options in an alternate format was allowed, as long as the representation did not favor one answer choice over another. For instance, if presenting the answer choices to a student on a communication board or using objects to represent the answer choices, the correct answer choice could not always be closest to the student or in the same position each time.
Use of Graphic Organizers	If the student was accustomed to using specific graphic organizers, manipulatives, or other tools during instruction, the use of those tools was allowable during the DLM alternate assessment.
Use of Blank Paper	If the student required blank, lined, or unlined paper, this could be provided. Once there was any writing on the paper, it became a secure testing document and needed to be disposed of by shredding at the conclusion of the testing session.
Generic Definitions*	If the student did not understand the meaning of a word used in the assessment, the test administrator could define the term generically and allow the student to apply that definition to the problem or question in which the term was used. Exceptions to this general rule were noted in the TIP for specific testlets.

*Note: \*Allowed using speech, sign, or language translation unless prohibited for a specific testlet.*

Although many supports and practices were allowable for computer-delivered and teacher-administered testlets, there were also practices that test administrators were trained to avoid, including the following:

- Repeating the item activity again after a student has responded or in any other way prompting the student to choose a different answer
- Using physical prompts or hand-over-hand guidance to the correct answer
- Removing answer choices or giving hints to the student
- Rearranging objects to prompt the correct answer – for example, putting the correct answer closer to the student

Test administrators were encouraged to direct any questions regarding whether a support was allowable to the DLM Service Desk or to their SEA.

#### **IV.4. SECURITY**

This section describes secure assessment administration, including test administrator training, security during administration, and the KITE system; secure storage and transfer of data; and

plans for forensic analyses for consortium-wide investigation of potential security issues. Test security procedures during item development and review are described in Chapter III.

#### ***IV.4.A. TRAINING AND CERTIFICATION***

Test security is promoted through required training and certification requirements for test administrators. Test administrators are expected to deliver DLM assessments with integrity and to maintain the security of testlets. Training for test administration detailed the test security measures (see Chapter X). Each year, test administrators must renew their DLM Security Agreement through Educator Portal. The text of the agreement is provided in Figure 34. Test administrators are not granted access to information in the Test Management portion of Educator Portal if they have not indicated their agreement with these terms.

The Dynamic Learning Maps (DLM) Alternate Assessment provides opportunities for flexible assessment administration. However, all DLM assessments - including instructionally embedded assessments chosen by the teacher and delivered during the year 2015 are secure tests.

Test administrators and other educational staff who support DLM implementation are responsible for following the DLM test security standards:

1. Assessments (testlets) are not to be stored or saved on computers or personal storage devices; shared via email or other file sharing systems; or reproduced by any means.
2. Except where explicitly allowed as described in the Test Administration Manual, electronic materials used during assessment administration may not be printed.
3. Those who violate the DLM test security standards may be subject to their state's regulations or state education agency policy governing test security.
4. Educators are encouraged to use resources provided by DLM, including practice activities and released testlets, to prepare themselves and their students for the assessments.

Questions about security expectations should be directed to the local DLM Assessment Coordinator.

I have read this security agreement and agree to follow the standards.

I have read this security agreement and DO NOT agree to follow the standards.

Please type your full name and click Save

Figure 34. Test security agreement text.

Although each state may have additional security expectations and security-related training requirements, all test administrators in each state are required to meet these minimum training and certification requirements.

#### ***IV.4.B. MAINTAINING SECURITY DURING TEST ADMINISTRATION***

There are several aspects of the DLM assessment system design that support test security and test administrator integrity during use of the system. During the spring testing window, each student is tested on only one of three linkage levels for each EE, and the selection of EEs is driven by the adaptive algorithm. Because of the variation in the testlets assigned to different students, test content has more limited exposure than a standardized, single-form test. Because TIPs are the only printed material, there is limited risk of exposure through printed material.

Guidance is provided in the *Test Administration Manual 2015-2016* and on TIPs regarding allowable practices, limits on their use, and proper disposal procedures. This guidance is intended to promote implementation fidelity and reduce the risk of cheating or other types of misadministration. See Chapter IX for test administration evidence related to implementation fidelity.

Agile Technology Solutions (ATS) has procedures in place to handle alleged security breaches. Any reported test security incident is assumed to be a breach and is handled accordingly. In the event of a test security incident, access is disabled at the appropriate level. Depending on the situation, the testing window could be suspended or test sessions could be removed. Test forms could also be removed if exposed, or if data is exposed by a form. If necessary, passwords would be changed for users at the appropriate level.

#### ***IV.4.C. SECURITY IN THE KITE SYSTEM***

As described earlier in this chapter, the KITE system consists of four applications. Educators and students see two of these applications: Educator Portal and KITE Client (Test Delivery Engine). A third application, Content Builder, is the content authoring system that stores test content and associated meta-data. These three applications are relevant to discussions of security in the KITE system. The fourth application, Learning Map Tool, hosts the learning maps models for ELA and mathematics. The KITE system is developed and managed by ATS at the University of Kansas. ATS also administers the DLM Service Desk, which provides customer support for KITE system users in the field.

Operational access to all servers is controlled by keys that are provided only to system administrators who manage the production data center in the operations team. Access to the networking equipment and hardware consoles is limited to the data center itself; remote access to these devices is limited to the data center-specific administration host.

All KITE applications handle educator and administrative passwords using industry-standard encryption techniques. The password policy requires eight characters, including a number, uppercase letter, and a lowercase letter. Passwords expire annually. Returning users may use their username and password from previous years. All applications generate access records that can be reviewed by system administrators to track access. Access to individual KITE applications is controlled according to the policies set forward for that application and the data the application maintains. All access policies and accounts are reviewed periodically to ensure that access to systems is limited to the appropriate populations.

In accordance with Family Educational Rights and Privacy Act (FERPA) rules, educators', administrators', and operations' access to personal student data is limited. The person has to have a legitimate educational interest in the student records to access them. All users in the system are provided the minimum amount of access required. For example, educators can view only their students' records, and users with building-level roles can view and edit student records within a building. A user's role in an organization defines the level of access to records within that organization. Roles may only be assigned by an existing user at a higher level within



the organization. For example, a district-level role may only be assigned by a user with a state-level role; district users may not assign a parallel role to other users. Security levels, groups, and the access provided are reviewed periodically to ensure continued compliance.

KITE Client is a secure browser that prevents access to unauthorized content during a testing session. The KITE web interfaces use industry-standard Secure Socket Layer and Transport Layer Security encryption to securely transfer data to and from the end user from a browser. The KITE system uses load balancing hardware and third party services to both prevent and mitigate the effects of a distributed denial of service attack if one should occur.

#### ***IV.4.D. SECURE TEST CONTENT***

Test content is stored in KITE Content Builder. All items used for released testlets exist in a separate pool from items used for summative purposes, ensuring that no items are shared among secure and non-secure pools. Only authorized users of the KITE assessment system have access to view items. Testlet assignment logic prevents a student from being assigned the same testlet more than once, except in cases of manual override for test-reset purposes.

#### ***IV.4.E. DATA SECURITY***

Beyond uploads to Educator Portal, there is occasionally a need to transfer secure data between the University of Kansas and the partner states. The consortium uses the University of Kansas' secure file transfer protocol (SFTP) system called the Hawk Drive to transfer files securely. This method is used when local educators need to share personally-identifiable information (PII) with the DLM Service Desk agents and when DLM staff deliver score reports and data files to states. Notification of SFTP folder links and passwords are made separately.

The consortium collects PII protocols and usage rules from member states, as illustrated in Appendix D. The protocols are documented on the state summary sheet as part of the collection of policy information about the state. The consortium documents any applicable state laws regarding PII, state PII handling rules, and state-specific PII breach procedures. The information is housed in the shared resources where Service Desk agents and the Implementation team access the information as needed. The protocols are followed with precision due to the sensitive nature of PII and the significant consequences tied to breaches of the data.

The procedures that are implemented in the case of a security incident, privacy incident, or data breach that involves PII or sensitive personal information are implemented by an investigation team that focuses first on mitigation of immediate risk, followed by identification of solutions to identified problems and communication with state partners. A document describing the specific procedures is available on the state partner website (see Appendix D).

#### ***IV.4.F. STATE-SPECIFIC POLICIES AND PRACTICES***

Some states adopt more stringent requirements, above and beyond consortium requirements, for access to test content and for the handling of secure data. Each DLM agreement with a state partner includes a Data Use Agreement. The Data Use Agreement addresses the data security

responsibilities of the consortium in regard to United States Department of Education Regulations 20 U.S.C. § 1232g; 34 CFR Part 99, also known as FERPA. The agreement details the role of the consortium as the holder of data and the rights of the state as the owner of the data. In many cases, the standard Data Use Agreement is modified to include state-specific data security requirements. The consortium documents these requirements on the state summary sheet, and the Implementation and Service Desk teams implement the requirements.

The consortium's Implementation team collects state education authorities' policy guidance on a range of state policy issues such as individual student test resets, district testing window extensions, and allowable sharing of PII. In all cases, the needed policy information is collected on a state summary sheet and recorded in a software program jointly accessed by Service Desk agents and the Implementation team. The Implementation team reviews the state testing policies during Service Desk agent training and provides updates during the state testing windows to supervisors of the Service Desk agents. As part of the training, the Service Desk agents are directed to contact the Implementation team with any questions that require state input, or if the state needs to develop or amend a policy.

#### ***IV.4.G. FORENSIC ANALYSIS PLANS***

There are a large number of possible forensic analyses available for investigating test data for possible security breaches, all of which require the collection of specific types of data. Over time, testing programs develop and refine their data collection architecture and mechanisms for the purpose of doing more sophisticated and useful data forensics. As 2016 was the first operational year for the DLM science assessments, limited forensic analyses were conducted for the following reasons:

- Limited data were available. While the goal is to collect data in the future to allow more meaningful analyses (e.g., keystroke data, item level timestamps), the data that was collected during the 2015–2016 operational year was limited to date and time stamps on the testlets submitted.
- Validity of results from forensic analyses may not be as well supported as they would in subsequent operational testing administrations. Even with ample field testing and practice opportunities, the DLM assessment system is a new approach to assessing the science skills of the population it serves. As such, there may be unanticipated administration situations in the system itself and in the classroom that reflect adjustments to the new assessment system rather than an intentional act or irregularity.

Overall, based on the limited data available for 2015–2016, forensic analyses are not planned until sufficient suitable data are available, likely in 2017 or later. Future analyses may include evaluation of response times to flag outliers, evaluation of answer-changing behavior, analysis of the relationship of FC complexity band and the linkage level of the student's last testlet, and identification of students who began the assessment at a lower linkage level and continually routed up a linkage level until reaching the successor level. Forensic analysis plans have been reviewed by the DLM Technical Advisory Committee (TAC; See Appendix D) and will be updated with the TAC and state partners as additional data become available.



## IV.5. IMPLEMENTATION EVIDENCE FROM 2015–2016 TEST ADMINISTRATION

This section describes evidence collected for 2015–2016 during the operational implementation of the DLM science alternate assessment. The categories of evidence include data relating to the adaptive delivery of testlets in the spring window, administration errors, user experience, and accessibility. Chapter IX has additional descriptions of evidence in support of the validity argument.

### IV.5.A. ADAPTIVE DELIVERY IMPLEMENTATION EVIDENCE

During the spring 2016 test administration, the science assessment was adaptive between testlets, as described in section IV.1.D above.

The correspondence between the FC complexity bands and first assigned linkage levels are shown in Table 35 above and summarized in Table 42.

Table 42. Correspondence of Complexity Bands and Linkage Level

First Contact Complexity Band	Linkage Level
Foundational	Initial
Band 1	Initial
Band 2	Precursor
Band 3	Target

Following the spring 2016 administration, 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 testlet administered and the second was calculated over all students within a grade, content area, and complexity band. The aggregated results can be seen in Table 43.

For the majority of students across all grade bands who were assigned to the Foundational and Band 1 complexity bands by the FC, testlets did not adapt to a higher linkage level after the first assigned testlet. Generally, there was a more even split between students assigned at Band 2 whose testlets did not adapt a linkage level and those students whose testlets did adapt up or down a linkage level between the first and second testlets. However, more students in high school or biology tended to adapt up a level with fewer students adapting down. Finally, for the majority of students assigned to Band 3, linkage levels between first and second testlets did not adapt.

Table 43. Adaptation of Linkage Levels Between the First and Second Testlets

Grade Band	Foundational <sup>a</sup>		Band 1 <sup>b</sup>		Band 2			Band 3	
	Adapted Up (%)	Did Not Adapt (%)	Adapted Up (%)	Did Not Adapt (%)	Adapted Up (%)	Did Not Adapt (%)	Adapted Down (%)	Did Not Adapt (%)	Adapted Down (%)
3-5	37.7	62.3	45.4	54.6	28.1	39.3	32.6	61.5	38.5
6-8	24.7	75.3	29.7	70.3	32.4	32.2	35.4	56.7	43.3
9-12	30.4	69.6	28.5	71.5	19.7	36.4	43.9	76.5	23.5
Biology	35.8	64.2	31.9	68.1	25.3	30.9	43.8	76.3	23.7

<sup>a b</sup> Foundational and Band 1 correspond to testlets at the lowest linkage level, so testlets could not adapt down a linkage level.

#### ***IV.5.B. ADMINISTRATION ERRORS***

Monitoring of testlet assignment uncovered a few incidents that affected student assignment to tests, including misrouting errors due to a local caching server issue and scoring errors, which may have indirectly affected routing because the thresholds are based on percentage of items answered correctly within a testlet. Scoring errors were corrected prior to calculation of summative results. For more information regarding the incidents identified, see Appendix D.

Table 44 provides a summary of the number of students affected by each type of incident, as delivered to states in the Incident Supplemental File. The number of students participating in science who were affected by each incident ranged from 19 to 1,381. In cases where misrouting was identified during the testing window, states were provided with lists of students affected. State representatives were given an option to revert each student's assessment back to the end of the last correctly completed testlet (i.e., the point at which routing failed) and complete the remaining testlets as intended. All students were able to resume testing after the fix was made.

Table 44. Number of Students Affected by Each 2016 Incident

<b>Incident Code</b>	<b>Incident Description</b>	<b>Frequency</b>
1	Potential misrouting due to use of the local caching server.	19
2	Potential misrouting due to missing responses not being treated as incorrect.	252
3	Potential misrouting due to an incorrectly keyed item.	1,381

All reported incidents were shared with the TAC in May 2016, and their feedback was solicited regarding potential impact and next steps for remediation and correction. The TAC recommended that a special circumstance incident file be prepared for states and delivered with the General Research File (GRF; see Chapter VII) to inform the states of all students affected by each issue. States were able to use this file to make determinations about potential invalidation of records at the student level based on state-specific accountability policies and practices.

#### ***IV.5.C. USER EXPERIENCE WITH ASSESSMENT ADMINISTRATION AND KITE SYSTEM***

A survey disseminated to classroom educators in spring 2016 evaluated the user experience of educators who had administered a DLM alternate assessment during the 2015–2016 school year spring window. User experience with the KITE system is summarized in this section, and additional survey contents are reported in the Accessibility section below and in Chapter IX (Validity).

For this section, the data from the overall test administrator survey was filtered to include only respondents from states administering science assessments. While all respondents were from those states, and therefore had the opportunity to administer the science assessment, the survey did not collect information about which subjects the test administrators actually assessed. DLM staff are working to develop and provide subject-specific surveys in future years to more accurately assess the test administration experience.

A total of 1,407 educators from states participating in the science DLM assessment responded to the survey (an estimated response rate of 17.2%). Most of the respondents reported that they had assessed a relatively small number of students during the testing window; 61.5% reported assessing four or fewer students.

The remainder of this section describes educators' responses to the portions of the survey addressing educator experience with DLM assessments and the KITE Client software.

#### IV.5.C.i. Educator Experience

Respondents were asked to reflect on their own experience with the assessments and their comfort level and knowledge with regard to administering them. Most of the questions required respondents to rate results on a four-point scale: strongly disagree, disagree, agree, or strongly agree. Responses are summarized in Table 45. The first two questions (regarding comfort level with the administration of both computer- and teacher-administered testlets) were only displayed if respondents had previously disclosed that they had administered the appropriate kind of testlet.

Table 45. Educator Responses Regarding Test Administration (N=1,407 unless otherwise stated)

Statement	SD		D		A		SA		Missing	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Confidence in ability to deliver computer-administered testlets (N=935)	24	2.57	33	3.53	341	36.47	432	46.20	105	11.23
Confidence in ability to deliver teacher-administered testlets (N=499)	9	1.80	27	5.41	177	35.47	219	43.89	67	13.43
Test administrator training prepared respondent for responsibilities of test administrator	125	8.88	275	19.55	680	48.33	152	10.80	175	12.44

Statement	SD		D		A		SA		Missing	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Respondent knew how to use accessibility features, allowable supports, and options for flexibility	48	3.41	143	10.16	853	60.63	190	13.50	173	12.30
Testlet Information Pages helped respondent to deliver the testlets	118	8.39	274	19.47	692	49.18	148	10.52	175	12.44

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

Educators responded that they were very confident with administering either kind of testlet, with 82.7% reporting responses of agree or strongly agree for computer-administered testlets, and 79.4% reporting responses of agree or strongly agree for teacher-administered testlets. Respondents mostly believed that the required test administrator training prepared them for their responsibilities as a test administrator, with 59.1% responding with agree or strongly agree. Additionally, most educators responded that they knew how to use accessibility features, allowable supports, and options for flexibility (74.1%) and that the TIPs helped them to deliver the testlets (59.7%).

#### IV.5.C.ii. KITE System

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

KITE Client is the software used for the actual administration of DLM testlets. Educators were asked to consider their experiences with KITE Client and respond to each question on a five-point scale: very hard, somewhat hard, neither hard nor easy, somewhat easy, or very easy.

Table 46 summarizes educators' responses to these questions.

Table 46. Ease of Using KITE Client (N = 1,407)

Statement	VH		SH		N		SE		VE		Missing	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Enter the site	33	2.35	106	7.53	197	14.00	379	26.94	484	34.40	208	14.78
Navigate within a testlet	28	1.99	103	7.32	183	13.01	388	27.58	496	35.25	209	14.85
Submit a completed testlet	17	1.21	45	3.20	157	11.16	350	24.88	626	44.49	212	15.07

Statement	VH		SH		N		SE		VE		Missing	
	n	%	n	%	n	%	n	%	n	%	n	%
Administer testlets on various devices	39	2.77	85	6.04	381	27.08	331	23.53	350	24.88	221	15.71

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

Respondents found it to be either somewhat easy or very easy to enter the site (61.3%), to navigate within a testlet (62.8%), to submit a completed testlet (69.4%), and to administer testlets on various devices (48.4%).

Educator Portal is the software used to store and manage student data and to enter PNP and FC information. Educators were asked to assess the ease of navigating and using Educator Portal for its intended purposes. The data are summarized in Table 47 on the same scale that was used to rate experience with KITE Client.

Table 47. Ease of Using Educator Portal (N = 1,407)

Statement	VH		SH		N		SE		VE		Missing	
	n	%	n	%	n	%	n	%	n	%	n	%
Navigate the site	133	9.45	374	26.58	293	20.82	313	22.25	119	8.46	175	12.44
Enter PNP and First Contact information	68	4.83	234	16.63	306	21.75	424	30.14	199	14.14	176	12.51
Manage student data	127	9.03	351	24.95	331	23.53	313	22.25	114	8.10	171	12.15
Manage your account	84	5.97	268	19.05	378	26.87	382	27.15	123	8.74	172	12.22

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

Overall, respondents found it to be either somewhat easy, very easy or neither hard nor easy to navigate the site (51.5%), to enter PNP and FC information (66.0%), to manage student data (53.9%), and to manage his or her account (62.8%).

Finally, respondents were asked to rate their overall experience with KITE Client and Educator Portal on a four-point scale: Poor, Fair, Good, and Excellent. Results are summarized in Table 48.

Table 48. Overall Experience with KITE Client and Educator Portal (N = 1,407)

	Poor		Fair		Good		Excellent		Missing	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
KITE Client	127	9.03	304	21.61	529	37.60	236	16.77	211	15.00
Educator Portal	218	15.49	448	31.84	446	31.70	87	6.18	208	14.78

The majority of respondents reported a positive experience with KITE Client; 37.6% of respondents ranked their experience as good and 16.8% of respondents ranked their experience as excellent. A majority reported a fair to positive experience with Educator Portal, with 31.8% ranking their experience as fair, 31.7% as good, and 6.2% ranking their experience as excellent. Feedback from surveys such as this one, Service Desk tickets, and input from governance board members are all used to improve future training and resources and to prioritize system enhancements that improve user experience.

#### IV.5.C.iii. Accessibility

Guidance around accessibility provided by Dynamic Learning Maps distinguishes between supports that (a) can be used by selecting online features via the PNP, (b) require additional tools or materials, and (c) are provided by the test administrator outside the system. Table 49 shows selection rates for three categories of PNP supports, sorted by rate of use within each category.

Table 49. Personal Needs and Preferences (PNP) Supports Selected for Students, Spring 2016 (N = 22,010)

Support	<i>n</i>	%
<b>Supports Activated by PNP</b>		
Magnification	1,381	6.6
Overlay color	903	4.3
Color contrast	994	4.7
Invert color choice	679	3.2
Read-aloud (text-to-speech) <sup>a</sup>	2,075	9.9
<b>Supports Requiring Additional Tools/Materials</b>		
Individualized manipulatives	6,813	32.0
Calculator	3,865	18.0



<b>Support</b>	<b><i>n</i></b>	<b>%</b>
Single-switch system	1,197	5.7
Alternate form-visual impairment	464	2.2
Two-switch system	268	1.3
Uncontracted braille	41	0.2
<b>Supports Provided Outside the System</b>		
Human read-aloud	18,388	88.0
Test administration enters responses for students	8,927	43.0
Partner-assisted scanning	1,324	6.3
Sign interpretation	366	1.7
Language translation	281	1.3

Note: Multiple selections could be made.

<sup>a</sup> During 2016, read-aloud was not available in the test delivery engine but educators were not prevented from recording the selection on the PNP.

The first category, Supports Activated by PNP, includes supports that are provided within KITE Client. This category of support includes features delivered online. Magnification, which allows educators to choose the amount of screen magnification during testing ( $\times 2$ ,  $\times 3$ ,  $\times 4$ , or  $\times 5$ ), was used by 6.6% of students. Without magnification, the font is Report School, size 22. Overlay color, used by 4.3% of students, allows educators to change the background color of the test from white to an alternate color (blue, green, pink, gray, or yellow). Color contrast allows educators to change the color scheme for the background and font and was used by 4.7% of students. Invert color choice allows educators to change the background color to black and font color to white, which was used by 3.2% of students. Read aloud (text-to-speech; TTS) was used by 9.9% of students. Read aloud (TTS) consists of synthetic spoken audio (read aloud with highlighting).

The second category, Supports Requiring Additional Tools/Materials, includes supports that are recorded in the PNP but provided outside of KITE Client and require additional tools or materials. Individualized manipulatives were used by 32.0% of students. Individualized manipulatives are familiar manipulatives that educators use during instruction. Additional information about individualized manipulatives is provided in the TIP. A single-switch system, used by 5.7% of students, is an interface that emulates the Enter key on the keyboard. Educators set scanning settings for the single-switch system in the PNP. An alternate form-visual impairment was used by 2.2% of students who do not read braille but are blind or have a visual impairment that prevents interaction with the onscreen content. This option is available for

some specific EEs and linkage levels. Alternate forms are not provided at every single EE and linkage level. Two-switch systems were used by 1.3% of students. Two-switch systems consist of two switches and a switch interface that are used to emulate the Tab key to move between choices and the Enter key to select the choice when highlighted. Uncontracted braille was used by 0.2% of students. Uncontracted braille forms are delivered at the state or district level and in braille-ready files or embossed files.

The third category, Supports Provided Outside the System, includes supports offered outside the KITE system that require actions by the test administrator. Human read-aloud was used by 88.0% of students. In human read aloud, test administrators read the assessment aloud to students. Responses were entered by the test administrator for 43.0% of students, an option that is intended for use when students are unable to independently and accurately record their responses in the KITE system. Students indicated their responses through their typical response mode, and test administrators keyed in those responses. Partner-assisted scanning was used by 6.3% of students. Test administrators signed test content for 1.7% of students who used American Sign Language, Exact English, or personalized sign systems. Test administrators translated the text for 1.3% of students who were English language learners or responded best to a language other than English.

Table 50 describes teacher responses to the survey about the student accessibility experience. Teachers were asked to respond to three items using a four-point Likert-type scale (strongly disagree, disagree, agree, strongly agree). The majority of teachers agreed or strongly agreed that the student was able to effectively use accessibility features (73.9%), that accessibility features were similar to ones the student used for instruction (71.3%), and that allowable options for flexibility were necessary when administering the test to meet students' needs (64.8%). These data support the conclusions that the accessibility features of the DLM alternate assessment were effectively used by students, emulated accessibility features used during instruction, and met student needs for test administration.

Table 50. Teacher Report of Student Accessibility Experience (Year-End Model)

Statement	SD		D		A		SA	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Student was able to effectively use accessibility features	213	11.0	292	15.1	1073	55.5	356	18.4
Accessibility features were similar to ones student uses for instruction	175	9.1	375	19.5	1163	60.5	208	10.8
Allowable options for flexibility were needed when administering test to meet student needs	197	8.7	597	26.5	985	43.6	478	21.2

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

## IV.6. CONCLUSION

The DLM system was designed to promote instructional relevance and responsiveness to individual student needs. The dynamic nature of the DLM test administration is reflected in the initial input through the FC survey and later, in the linkage level adaptations based on student prior performance. Assessment delivery options allow for necessary flexibility for student communication mode and linkage level while also being controlled to maximize standardization and support valid scores. To summarize, the DLM system supports necessary flexibility while maintaining standard approaches that support the assessment claims and goals (Chapter I). Feedback collected about the assessment's administration is used to support continuous improvement of the training and resources provided as well as to plan upgrades to the system to improve the assessment experience.

## V. MODELING

The Dynamic Learning Maps (DLM) project draws upon a well-established research base in cognition and learning theory but relatively uncommon operational psychometric methods to provide feedback about student performance. Furthermore, the DLM alternate assessment in science draws upon the existing methods employed by the English language arts (ELA) and mathematics assessments to provide assessment results. The approach uses innovative operational psychometric methods to provide feedback about student mastery of skills and is grounded in a well-established body of research. This chapter describes the psychometric model that underlies the DLM assessment system and describes the process used to estimate item and student parameters from student assessment data.

### V.1. PSYCHOMETRIC BACKGROUND

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

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

Because of the underlying map structure and the goal 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, which draws upon research in cognition and learning theory to provide feedback about student performance. Diagnostic classification models (DCMs) are confirmatory latent class models that characterize the relationship of observed responses to a set of categorical latent variables (e.g., Rupp & Templin, 2008; Rupp, Templin, & Henson, 2010). DCMs are also known as cognitive diagnosis models (e.g., Leighton & Gierl, 2007) or multiple classification latent class models (Maris, 1999) and are mathematically equivalent to Bayesian networks (e.g., Almond, Mislevy, Steinberg, Yan, & Williamson, 2015; Mislevy & Gitomer, 1995; Pearl, 1988). This is the main difference from more traditional psychometric models, such as item response theory, which model a single continuous latent variable. DCMs provide information about student mastery on multiple latent variables or skills of interest.

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

For each item, there is a set of conditional item response probabilities that correspond to the student's possible mastery patterns. When an item measures a single binary attribute, there are only two possible statuses any examinee could have: (1) a master of the attribute, or (2) a non-master of the attribute.

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

## V.2. ESSENTIAL ELEMENTS AND LINKAGE LEVELS

Because the primary goal of the DLM science alternate assessment is to measure what students with the most significant cognitive disabilities know and can do, alternate grade band expectations called Essential Elements (EEs) were created to provide students in the population access to the general education grade-level academic content. The EEs for science were derived from the *Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (National Research Council, 2012; *Framework*) and the Next Generation Science Standards (2013; NGSS). See Chapter II for a complete description. Each EE has an associated set of linkage levels that are ordered by increasing complexity. There are three linkage levels for each EE in science: Initial, Precursor, and Target.

An example of an EE with three linkage levels is given in Figure 35. The EE in the example is from fifth grade and is labeled SCI.EE.5-LS1-1. See Chapter II for more detail on the development of the linkage levels and how they relate to the DLM design.

<b>Essential Element: SCI.EE.5-LS1-1</b>
<b>Target Level:</b> Provide evidence that plants need air and water to grow.
<b>Precursor Level:</b> Provide evidence that plants grow.
<b>Initial Level:</b> Distinguish things that grow from things that don't grow.

Figure 35. EE and linkage levels for SCI.EE.5.LS1-1 (fifth grade science).

### V.3. OVERVIEW OF DLM MODELING APPROACH

There are many statistical models available for estimating the probability of mastery for attributes in a DCM. The statistical model used to determine the probability of mastery for each linkage level for DLM assessments is latent class analysis, which provides a general statistical framework for obtaining probabilities of class membership for each measured attribute (Macready & Dayton, 1977). Student mastery statuses for each linkage level are obtained from an Expectation-Maximization procedure that contributes to an overall profile of mastery.

#### V.3.A. DLM MODEL SPECIFICATION

Due to the administration design, where overlapping data from students taking testlets at multiple linkage levels within an EE were uncommon, simultaneous calibration of all linkage levels within an EE was not possible. Instead, each linkage level was calibrated separately for each EE using separate latent class analyses. Additionally, because items were developed to 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. As such, each class (i.e., masters and non-masters) has a single probability of responding correctly to all items measuring the linkage level, as depicted in Table 51. Similarly, for each item measuring the linkage level, a student has the same probability of providing a correct response. Chapter III details item review procedures intended to support the fungibility assumption as well as field test results that provide preliminary evidence in support of this assumption. Chapter X discusses future studies intended to continue evaluating the fungibility assumption.

Table 51. Depiction of Fungible Item Parameters for Items Measuring a Single Linkage Level

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

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

The DLM scoring model for the 2015-2016 science administration was as follows. Each linkage level within each EE was considered the latent variable to be measured (the attribute). Using latent class analysis, a probability of mastery on a scale of 0 to 1 was calculated for each linkage level within each EE. Students were then classified into one of two classes for each linkage level of each EE: either master or non-master. As described in Chapter VI, a posterior probability of at least 0.8 was required for mastery classification.

All items in a linkage level were assumed to measure that linkage level, meaning the Q-matrix for the linkage level was a column of ones. As such, each item measured one latent variable, resulting in two parameters per item: (1) the probability of answering the item correctly for examinees who have not mastered the linkage level (i.e., the reference group), and (2) the probability of answering the item correctly for examinees who have mastered the linkage level. As per the assumption of item fungibility, a single set of probabilities was estimated for all items within a linkage level. Finally, a structural parameter was also estimated, which was the proportion of masters for the linkage level (the analogous map parameter). In total, three parameters per linkage level are specified in the DLM scoring model: a fungible probability for non-masters, a fungible probability for masters, and the proportion of masters. An explanation of the full model is provided below.

### **V.3.B. MODEL CALIBRATION**

Across all grade spans and courses, there were 34 EEs, all with 3 linkage levels, resulting in a total of  $34 \times 3 = 102$  separate calibration models. Each separate calibration included all items available for the EE and linkage level. Each model was estimated using marginal maximum likelihood using a program that was developed in the R Project for Statistical Computing (R Core Team, 2013).

Latent class analysis was used to obtain the posterior probabilities of mastery, or the likelihood a student mastered the skill being measured. As such, it did not provide scaled score values, but rather a probability on a scale of 0 to 1 representing the certainty of skill mastery. Values closer



to 0 or 1 represent greater certainty of non-mastery or mastery, respectively, whereas values closer to 0.5 represent maximum uncertainty.

A latent class analysis was conducted for each linkage level for each EE. The calibration of the model and final scoring procedure used an Expectation-Maximization algorithm. If the probability of a correct response on item  $i$  for a person in class  $j$  is defined as  $\pi_{ij}$ , the likelihood of a given response pattern for an individual,  $h$ , over  $J$  classes and  $I$  items is defined as:

$$f(\mathbf{X}_h) = \sum_{j=1}^J \eta_j \prod_{i=1}^I \pi_{ij}^{x_i} (1 - \pi_{ij})^{1-x_i}$$

This likelihood (or the log-likelihood if the log is taken) can be maximized using an Expectation-Maximization algorithm using three estimating equations. The expectation step estimates the posterior probability for each student. It is expressed with the following formula (using notation consistent with Bartholomew, Knott, & Moustaki, 2011):

where  $h(j | \mathbf{X}_h)$  represents the posterior probability of a person's class membership given their

$$h(j | \mathbf{X}_h) = \frac{\eta_j \prod_{i=1}^I \pi_{ij}^{x_{ih}} (1 - \pi_{ij})^{1-x_{ih}}}{f(\mathbf{X}_h)}$$

responses. The numerator is the person's probability of item responses for a given class,  $\prod_{i=1}^I \pi_{ij}^{x_{ih}} (1 - \pi_{ij})^{1-x_{ih}}$ , times the probability of membership in that given class,  $\eta_j$ . The denominator ( $f(\mathbf{X}_h)$ ) is the probability of that person's item responses, or the full likelihood, defined above.

The Maximization step estimates the model parameters, including the item parameter,  $\pi_{ij}$  for each item  $i$  and class  $j$ , and the proportion of people in a given class,  $\eta_j$ .

The item parameter was estimated using the following formula:

$$\pi_{ij} = \frac{\sum_{h=1}^N x_{ih} h(j | \mathbf{X}_h)}{N \eta_j}$$

where  $h(j | \mathbf{X}_h)$  represents the posterior probability of a person's class membership given their responses, which was estimated during the Expectation step. The numerator is the sum of the item responses across all respondents,  $x_{ih}$ , weighted by the posterior probability of each respondent being in that class. The denominator is the number of respondents,  $N$ , times the proportion of people estimated to be in the class  $j$ . Thus, the item parameters can be thought of as item  $p$ -values, conditional on group membership. Because the assessment system assumed a fungible item model, all items measuring a linkage level had the same parameter for each class.

The parameter  $\eta_j$  was estimated using the following formula:

$$\eta_j = \frac{\sum_{h=1}^N h(j | \mathbf{X}_h)}{N}$$

where  $h(j|X_h)$  represents the posterior probability of a person's class membership given their responses, which was estimated during the Expectation step. The numerator is the sum of the class membership probabilities across all respondents, and the denominator  $N$  is the number of respondents.

Model calibration in 2016 occurred in June and incorporated operational item responses from the 2015-16 testing window. The model was calibrated using the Expectation-Maximization algorithm until the convergence criteria, change in log-likelihood to  $< 0.00001$ , was met. During the calibration process, initial values of 0.9 and 0.1 for the item parameters were provided for each class, masters and non-masters respectively, to prevent their definitions from switching during estimation. The initial value of  $\eta$  was set to 0.5 for each class.

The final calibrated model parameters from the Maximization step described above were used to run the Expectation step a final time using all operational item responses obtained during the spring window. This resulted in the final student posterior probabilities for each linkage level, which were used for scoring.

#### **V.4. DLM SCORING: MASTERY STATUS ASSIGNMENT**

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 was judged to have mastered a given linkage level within an EE, then the student was assumed to have mastered all lower levels within that EE. This scoring rule relies strongly on the expert opinion used to construct and order the linkage levels that guided item and testlet development. Chapter III provides evidence from the science field test that supports the ordering of linkage levels. Additional validation studies for this scoring rule are currently underway.

In addition to the calculated posterior probability of mastery, students were able to demonstrate mastery of each EE in two additional ways: (1) having answered 80% of all items administered at the linkage level correctly, or (2) the "two-down" scoring rule. The "two-down" scoring rule was implemented to guard against students assessed at the highest linkage level being overly penalized for incorrect responses. Because students did not test at more than one linkage level within an EE, students who tested at the Target level, but did not demonstrate mastery, were assigned mastery status of the linkage level two below (Initial) to prevent them from being penalized for testing at the highest level and not demonstrating mastery. Students who did not demonstrate mastery at the Initial or Precursor levels were considered non-masters of all linkage levels within the EE due to the two-down rule being inapplicable. The same scoring method was implemented for the ELA and mathematics assessments. This scoring method was discussed and determined to be a reasonable approach by the DLM Technical Advisory Committee during a conference call on July 21, 2015.

In order to evaluate the degree to which each mastery assignment rule contributed to students' linkage level mastery status, the percentage of mastery statuses obtained by each scoring rule was calculated, as shown in Figure 36. Posterior probability was given first priority. If mastery was not demonstrated by the posterior probability threshold being met, the next two scoring rules were imposed. Nearly 80% of mastered linkage levels were derived from the posterior probability obtained from the modeling procedure. The other approximately 20% of linkage levels were assigned mastery status by the minimum mastery, or "two-down" rule, and the remaining percentages at each grade span were determined by the percent correct rule. These results indicate that for science, the percent correct rule likely had strong overlap (but was ordered second in priority) with the posterior probabilities, in that correct responses to all 3-4 items measuring the linkage level were likely necessary to achieve a posterior probability above the 0.8 threshold. The percent correct rule does, however, provide mastery status in those instances where providing correct responses to all items still resulted in a posterior probability below the mastery threshold.

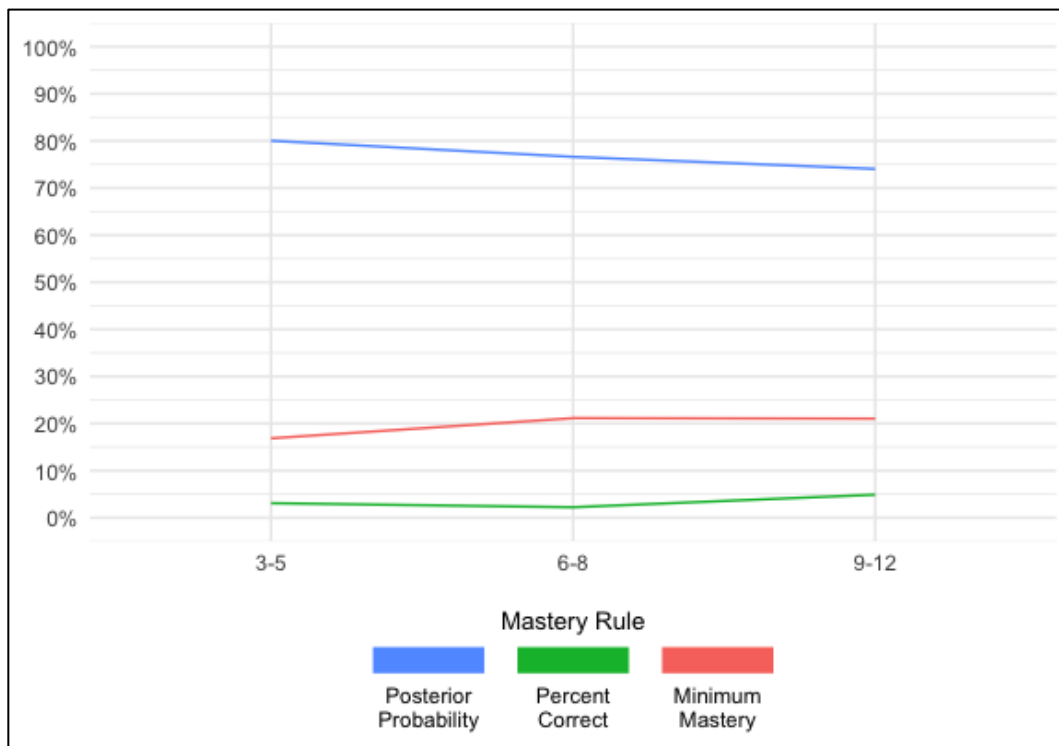


Figure 36. Linkage level mastery assignment by mastery rule for each science grade bands.

## V.5. CONCLUSION

In summary, the DLM modeling approach makes use of well-established research in the areas of Bayesian inference networks and diagnostic classification modeling to determine student mastery of skills measured by the assessment. Latent class analyses are conducted for each linkage level of each EE to determine the probability of student mastery. Items within the linkage level are assumed to be fungible, with equivalent item probability parameters for each

class, due to the conceptual approach used to construct DLM testlets. For each linkage level, a mastery threshold of 0.8 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. Two additional scoring procedures are implemented in addition to posterior probabilities of mastery obtained from the model to ensure students are not overly penalized by the modeling approach in instances where the student only tested at a single linkage level, which include percent 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 based on their posterior probability values obtained from the modeling results.

## VI. STANDARD SETTING

The standard-setting process for the Dynamic Learning Maps (DLM) Science Alternate Assessment System consisted of the adoption of the policy level performance level descriptors (PLDs) originally developed for DLM English language arts and mathematics assessments, a three-day standard setting meeting, and follow-up Technical Advisory Committee (TAC) and state partner evaluation of the process, impact data, and cut points. The purpose of the standard setting activities was to derive recommended cut points for placing students into four performance levels based on results from the 2015–2016 DLM science assessments. This chapter provides a brief description of the development of the rationale for the standard setting approach; the policy PLDs; methods, preparation, procedures, and results of the standard setting meeting; and follow-up state review of the process and results. A more detailed description of the DLM standard setting activities and results can be found in the *2016 Standard Setting: Science*, Technical Report No. 16-03 (Nash, Clark, Karvonen, & Brussow, 2016). The chapter concludes with a full description of the development of grade- and content-specific PLDs that were developed after approval of the consortium cuts.

### VI.1. STANDARD SETTING OVERVIEW

The 2015–2016 school year was the first fully operational testing year for the DLM science assessments. The DLM Consortium's operational testing window ended on June 10, 2016, and the DLM staff conducted standard setting June 15–17, 2016, in Kansas City, Missouri. The standard setting event was a DLM Science Consortium–wide event with the purpose of establishing a set of cut points for the end-of-year assessment. Although DLM state partners voted on acceptance of final cut points, individual states had the option to adopt the consortium cut points or develop their own independent cut points.

The science assessment system follows a year-end testing model, which has a consistent blueprint that is covered in its entirety in the spring testing window. Assessments are available in grade bands (3-5, 6-8, high school) and End-of-Instruction biology.<sup>8</sup> Essential Elements (EEs) were designed to be targets reached by the end of the grade band. However, states in the DLM Science Consortium require assessment of science at different grade levels within the grade bands. As such, expectations for students in lower grades within a grade band could reasonably be lower than expectations for students at higher grades within the same band. Therefore, grade-specific achievement standards were the desired outcome. Based on TAC recommendation and a vote by state partners, cut points were set at tested grade levels within the elementary and middle school grade bands (cut points in grades 4, 5, 6, and 8).<sup>9</sup>

#### VI.1.A. STANDARD SETTING APPROACH: RATIONALE AND OVERVIEW

The approach to standard setting was developed to be consistent with the DLM Alternate Assessment System's design and to rely on established methods; recommended practices for

---

<sup>8</sup> States had the option of choosing which high school assessment to administer.

<sup>9</sup> No states tested science in grades 3 or 7.

developing, implementing, evaluating, and documenting standard settings (Cizek, 1996; Hambleton, Pitoniak, & Copella, 2012); and the *Standards on Educational and Psychological Testing* (2014). The DLM standard setting approach (Clark, Nash, Karvonen, & Kingston, in press) used mastery classifications of skills and was consistent with the approach used to set English language arts and mathematics standards in 2015. The panel process drew from several established methods, including generalized holistic (Cizek & Bunch, 2006) and body of work (Kingston & Tiemann, 2012).

Because the DLM assessment makes use of diagnostic classification modeling rather than traditional psychometric methods, the DLM standard setting approach relied on aggregation of dichotomous classifications of linkage-level mastery for each EE in the blueprint. Drawing from the generalized holistic and body-of-work methods, panels used a profile approach to classify student mastery of linkage levels into performance levels. Profiles provided a holistic view of student performance by summarizing across the EEs and linkage levels. Cut points were determined by evaluating the total number of mastered linkage levels. Although the number of mastered linkage levels is not an interval scale, the process for identifying the DLM cut points is roughly analogous to assigning a cut point along a scale-score continuum.

Figure 37 summarizes the complete set of sequential steps included in the DLM standard setting process. This includes steps conducted before, during, and after the on-site meeting in June 2016.

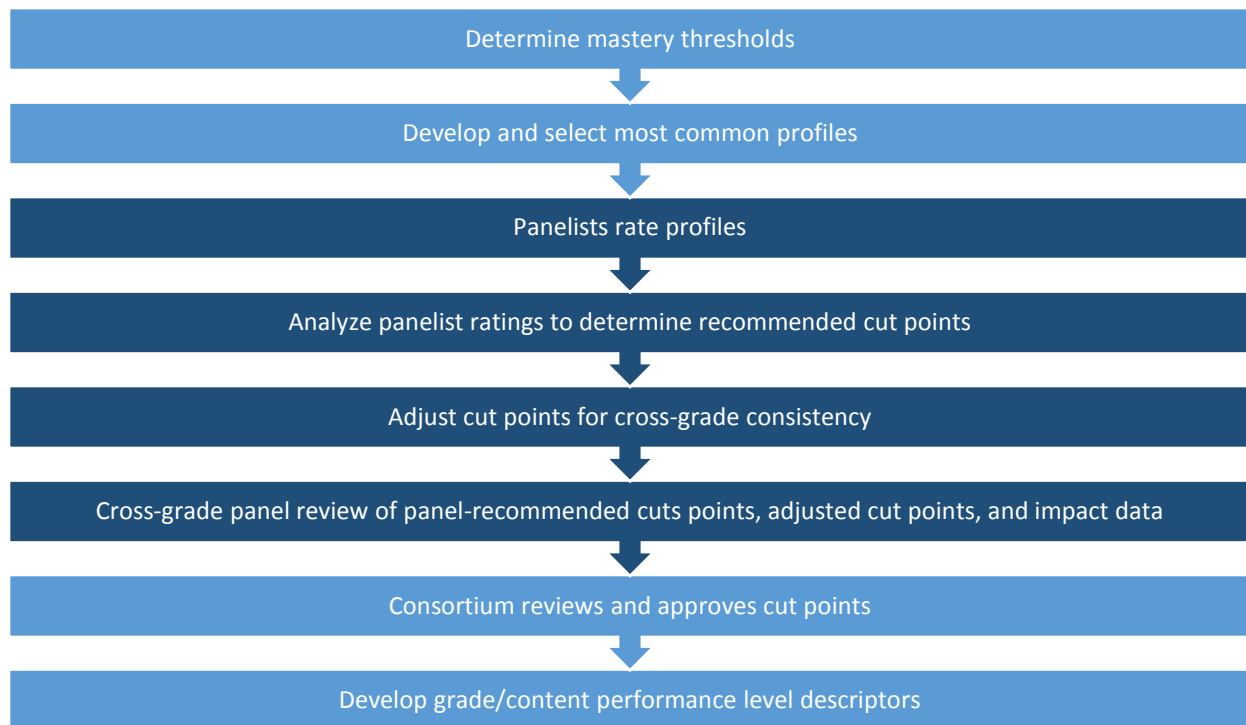


Figure 37. Steps of the DLM standard-setting process.

Note: Dark shading represents steps conducted at the standard-setting meeting in June 2016.

### ***VI.1.B. POLICY PERFORMANCE LEVEL DESCRIPTORS***

DLM science state partners chose to use the existing DLM performance levels and policy PLDs originally developed for English language arts and mathematics.

DLM state partners developed policy PLDs through a series of conversations and draft PLD reviews between July and December 2014. In July 2014, the state partners discussed general concepts that should be reflected in the PLDs and reviewed several examples of descriptors for three, four, and five performance levels. In fall 2014, the state partners indicated the number of levels they would require and gave feedback on additional iterations of PLDs that had been revised based on previous input. By December 2014, the PLDs were finalized. All states participating in the 2014–2015 operational assessment required four performance levels. The final version of policy PLDs is summarized in Table 52. The consortium-level definition of proficiency was at target.



Table 52. Final Performance Level Descriptors for the DLM Consortium

<b>Performance Level Descriptors</b>
The student demonstrates <i>emerging</i> understanding of and ability to apply content knowledge and skills represented by the Essential Elements.
The student's understanding of and ability to apply targeted content knowledge and skills represented by the Essential Elements is <i>approaching the target</i> .
The student's understanding of and ability to apply content knowledge and skills represented by the Essential Elements is <i>at target</i> .
The student demonstrates <i>advanced</i> understanding of and ability to apply targeted content knowledge and skills represented by the Essential Elements.

### ***VI.1.C. PROFILE DEVELOPMENT***

Prior to the standard-setting meeting, student performance on linkage levels was used to create profiles of student learning.

The first step to develop profiles required obtaining mastery classifications at the linkage level. Based on input from TAC and DLM state partners, an agreed-on cut was applied to students' posterior probabilities from the diagnostic classification model calibration. For each linkage level, all students with a probability greater than or equal to .8 would receive a linkage level mastery status of 1, or mastered. All students with a probability lower than .8 would receive a linkage level mastery status of 0, or not mastered.<sup>10</sup>

Given the linkage level mastery data, profiles of student mastery were created that summarize linkage level mastery by EE. Profiles were created using data for each grade band. Each profile listed all the linkage levels for all the EEs from the blueprint, with green-shaded boxes indicating the mastered linkage levels and blue-shaded boxes indicating the tested but not yet mastered linkage levels. Figure 38 provides an example profile for a hypothetical student.

---

<sup>10</sup> Maximum uncertainty occurs when the probability is .5, and maximum certainty occurs when the probability approaches 0 or 1. Considering the risk of false positives and false negatives, the threshold used to determine mastery classification was set at .8.


End of Year Learning Profile			
<b>SUBJECT:</b> Science <b>MODEL:</b> Year-End		<b>GRADE:</b> Middle school science <b>PROFILE ID:</b> 0122	
		 <b>YEAR:</b> 2015-16 <b>TOTAL LL:</b> 14	
Essential Element	Level Mastery		
	1	2	3 (Target)
SCI.MS.PS.1.2	Identify change	Gather data on properties before and after chemical changes	Interpret data on properties before and after chemical changes
SCI.MS.PS.2.2	Identify ways to change movement	Investigate and identify ways to change motion	Investigate and predict changes in motion
SCI.MS.PS.3.3	Identify objects and materials that minimize thermal energy transfer	Investigate objects/materials and predict changes in thermal energy transfer	Refine a device to minimize or maximize thermal energy transfer
SCI.MS.LS.1.3	Recognize major organs	Model how organs are connected	Make a claim how structure and function support survival
SCI.MS.LS.1.5	Match organisms to habitats	Identify factors that influence growth	Interpret data to show that resources influence growth
SCI.MS.LS.2.2	Identify food that animals eat	Classify animals by what they eat	Identify producers and consumers in a food chain
SCI.MS.ESS.2.2	Identify differences in weather conditions from day to day	Identify geoscience processes that impact landforms	Explain how geoscience processes change Earth's surface
SCI.MS.ESS.2.6	Interpret weather information to identify conditions	Interpret weather information to compare conditions	Interpret weather information to make predictions
SCI.MS.ESS.3.3	Recognize resources that are important for life	Recognize ways that humans impact the environment	Monitor and minimize an impact on the environment

Figure 38. Example standard setting profile for a hypothetical student.

Note: Green shading represents linkage level mastery. Blue shading represents no evidence of mastery for the Essential Element.

Profiles were available for all students who participated in the spring window by May 12, 2016 ( $N = 20,448$ ; grades 3-5,  $n = 5,455$ ; grades 6-8,  $n = 5,622$ ; grades 9-12,  $n = 5,098$ ; End-of-Instruction biology course,  $n = 1,312$ ). The frequency with which each precise profile (i.e., pattern of linkage level mastery) occurred in this population was computed. Based on these results, the three most common profiles were selected for each possible total linkage level mastery value (i.e., total number of linkage levels mastered) for each grade or course. In instances where data were not available at a specific linkage level value, (e.g., no students mastered exactly 26 linkage levels for the grade or course), profiles were based on simulated data. To simulate profiles, the DLM science content team used adjacent profiles for reference and created simulated profiles that represented likely patterns of mastery, consistent with the approach used for English language arts and mathematics. Less than 4% of all the science profiles developed were simulated.<sup>11</sup>

<sup>11</sup> Further detail on specific procedures for preparing standard-setting profiles may be found in Chapter 1 of Technical Report No. 16-03.

#### ***VI.1.D. PANELISTS***

The DLM staff worked with participating states in March 2016 to recruit standard setting panelists. State partners were responsible for communicating within their state to recruit potential panelists. Panelists sought were teachers and school and district administrators with both content knowledge and expertise in the education and outcomes of students with the most significant cognitive disabilities (SCD). Other subject-matter experts, such as faculty of higher-education institutions or state/regional educational agency staff, were also suggested for consideration.

The 32 panelists who participated in standard setting represented varying backgrounds. Table 53 and Table 54 summarize their demographic information. Most of the selected panelists were classroom teachers. Panelists had a range of years of experience with science and working with students with SCD.

Nearly half of the participants had experience with setting standards for other assessments ( $n = 15$ ). Some panelists already had experience with the DLM assessment, either from writing items ( $n = 8$ ) or externally reviewing items and testlets ( $n = 10$ ). Only one panelist reported having less than one year or no experience with alternate assessments; that panelist was university faculty/staff with 19 years of experience with science content.<sup>12</sup>

---

<sup>12</sup> Further detail on standard setting volunteers, selection process, and panel composition may be found in Chapter 3 of Technical Report No. 16-03.

Table 53. Demographic Characteristics of Panelists

Demographic Category	Count
<b>Gender</b>	
Female	29
Male	3
<b>Race</b>	
African American	3
American Indian/Alaska Native	3
Asian	2
Hispanic/Latino	2
Native Hawaiian/Pacific Islander	1
White	21
<b>Professional Role</b>	
Classroom teacher	23
District staff	6
State education agency staff	2
University faculty/staff	2
Other	8
<b>Total</b>	<b>32</b>

Table 54. Panelists' Years of Experience

	<i>M</i>	<b>Min</b>	<b>Max</b>
Students with the most significant cognitive disabilities	14.3	2.0	30.0
Science	13.2	1.0	30.0

### ***VI.1.E. MEETING PROCEDURES***

Panelists participated in a profile-based standard setting procedure to make decisions about cut points. The panelists participated in four rounds of activities in which they moved from general to precise recommendations about cut points.

The primary tools of this procedure were range-finding folders and pinpointing folders. The range-finding folders contained profiles of student work that represented the range of total linkage levels. Pinpointing folders contained profiles for specific areas of the range.

Throughout the procedure, the DLM staff instructed panelists to use their best professional judgment and consider all students with SCD to determine which performance level best described each profile. Each panel had one grade-level or course set of cut points to determine.

The subsequent sections provide details of the final procedures, including quality assurance used for determining cut points.<sup>13</sup>

### **VI.1.E.i. Training**

Panelists were provided with training both before and during the standard setting workshop. Advance training was available online on-demand in the 10 days prior to the standard setting workshop. The advance training addressed the following topics:

1. Students who take the DLM assessments,
2. Content of the assessment system, including EEs for science, domains and topics, linkage levels, and alignment,
3. Accessibility by design, including the framework for the DLM Alternate Assessment System's cognitive taxonomy and strategies for maximizing accessibility of the content; the use of the Access (Personal Needs and Preferences) Profile (PNP) to provide accessibility supports during the assessment; and the use of the First Contact survey to determine linkage level assignment,
4. Assessment design, including item types, testlet design, and sample items from various linkage levels in science,
5. An overview of the assessment model, including test blueprints and the timing and selection of testlets administered, and
6. A high-level introduction to two topics that would be covered in more detail during on-site training: the DLM approach to scoring and reporting and the steps in the standard setting process.

Additional panelist training was conducted at the standard setting workshop. The purposes of on-site training were twofold: (1) to review advance training concepts that panelists had indicated less comfort with, and (2) to complete a practice activity to prepare panelists for their responsibilities during the panel meeting. The practice activity consisted of range finding using training profiles for just a few total linkage levels mastered (e.g., 5, 10, 15, 20).

Overall, panelists participated in approximately 8 hours of standard setting-related training before beginning the practice activity.

### **VI.1.E.ii. Range Finding**

During the range-finding process, panelists reviewed a limited set of profiles to assign general divisions between the performance levels using a two-round process. The goal of range finding

---

<sup>13</sup> Further information regarding all meeting procedures and fidelity of the final procedures to the planned procedures can be found in Chapter 4 of the Technical Report No. 16-03.

was to locate ranges (in terms of number of linkage levels mastered) where panelists agreed that approximate cut points should exist.

First, panelists independently evaluated profiles and identified the performance level that best described each profile. Once all panelists completed their ratings, the facilitator obtained the performance level recommendations for each profile by a raise of hands.

After a table discussion of how panelists chose their ratings, the panelists were given the opportunity to adjust the independent ratings they chose. A second round of ratings were then recorded and shared with the group.

Using the second round's ratings, built-in logistic regression functions calculated the probability of a profile being categorized in each performance level, conditioned on number of linkage levels mastered, and the most likely cut points for each performance level were identified. In instances where the logistic regression function could not identify a value (e.g., the group unanimously agreed on the categorization of profiles to performance levels), psychometricians evaluated the results to determine the approximate cut point based on panelist recommendations.<sup>14</sup>

### **VI.1.E.iii. Pinpointing**

Pinpointing rounds followed after range finding. During pinpointing, panelists reviewed additional profiles to refine the cut points. The goal of pinpointing was to identify specific cut points in terms of number of linkage levels mastered within the general ranges determined in range finding.

First, panelists reviewed profiles for seven total scores including and around the cut point value identified during range finding (e.g., total scores at the cut point and +/-3 total linkage levels mastered). Next, panelists independently evaluated these profiles and assigned each a performance level. Once all panelists completed their ratings, the facilitator obtained the recommendations for each profile by a show of hands.

After discussion of the ratings, a second round of ratings commenced. Panelists were given the opportunity to adjust their independent ratings if they chose. Using the second round's ratings, built-in logistic regression functions calculated the probability of a total score being categorized in each performance level conditional on the number of linkage levels mastered, and the most likely cut points for each performance level were identified. In instances where the logistic regression function could not identify a value (e.g., the group unanimously agreed on the categorization of profiles to performance levels), psychometricians evaluated the results to determine the final recommended cut point based on the panelist recommendations (see Footnote 14).

---

<sup>14</sup> Chapter 4 of the Technical Report No. 16-03 provides greater detail on range finding and pinpointing and includes details on the number of linkage levels per grade and course.

### VI.1.E.iv. Adjusting the Cut Points

To mitigate the effect of sampling error and issues related to a system of cut points across a series of grade levels, statistical adjustments were made to the panel-recommended cut points in an effort to systematically smooth distributions within the system of cut points being considered. No adjustments were made for End-of-Instruction biology because both the standards assessed and the students taking this assessment were assumed to be very different.<sup>15</sup>

### VI.1.E.v. Vertical Articulation Panel

Finally, a vertical articulation panel was convened to ensure that cut points progressed logically as content expectations increased according to grade level. Once the panel-recommended cut points were set, two representatives from each panel (except End-of-Instruction biology<sup>16</sup>) convened for a review and discussion of the grade-level panels' recommended cut points, the statistically adjusted cut points (methodology discussed in a subsequent section), and the associated impact data for each. The process began with a discussion of panelists' content-based rationales for their ratings and their panel's recommended cut points. Next, panel-recommended cut points and statistically adjusted cut points, with impact data for each, were presented for all grade-level panels and high school. After a whole-group discussion about the system of cut points focusing on content-based rationales for results, the panel's conclusions and final recommendation were documented.

## VI.2. RESULTS

This section summarizes the panel-recommended and statistically adjusted cut points and the impact data and evaluation results.<sup>17</sup>

### VI.2.A. PANEL-RECOMMENDED AND ADJUSTED CUT POINTS

Table 55 includes a summary of the cut-point recommendations reached by the panelists following the range-finding and pinpointing process.

Table 55. Panel Cut-Point Recommendations

Grade	Emerging/ Approaching	Approaching/ Target	Target/ Advanced	Maximum Number of Linkage Levels
4	9	16	22	27

<sup>15</sup> The specific steps applied to each subject within each grade level can be found in Chapter 5 of the Technical Report No. 16-03.

<sup>16</sup> End-of-Instruction biology was not included in the vertical articulation process as it was not expected that students in one course were representative of the students in the general high school grade band and there was no reason to expect that a single End-of-Instruction biology assessment was somehow contiguous to a previous grade-level, multi-domain assessment.

<sup>17</sup> Additional detailed results are provided in Chapter 5 of the Technical Report No. 16-03.



Grade	Emerging/ Approaching	Approaching/ Target	Target/ Advanced	Maximum Number of Linkage Levels
5	11	18	25	27
6	9	15	22	27
8	11	16	23	27
9-12	9	17	24	27
Biology	9	15	22	30

To mitigate the effect of sampling error and the issues related to a system of cut points across a series of grade levels, statistical adjustments were made to the panel-recommended cut points into systematically smooth distributions within the system of cut points being considered.

Table 56 summarizes the adjusted cut points that used the methods described above and the impact data for those adjusted cut points.<sup>18</sup>

Table 56. Adjusted Cut-Point Recommendations

Grade	Emerging/ Approaching	Approaching/ Target	Target/ Advanced	Maximum Number of Linkage Levels
4	9	15	21	27
5	10	17	25	27
6	10	15	21	27
8	10	16	23	27
9-12	8	16	23	27

### ***VI.2.B. VERTICAL ARTICULATION PANEL PROCESS***

The vertical articulation panel provided a strong cross-grade, content-based rationale for recommending all of the adjusted cut points, with the exception of one cut point. Specifically, they recommended retaining the panel-recommended cut point for the sixth grade cut between emerging and approaching the target. As the adjusted cut points at this level for sixth and eighth grades were the same, they chose to retain the panel-recommended cut to maintain a higher performance expectation for students in the eighth grade.

### ***VI.2.C. DLM STAFF-RECOMMENDED CUT POINTS AND IMPACT DATA***

DLM staff accepted the recommendations made by the vertical articulation panel and recommended those cut scores for all subsequent reviews made by the TAC and DLM science states. The final DLM staff-recommended cut points for science are shown in Table 57. Figure 39 displays the results of the DLM staff-recommended cut points in terms of impact for each

---

<sup>18</sup> Cut points for biology were not statistically adjusted.

grade and course.<sup>19</sup> These cut points were used to assign students to performance levels after the spring 2016 operational administration of the science assessment. Table 58 includes the demographic data for students included in the impact data. Similar percentages for subgroups of students were found for the population of students included in the final reporting of the 2015-2016 assessment (Chapter VII).

Table 57. DLM Staff–Recommended Cut Points for Science

<b>Grade</b>	<b>Emerging/ Approaching</b>	<b>Approaching/ Target</b>	<b>Target/ Advanced</b>	<b>Maximum Number of Linkage Levels</b>
4	9	15	21	27
5	10	17	25	27
6	9	15	21	27
8	10	16	23	27
9-12	8	16	23	27
Biology	9	15	22	30

---

<sup>19</sup> Chapter 5 of the Technical Report No. 16-03 reports the frequency distributions for the panel-recommended cut points.

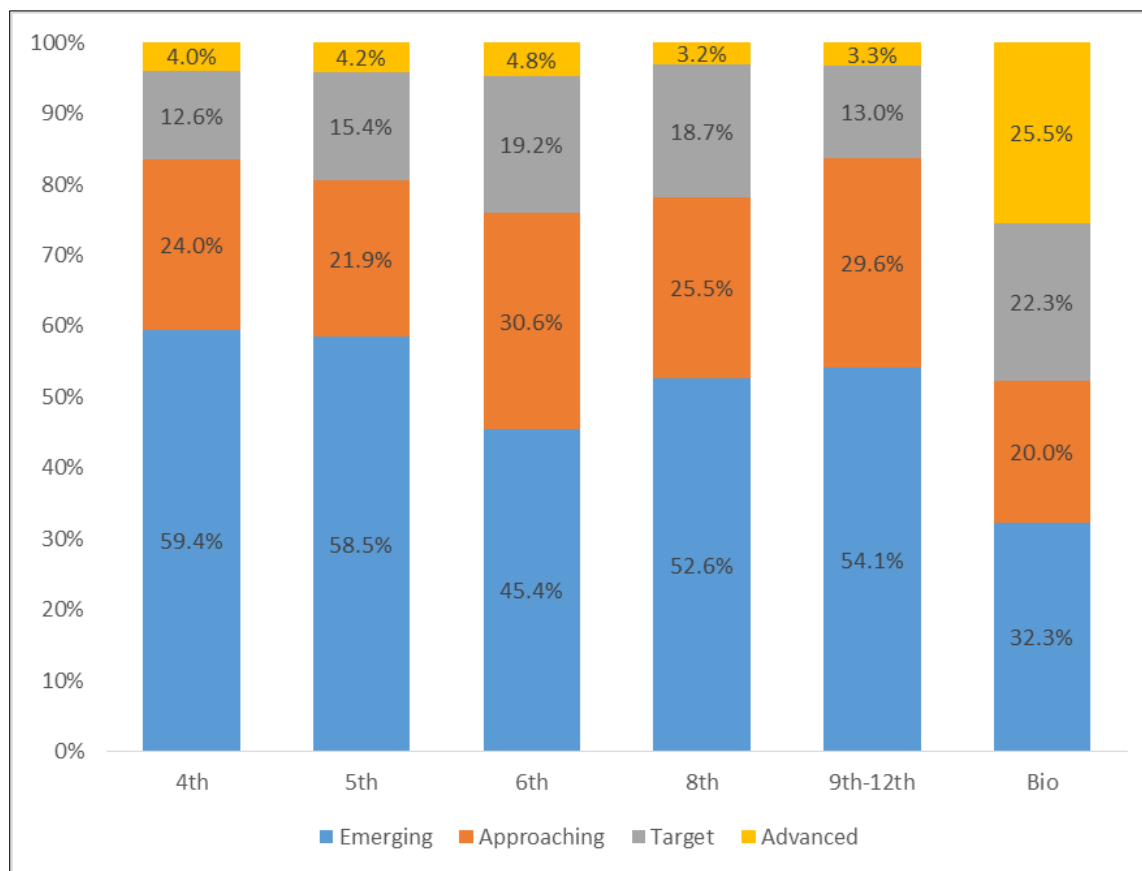


Figure 39. Science impact data using DLM staff-recommended cut points.

Table 58. Demographic Information for Students Included in Impact Data

Demographic	<i>n</i>	%
<b>Gender</b>		
Female	8,657	35.1
Male	15,981	64.9
Missing data	4	< 0.1
<b>Primary Disability</b>		
Intellectual disability	3,643	14.8
Autism	1,760	7.1
Multiple disabilities	855	3.5
Other health impairment	613	2.5
Specific learning disability	194	0.8
Other	1,895	7.7
Missing data	15,682	63.6
<b>Comprehensive Race</b>		
White	15,723	63.8

<b>Demographic</b>	<i>n</i>	%
African American	5,377	21.8
Asian	644	2.6
American Indian	829	3.4
Alaska Native	32	0.1
Two or More Races	1,873	7.6
Native Hawaiian/Pacific Islander	51	0.2
Missing data	113	0.5
<b>Hispanic Ethnicity</b>		
No	21,733	88.2
Yes	2,845	11.5
Missing data	64	0.3
<b>ESOL Participation</b>		
Not an ESOL eligible student and not an ESOL monitored student	23,819	96.7
ESOL eligible or monitored student	823	3.3
<b>Science Band<sup>a</sup></b>		
Foundational	2,065	8.4
Band 1	4,530	18.4
Band 2	4,963	20.1
Band 3	13,084	53.1
<b>Total</b>	<b>24,642</b>	

<sup>a</sup> Science band for the 2015–2016 administration was based on First Contact expressive communication survey questions only. See Chapter IV for more detail.

#### ***VI.2.D. EXTERNAL EVALUATION OF STANDARD SETTING PROCESS AND RESULTS***

A DLM TAC member was on site for the duration of the standard setting event and reported that the standard setting meeting was well planned and implemented, the staff were helpful to the panelists, and the panelists worked hard to set standards. The full TAC accepted a resolution about the adequacy, quality of judgments, and extent to which the process met professional standards.<sup>20</sup>

The panel-recommended cut points, the DLM staff–recommended cut points, and the associated impact data for both sets of cut points were presented to the TAC and consortium states for review. The TAC supported the DLM adjustment method and resulting cut points. Following the states’ review process and discussion with the DLM team, the states voted to accept the DLM staff–recommended cut points as the final consortium cut points with no further adjustment.

---

<sup>20</sup> The TAC chair memorandum and TAC resolution are provided in the Technical Report No. 16-03, Appendix E.

### VI.3. GRADE-LEVEL PERFORMANCE LEVEL DESCRIPTORS

Based on the general approach to standard setting, which relied on mastery profiles to anchor panelists' content-based judgments, grade- and content-specific PLDs were not used during standard setting. Instead, these grade-specific PLDs emerged based on the final cut points and were syntheses of content from the more fine-grained linkage level descriptors. Grade-specific PLDs were completed after standard setting in 2016. Standard setting panelists began the process by drafting lists of skills and understandings that they determined were characteristic of specific performance levels after cut points had been established. In general, these draft lists of skills and understandings were based on the linkage levels described in the mastery profiles used for standard setting—either separate linkage level statements or syntheses of multiple statements. These draft lists of important skills were collected and used as a starting point for the DLM content team as it developed language for grade-specific descriptions for each performance level. The purpose of these content descriptions was to provide information about the knowledge and skills that are typical for each performance level. The content team prepared to draft PLDs by consulting published research related to PLD development (e.g., Perie, 2008) and reviewing PLDs developed for other assessment systems in order to consider grain size of descriptive language and variety of formats for publication. In addition to the draft lists generated by standard setting panelists, the content team used the following materials as they drafted specific language for each grade- and content-specific PLD:

- The DLM test blueprint
- The cut points set at standard setting for each grade
- Sample mastery profiles used as part of standard setting
- Essential Element Concept Maps (EECMs) for each EE included on the blueprint for each grade level
- Linkage level descriptions for every EE

The content team reviewed the EEs, EECMs, and linkage level descriptors for the profiles to determine skills and understandings assessed at the grade level. These skills and understandings came from each conceptual area assessed at the specific grade level and vary from one grade to the next. Then the content team reviewed the draft skill lists created by standard setting panelists and final cut points approved by the consortium. The content team then used the sample mastery profiles to consider the types and ranges of student performances that could lead to placement into specific performance levels. Using these multiple sources of information, the content team evaluated the placement of skills into each of the four performance levels.

While not an exhaustive list of all the content related to each EE, the synthesis of standard setting panelist judgments and content team judgments provided the basis for descriptions of the typical performance of students showing mastery at each performance level. As the content team drafted PLDs for each grade, they reviewed the descriptors in relationship to each other to ensure that there was differentiation in skills from one grade to the next.

The content team prepared initial drafts of the grade- and content-specific descriptions for grades 4, 5, 6, 8, high school, and EOI biology. Project staff reviewed these drafts internally. The DLM state partners reviewed the draft PLDs after the December 2016 consortium governance meeting. Project staff asked state partners to review the progression of descriptors from grade to grade within the four performance levels in grades 4, 5, 6, 8, high school, and EOI biology to provide general feedback to the initial drafts. Feedback from state partners was minimal. The feedback focused on placement of content-specific descriptors within the four performance levels and domains, and formatting.

After the review period ended, the content team responded to feedback received by reviewing placement of content specific descriptors within the four performance levels and domains. A full editorial review was completed on the draft PLDs after the review period. Final versions of the grade and content PLDs are available on the DLM website (<http://dynamiclearningmaps.org/content/assessment-results>). Appendix E contains examples of grade and content PLDs.

## VII. ASSESSMENT RESULTS

Following the discussion of the standard-setting process in Chapter VI, Chapter VII reports the 2015–2016 spring operational results of the Dynamic Learning Maps (DLM) science alternate assessment. This chapter presents student participation data, final results in terms of the percentage of students at each performance level (impact), and subgroup performance by gender, race, ethnicity, and English language learner status. This chapter also reports the distribution of students by the highest linkage level mastered. Finally, this chapter and Appendix F describe all the various types of score reports, data files, and interpretive guides.

### VII.1. STUDENT PARTICIPATION

The 2015–2016 spring science assessments were administered to a total of 20,214 students, including states administering the End-of-Instruction (EOI) biology assessment and districts affiliated with the Bureau of Indian Education, as shown in Table 59. The 168,367 assessment sessions (testlets administered) were administered by 7,846 educators in 5,589 schools and 2,008 school districts.

Table 59. Student Participation by State or Agency

State/Agency	Students
Choctaw	7
Illinois	4,639
Iowa	919
Kansas	1,150
Miccosukee	5
Mississippi	1,643
Missouri	4,617
Oklahoma	2,434
West Virginia	850
Wisconsin	3,950
<b>Total</b>	<b>20,214</b>



More than 6,000 students participated in both the elementary grade band and the middle school grade band<sup>21</sup> (see Table 60). In high school, more than 7,500 students participated. The high school grade band includes students participating in the DLM alternate EOI biology assessment in lieu of the DLM high school science assessment.

Table 60. Student Participation by Grade

Grade	Students
<b>Elementary</b>	
3	435
4	1,343
5	4,406
<b>Middle</b>	
6	798
7	599
8	5,029
<b>High</b>	
9	1,210
10	1,972
11	3,956
12	466
<b>Total</b>	<b>20,214</b>

Table 61 summarizes the demographic characteristics of students who participated in the spring 2015–2016 assessments. The majority of participants were male (64.8%), and the majority of participants were white (64.6%). Only 0.8% of students were reported as being eligible for or monitored for English language learner services. Please note that because teachers were not required to complete all of the student demographic information, some variables in the following tables have missing data.

---

<sup>21</sup> In an effort to increase science instruction beyond the tested grades, several states promoted participation in the science assessment at all grade levels (i.e., did not restrict participation to the grade levels required for accountability purposes). Grade levels 3 and 7 are not tested for accountability purposes in the current DLM science states.

Table 61. Demographic Characteristics of Participants

Subgroup	<i>n</i>	%
<b>Gender</b>		
Female	7,122	35.23
Male	13,090	64.76
Missing	2	.01
<b>Race</b>		
White	13,049	64.55
African American	4,335	21.45
Alaska Native	38	0.19
American Indian	646	3.20
Asian	526	2.60
Native Hawaiian or Pacific islander	44	0.22
Two or More Races	1,483	7.34
Missing	93	0.46
<b>Hispanic Ethnicity</b>		
No	17,847	88.29
Yes	2,300	11.38
Missing	67	0.33
<b>English Language Learner (ELL) Participation</b>		
Not ELL eligible or monitored	367	1.82
ELL eligible or monitored	160	0.79
Missing	19,687	97.39

## VII.2. STUDENT PERFORMANCE

Student performance on DLM assessments is interpreted using cut points determined during standard setting (see Chapter VI) that 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).

As described in Chapter VI, students were considered masters of a linkage level if (1) their posterior probability from the diagnostic classification model was greater than or equal to .8 or (2) the proportion of items that they responded to correctly within the linkage level was greater than or equal to .8. If the student did not demonstrate mastery at the level assessed, mastery was assigned two linkage levels below the level assessed. In addition, students were considered masters of all linkage levels below the level at which they demonstrated mastery.

Mastery status values were aggregated within and across EEs to obtain the total number of linkage levels the student mastered. Although the total number of mastered linkage levels is not

a raw or scale score, the number of linkage levels mastered across EEs assessed was the metric used for setting performance level cut points.

For the 2015–2016 administration, student performance was reported using the four performance levels approved by the DLM Consortium:

- 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 62 reports the performance distributions (i.e., the percentage of students at each performance level) from the 2015–2016 spring administration for science.<sup>22</sup>

The percentage of students who achieved at the Target or Advanced performance levels was under 20% for all grades. At the elementary level, 64% of students performed at the Emerging performance level and slightly less than 60% of students performed at the Emerging performance level in the middle and high school levels. The majority of students were categorized as either Emerging or Approaching the Target performance levels, with the exception of students in the one science state using EOI biology, where there was a more even distribution across the four performance levels.

Table 62. Percentage of Students by Grade and Performance Level (n = 20,214)

Grade or Course	N	Emerging (%)	Approaching (%)	Target (%)	Advanced (%)	Target/Advanced (%)
<b>Elementary School</b>						
3	435	74.71	14.25	8.51	2.53	11.04
4	1,343	65.15	19.29	10.2	5.36	15.56

<sup>22</sup> As several states allowed participation in the science assessment at all grade levels (i.e., not restricted to grade-level testing required for accountability purposes), many students participated in grade levels for which cut points were not set during standard setting. Specifically, cut points were not set at grades 3 and 7 (see Chapter VI of this manual). Students testing at these grade levels were assigned performance levels using cut points set at the next highest grade level. As DLM provided states scores for all students who participated in the science assessment, all students are included in this chapter.

Grade or Course	N	Emerging (%)	Approaching (%)	Target (%)	Advanced (%)	Target/Advanced (%)
<b>Elementary School</b>						
5	4,406	62.55	18.75	15.02	3.68	18.7
<b>Middle School</b>						
6	798	56.52	24.31	15.04	4.14	19.18
7	599	66.28	18.2	12.52	3.01	15.53
8	5,029	56.07	23.5	17.22	3.2	20.42
<b>High School</b>						
9	1,090	61.93	25.78	9.63	2.66	12.29
10	1,460	56.64	27.95	12.47	2.95	15.42
11	3,830	61.44	24.80	11.15	2.61	13.76
12	419	66.83	19.57	12.17	1.43	13.60
<b>Biology</b>	805	22.48	16.27	22.24	39.01	61.25

### VII.2.B. SUBGROUP PERFORMANCE

The distribution of students across performance levels for subgroups, including groups based on gender, race, ethnicity, and English language learner status, was determined in order to set a baseline for the evaluation of changing achievement gaps in future years.

Table 63 summarizes the disaggregated subgroup frequencies collapsed across all grades. Although states each have their own rules for minimum student counts needed to support public reporting of results, small counts are not suppressed here because results are aggregated across states and individual students cannot be identified. Rows labeled as Missing indicate the student's demographic data was not entered into the system.

Table 63. Students at Each Performance Level by Demographic Group (n = 20,214)

Demographic Group	Emerging		Approaching		Target		Advanced	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
<b>Gender</b>								
Female	4,349	61.06	1,564	21.96	923	12.96	286	4.02
Male	7,590	57.98	2,920	22.31	1,918	14.65	662	5.06
Missing	1	50.00	0	0	0	0	1	50.00
<b>Race</b>								

Demographic Group	Emerging		Approaching		Target		Advanced	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
White	7,531	57.71	2,962	22.7	1,907	14.61	649	4.97
African American	2,616	60.35	976	22.51	580	13.38	163	3.76
Alaska Native	22	57.89	6	15.79	5	13.16	5	13.16
American Indian	297	45.98	131	20.28	139	21.52	79	12.23
Asian	393	74.71	88	16.73	39	7.41	6	1.14
Native Hawaiian or Pacific islander	22	50.00	15	34.09	4	9.09	3	6.82
Two or More Races	1,009	68.04	284	19.15	149	10.05	41	2.76
Missing	50	53.76	22	23.66	18	19.35	3	3.23
<b>Hispanic Ethnicity</b>								
No	10,401	58.28	4,000	22.41	2,581	14.46	865	4.85
Yes	1,498	65.13	473	20.57	249	10.83	80	3.48
Missing	41	61.19	11	16.42	11	16.42	4	5.97
<b>English Language Learner (ELL) Participation</b>								
Not ELL eligible or monitored	155	42.23	89	24.25	94	25.61	29	7.9
ELL eligible or monitored	104	65.00	28	17.5	18	11.25	10	6.25
Missing	11,681	59.33	4,367	22.18	2,729	13.86	910	4.62

### 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 0, 1 (Initial), 2 (Initial and Precursor), or 3 (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 number of students who showed no evidence of mastery, Initial level mastery, Precursor level mastery and Target level mastery (as the highest level of mastery) was each 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 64 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 3<sup>rd</sup> grade EEs, 38% of the time the highest level students mastered was the Initial level. The percentage of students who mastered as high as the Target linkage level ranged from approximately 8% in 12<sup>th</sup> grade to 29% in biology.

Table 64. *Percentage of Students Demonstrating Highest Linkage Level Mastered Across EEs, by Grade/Course*

Grade or Course	Linkage Level			
	No Evidence (%)	Initial (%)	Precursor (%)	Target (%)
<b>Elementary</b>				
3	46.3	37.5	6.9	9.3
4	41.0	38.1	8.8	12.1
5	34.2	38.8	9.2	17.8
<b>Middle</b>				
6	34.3	40.1	12.7	12.9
7	37.2	37.7	11.4	13.6
8	31.6	39.2	13.1	16.1
<b>High</b>				
9	40.4	39.4	8.4	11.8
10	37.6	39.4	9.2	13.8
11	39.0	39.8	8.6	12.6
12	55.5	30.8	5.5	8.2
<b>Biology</b>	26.2	37.0	7.6	29.3

### VII.3. DATA FILES

Three data files, made available to DLM states, summarize results from the 2015–2016 year. The General Research File (GRF) contains student results, including each student’s highest linkage level mastered for each EE and final performance level for the subject. The Special Circumstances File provides information about EEs that were affected by extenuating circumstances for each student, (e.g. chronic absences), as defined by each state. Finally, the Incident File lists students who were affected by one of the known problems with testlet assignments during the spring 2016 window.

The GRF, the Special Circumstances File, and the Incident File organize information into columns with student records in rows. If combined, the number of columns is too large for some software to read. Therefore, the GRF and supplemental files are provided separately and follow different structures. The structures for each of these files are located on the online scoring

and reporting resource page. For more information, see the *Data Dictionary* (Dynamic Learning Maps Consortium, 2016a; in Appendix F).

A sample GRF with 10 fictitious records was provided to DLM state members during the 2015–2016 year to assist in the preparation of software and data systems within each state. *A Guide to Scoring and Reports* was also provided (Dynamic Learning Maps Consortium, 2015a; see Appendix F). The structures of the GRF and supplemental files were also discussed on several partner calls to orient state members to their formats.

The GRF also contained an “Invalidation Code” field that was used during the state review process. After the GRF and supplemental files were created, states were given the opportunity to review the GRF and make changes or invalidate records. States were able to make changes to demographic data in the GRF to ensure accuracy of demographic data displayed in the score reports. States were not able to make changes to any of the EE or final performance level scores. Additionally, states used the supplemental files to determine if an entire student record should be invalidated. This was done to allow states the ability to remove students who should not be included for specific reporting or accountability purposes. These decisions were made by states based on their own state policies and procedures. By invalidating a record and resubmitting the GRF to DLM staff, the student did not receive an individual student score report and was excluded from aggregated reports.

## **VII.4. SCORE REPORTS**

Assessment results were provided to all DLM member states to be reported to parents/guardians and to educators in state and local education agencies. Reports are intended to represent what students know and can do. Performance level descriptors provide useful information about student achievement, allowing inferences regarding student achievement, progress, and growth to be drawn at the domain level. Assessment scores provide information that can be used to guide instructional decisions. Individual student reports were provided to educators and parents/guardians. Several aggregated reports (classroom, school, district, and state) were provided to state and local education agencies.

### **VII.4.A. INDIVIDUAL REPORTS**

Individual student score reports for DLM English language arts and mathematics were originally developed through a series of focus groups conducted in partnership with the Arc, a community-based organization advocating for and serving people with intellectual and developmental disabilities and their families (*2014–2015 Technical Manual—Year-End*, 2016). First, several groups focused on parent/guardian perceptions of existing alternate assessment results and score reports (Nitsch, 2013). These findings informed the development of prototype DLM score reports. Prototypes were reviewed by state partners and revised based on multiple rounds of feedback. Refined prototypes were shared with parents/guardians, advocates, and educators through additional focus groups (Clark, Karvonen, Kingston, Anderson, & Wells-Moreaux, 2015) before finalizing the 2015 reports. Science student score reports followed the same template as the English language arts and mathematics reports.



Individual student score reports comprise (1) the Performance Profile, which aggregates linkage level mastery information for reporting on each science domain and for the subject overall, and (2) the Learning Profile, which reports specific linkage levels mastered for each assessed EE<sup>23</sup>. There is one individual student score report per student per subject.

The performance levels reported on the Performance Profile are Emerging, Approaching the Target, At Target, and Advanced. These labels, which reflect a student's overall performance, were determined through a standard-setting process in spring 2016, as discussed in Chapter VI. The Performance Profile also reports the percentage of skills, or linkage levels, the student mastered within each science domain. Bulleted lists of the skills mastered follow the results reported for the science domain. The Learning Profile shows each EE separated into the three linkage levels: Initial, Precursor, and Target. A key is provided to indicate which levels the student has mastered and not mastered. A sample excerpt of the Performance Profile part of the individual student score report is provided in Figure 40. Complete sample reports are provided in Appendix F.

---

<sup>23</sup> Only states that follow the integrated assessment model for DLM English language arts and mathematics (i.e. Iowa, Kansas and Missouri) receive the Learning Profile in all three subject areas. Year-end states (i.e., Illinois, Mississippi, Oklahoma, West Virginia, Wisconsin) requested this information be omitted for science to be consistent with their ELA and mathematics reports.

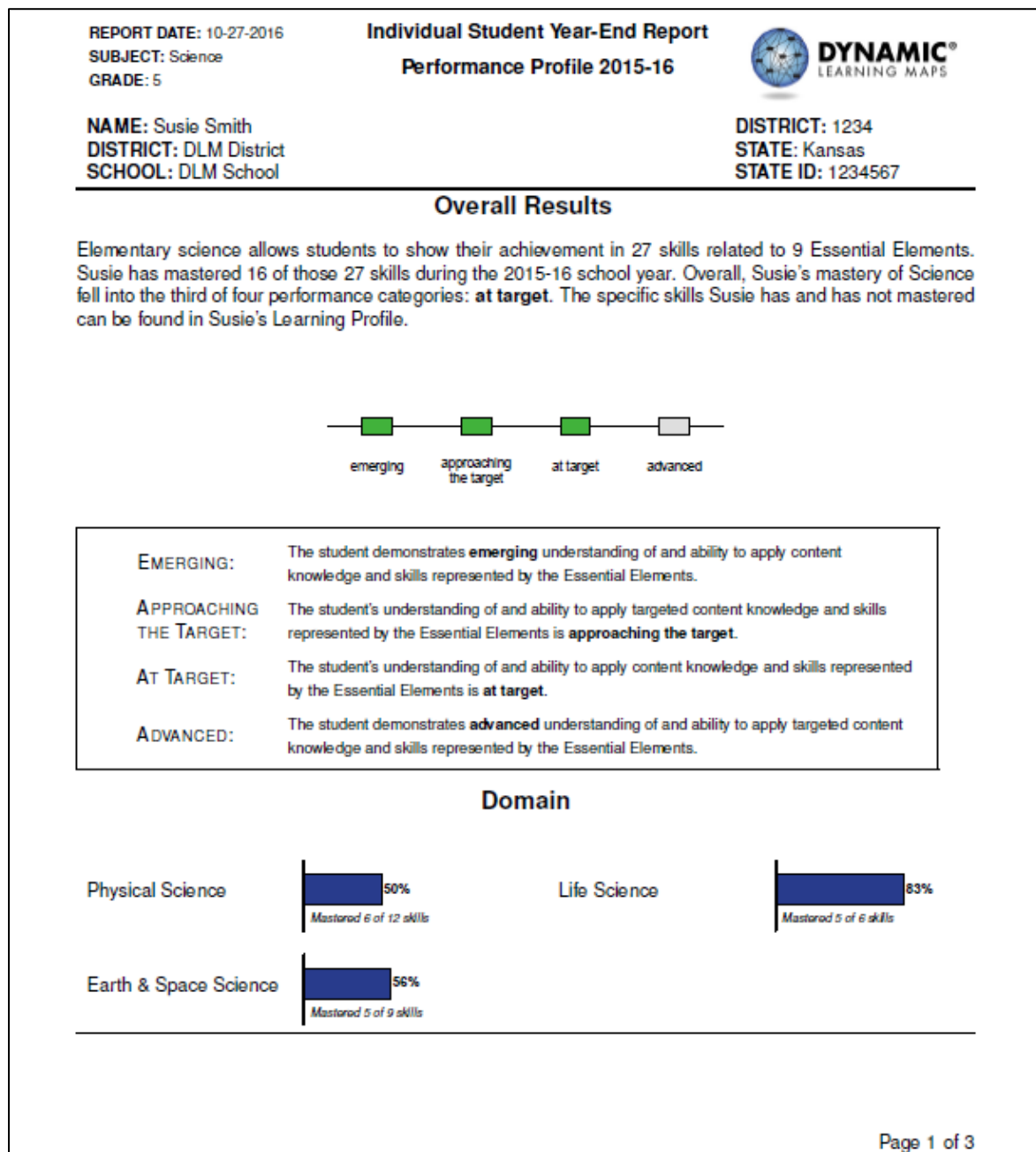


Figure 40. Page one of the performance profile for 2015-2016.

#### VII.4.B. AGGREGATED REPORTS

Student results are also aggregated into several other types of reports. At the classroom and school levels, roster reports list individual students with the number of EEs assessed, number of linkage levels mastered, and final performance level. District- and state-level reports provide frequency distributions, by grade level and overall, of students assessed and achieving at each performance level in science. Sample aggregated reports are provided in Appendix F.

### **VII.4.C. INTERPRETATION RESOURCES**

To support score interpretation and use, multiple supports were provided to aid score report interpretation:

- *The Parent Interpretive Guide* was designed to provide definition and context to student score reports (Dynamic Learning Maps Consortium, 2015b).
- Parent/guardian letter templates were developed within the DLM Consortium to be used by educators and state superintendents to introduce the student reports to parents/guardians.
- *The Teacher Interpretive Guide* was designed to support educators' discussions and build understanding for parents/guardians and other stakeholders (Dynamic Learning Maps Consortium, 2015d).
- *The Scoring and Reporting Guide for Administrators* targeted building and district-level administrators (Dynamic Learning Maps Consortium, 2015c).
- All of the resources listed above were compiled on a webpage, "Scoring and Reporting Resources" (<http://dynamiclearningmaps.org/srr/ye>). This page also contained an overview of scoring, score-report delivery, and data files. The overview was intended for state education agency staff who would be receiving DLM assessment results but did not have a lot of familiarity with the assessment. Finally, the resources page hosted score report prototypes for individual score reports and class, school, district, and state aggregated reports.

#### **VII.4.C.i. Parent Interpretive Guide**

*The Parent Interpretive Guide* (Dynamic Learning Maps Consortium, 2015b; see Appendix F) uses a sample individual student report and text boxes to explain that the assessment measures student performance on alternate achievement standards for students with the most significant cognitive disabilities—the DLM EEs. The guide goes on to describe how EEs detail what the individual student should know and be able to do at a particular grade level. In addition, the guide clarifies that students took assessments in science and that the report describes how the student performed on the assessment.

Because the Performance Profile section reports overall results in terms of the four performance levels, the sample report explains these performance level descriptions. The sample report clarifies that *At Target* means the student has met the alternate achievement standards in a given subject area at grade level. The Performance Profile goes on to define science domains and relates the student performance to those domains. Finally, the Performance Profile describes specific academic skills that the student demonstrated on the assessment within the context of grade-level academic content.

The sample report also provides additional information about the Learning Profile. The sample report shows how this section identifies what the student can do to build on the skills and knowledge demonstrated in the assessment and progress toward more complex grade-level

skills. The Learning Profile uses colored shading to illustrate which skills the student mastered and which skills were assessed but not mastered. Finally, the sample Learning Profile clarifies the target for performance using a bull's eye symbol to mark the Target performance level.

#### **VII.4.C.ii. Parent Letters**

The DLM Consortium developed templates for explanatory letters that educators and chief state school officers could use to introduce parents/guardians to the student reports (see Appendix F). These letters provide context for the reports, including what the DLM assessment is, when it was administered, and what results tell about student performance.

The letter from the chief state school officer emphasizes that setting challenging and achievable academic goals for each student is the foundation for a successful and productive school year. The letter acknowledges that students have additional goals that parents/guardians and the students' Individualized Education Program teams have established.

#### **VII.4.C.iii. Teacher Interpretive Guide**

An interpretive guide was provided for educators who would discuss results with parents/guardians or other stakeholders. The *Parent Interpretive Guide 2015-16* (Dynamic Learning Maps Consortium, 2015b), walks educators through directions for getting ready for a parent/guardian meeting, discussing the score report, and finding additional information. See Appendix F for the complete guide.

#### **VII.4.C.iv. Scoring and Reporting Guide for Administrators**

The guide designed for principals and district administrators covers each type of report provided for DLM assessments and explains how reports are distributed. The guide explains the contents of each report and provides hints about interpretation. See Appendix F for the complete guide.

### **VII.4.D. QUALITY CONTROL PROCEDURES FOR DATA FILES AND SCORE REPORTS**

Quality control procedures were implemented for all three data file types. To ensure that formatting and the order of columns were identical, column names in each file were compared with the data dictionary that was provided to states. Additional file-specific checks were conducted to ensure accuracy of all data files.

#### **VII.4.D.i. Quality Control Audit**

An audit of the quality control processes was held on March 25, 2016. Attendees included DLM psychometric staff; the director of the DLM program; the director of the Center for Educational Testing and Evaluation (CETE), which houses the DLM program; CETE psychometric staff; and the director of the Achievement and Assessment Institute, which houses CETE. Process documentation was created to ensure that established quality control procedures were clearly outlined and easily comprehensible. The audit meeting concluded that the quality control procedures currently in place were acceptable, though several enhancements were suggested

for the 2015–2016 reporting cycle. Changes suggested and implemented included creation of automated checks using the R programming language, use of networked workstations to coordinate score report generation and review, and the addition of reasonableness checks to ensure that data retrieved from the database did not contain any unexpected values.

#### **VII.4.D.ii. Automated Quality Control Checks**

To allow quality control checks to be performed more rapidly and efficiently, R programs were developed to perform quality control procedures on the GRF and on student score reports.

##### ***VII.4.D.ii.a GRF Automated Quality Control Program***

The first program written to perform automated checks was designed to perform quality control on the GRF. This program conducts a series of checks that can be organized into four main steps:

1. Check the data for reasonableness (checks detailed below).
2. Ensure that the number of linkage levels mastered for each student is less than or equal to the maximum possible value for that grade and subject.
3. Check all EE scores against the original scoring file.
4. Verify that students participating in EOI biology assessments are displayed with one row per course.

The automated program checks each row of data in the GRF and outputs any errors for review by the psychometric team.

The reasonableness checks ensure that the GRF column names accurately match the data dictionary provided to state partners and additionally check the following columns to ensure that data matches defined parameters: Student ID, State Student Identifier, Current Grade Level, Course, Student Legal First Name, Student Legal Middle Name, Student Legal Last Name, Generation Code, Username, First Language, Date of Birth, Gender, Comprehensive Race, Hispanic Ethnicity, Primary Disability Code, English for Speakers of Other Languages (ESOL) Participation Code, School Entry Date, District Entry Date, State Entry Date, State, District Code, District, School Code, School, Educator First Name, Educator Last Name, Educator Username, and Final Science Band. If invalid values are found, they are corrected as necessary by DLM staff or state partners during their 2-week review period.

##### ***VII.4.D.ii.b Student Score Reports Automated Quality Control Program***

An automated program was developed to support manual review of individual student score reports. The program was written to check key values used to generate the individual student score reports. As the score reporting program generates reports, it creates a “proofreader” file containing the values that are used to create each score report. These values are then checked against the GRF to ensure that they are being accurately populated into score reports.

Demographic values including student name, school, district, grade level, state, and state student identifier are checked to ensure a precise match. Values of skills mastered, performance level, domains tested and mastered, and EEs mastered and tested are also checked to ensure the correct values are populated, and values referring to the total number of skills, EEs, or domains available are checked to ensure they are the correct value for that grade, subject, and content area.

#### **VII.4.D.iii. Manual Quality Control Checks**

Upon its creation, each state's GRF was checked against a variety of sources to ensure that the information provided was accurate and complete. Each state's GRF was also checked to ensure it only included students belonging to that specific state. Values in the EE columns of the GRF were compared against the original scoring file to ensure their accuracy, and performance levels were recalculated and compared to ensure their accuracy.

Manual quality control checks were also performed. Given the large number of score reports generated, a stratified random sample of approximately 3-5% of the score reports generated were manually checked. Stratification was based on grade and state to ensure that any potential systematic issues due to differences in blueprints or testing models were detected.

During manual quality control checks, the Performance Profile and the Learning Profile portions (if applicable) of the individual student score reports were checked for accuracy. Performance Profiles were checked to make sure the correct performance level displayed and matched with the value in the GRF. The percentage of skills mastered in the Performance Profile was compared against the GRF and the Learning Profile portion of the student score report to ensure that all three contained the same values. Additionally, the number of domains listed in the Performance Profile were compared with the blueprint. For each EE on the student's Learning Profile, the highest linkage level mastered was compared with the value for the EE in the GRF. For both the Performance Profile and Learning Profile, the number of EEs listed on the report was compared against the number listed in the blueprint for that subject and grade or course. Demographic information in the header of the Performance Profile and Learning Profile was checked to ensure that it matched values in the GRF. Formatting and text within each report was given an editorial review as well.

Aggregated reports underwent similar manual checks, including the comparison of header information to GRF data and verification that all students rostered to an educator or school (for class and school reports, respectively) were present and that no extraneous students were listed. Performance levels (for class and school reports) and the number and percentage of students with a given performance level (for district and state reports) were checked against the same corresponding numbers or aggregated numbers from the GRF.

Once all reports were checked, all files to be disseminated to states underwent a final set of checks to ensure that all files were present. This last set of checks involved higher level assurances that the correct number of district files were present for each type of report according to the expected number calculated from the GRF, that file naming conventions were



followed, that all types of data files were present, and that all student reports had been generated.

All issues identified during quality control checks were corrected prior to distribution of data files and score reports to states.

## **VII.5. CONCLUSION**

The 2015–2016 spring science assessments were administered to a total of 20,214 students across 8 states and two Bureau of Indian Education-affiliated districts. With the exception of End-of-Instruction biology, fewer than 20% of students per grade achieved at the Target or Advanced levels. As this was the first year of the assessment and many states were still transitioning to instruction aligned with the DLM EEs, results were consistent with what states anticipated. States are provided with individual student reports and several types of aggregated reports, as well as data files to support states' own use of assessment results for accountability purposes.



## VIII. RELIABILITY

The Dynamic Learning Maps (DLM) Alternate Assessment System uses non-traditional psychometric models (diagnostic classification models) to produce student score reports. As such, evidence for the reliability of scores<sup>24</sup> is based on methods that are commensurate with the models used to produce score reports. As details on modeling are found in Chapter V, this chapter discusses the methods used to estimate reliability, the factors that are likely to affect the variability in reliability results, and an overall summary of reliability results.

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

The procedures used to assemble reliability evidence align with all applicable standards. Information about alignment with individual standards is provided throughout this chapter.

### VIII.1. BACKGROUND INFORMATION ON RELIABILITY METHODS

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

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

---

<sup>24</sup> The term results is typically used in place of scores to highlight the fact that DLM assessment results are not based on scale scores. For ease of reading, the term score is used in this chapter.

result, inherent in any discussion of reliability is the fact that the metric of reliability is accurate to the extent the assumptions of the test are met.

As the DLM Alternate Assessment System uses a different type of psychometric approach than is commonly found in contemporary testing programs, the reliability evidence reported may, at first, look different from that reported when test scores are produced using traditional psychometric techniques such as classical test theory or item response theory. Consistent with traditional reliability approaches, however, is the meaning of all indices reported for DLM assessments: When a test is perfectly reliable (i.e., it has an index value of 1), any variation in test scores comes from individual differences in the trait within the sample in which the test was administered. When a test has zero reliability, then any variation in test scores comes solely from random error.

As the name suggests, diagnostic classification models (DCMs) are models that produce classifications as probability estimates for student test takers. For the DLM system, the classification estimates are based on the set of content strands, alternate achievement standards, and levels within standards in which each student was tested. In DLM science, the standards are called Essential Elements (EEs), which are categorized into one of three domains: physical science, life science and Earth and space science. Each EE is divided into three linkage levels of complexity: Initial (I), Precursor (P), and Target (T).

For each linkage level embedded within each EE, DLM testlets were written with items measuring the listed linkage-level. Students took one testlet at a single linkage level within an EE. Therefore, a linkage-level scoring model was used to estimate examinee proficiency (See Chapter V for more information).

The DCMs used in psychometric analyses of student test data produced student-level posterior probabilities for each linkage level for which a student was tested, with a threshold of 0.8 specified for demonstrating mastery (See Chapter VI). To guard against the model being overly influential, two additional scoring rules were applied. Students could additionally demonstrate mastery by providing correct responses to at least 80% of the items measuring the EE and linkage level. Furthermore, because students often did not test at more than one linkage level within an EE, students who did not meet mastery status for any tested linkage level were assigned mastery status for the linkage level two levels below the lowest level in which they were tested (unless the lowest level tested was either the I or P levels, in which case students were considered non-masters of all linkage levels within the EE).

DLM score reports display linkage level mastery for each EE.<sup>25</sup> Linkage level results are also aggregated for EEs within each domain. Score reports also summarize overall performance in science with a performance level classification. The classification is determined by summing all linkage levels mastered and comparing the value with cut points established during standard

---

<sup>25</sup> Only displays on score reports issued to states participating in the DLM ELA and mathematics integrated model assessment program. Year-end states requested this information be omitted for science to be consistent with their ELA and mathematics reports.

setting. For more information on cut points, see Chapter VI. For more information on score reports, see Chapter VII.

Consistent with the levels at which DLM results are reported, this chapter provides six types of reliability evidence: (a) classification to overall performance level (performance level reliability); (b) the total number of linkage levels mastered for the content area (content-area reliability); (c) the number of linkage levels mastered within each domain (domain reliability); (d) the number of linkage levels mastered within each EE (EE reliability); (e) the classification accuracy of each linkage level within each EE (linkage-level reliability); and (f) classification accuracy summarized for the three linkage levels (conditional evidence by linkage level). As described in the next section, reliability evidence comes from simulation studies in which model-specific test data are generated for students with known levels of attribute mastery.

Each type of reliability evidence provides various correlation coefficients. Correlation estimates mirror estimates of reliability from contemporary measures such as the Guttman-Cronbach Alpha. For performance level and EE reliability, the polychoric correlation estimates the relationship between two ordinal variables: true performance level or number of linkage levels mastered and estimated value. For content-area reliability and domain reliability, the Pearson correlation estimates the relationship between the true and estimated numbers of linkage levels mastered. Finally, for linkage-level and conditional evidence by linkage level reliability, the tetrachoric correlation estimates the relationship between true and estimated linkage-level mastery statuses. The tetrachoric correlation is a special case of the polychoric in which the variables are discrete. Both the polychoric and tetrachoric correlations provide more accurate estimates of relationships between ordinal and discrete variables that would otherwise be attenuated using the traditional correlation (i.e., the Pearson coefficient).

Each type of reliability evidence produces correct classification rates (raw and chance corrected), which indicate the proportion of estimated classifications that match true classifications. The chance-corrected classification rate is labeled kappa and represents the proportion of error reduced above chance. Values of kappa above .6 indicate substantial-to-perfect agreement between estimated and true values (Landis & Koch, 1977).

With the classification methods of DCMs based on discrete statuses of an examinee, reliability-estimation methods based on item response theory estimates of ability are not applicable. In particular, standard errors of measurement (inversely related to reliability) that are conditional on a continuous trait are based on the calculation of Fisher's information, which involves taking the second derivative-model likelihood function with respect to the latent trait. When classifications are the latent traits; however, the likelihood is not a smooth function regarding levels of the trait and therefore cannot be differentiated (e.g., Henson & Douglas, 2005; Templin & Bradshaw, 2013). In other words, because diagnostic classification modeling does not produce a total score or scale score, traditional methods of calculating conditional standard errors of measurement are not appropriate. Rather, an alternative method is presented whereby reliability evidence is summarized for each linkage level. Since linkage levels are intended to represent varying levels of knowledge, skills, and understanding, reliability provided at each level is analogous to conditional reliability evidence.

### **VIII.1.A. 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).

Because the DLM psychometric model produces complex mastery results summarized at multiple levels of reporting (performance level, content area, domain, EE, and linkage levels) rather than a traditional raw or scaled score value, methods for evaluating reliability were based on simulation. Simulation has a long history of use in deriving reliability evidence. Large testing programs such as the Graduate Record Examination report reliability results based on simulation (e.g., Educational Testing Service, 2016). With respect to DCMs, simulation-based reliability has been used in a number of studies (e.g., Bradshaw, Izsák, Templin, & Jacobson, 2014; Templin & Bradshaw, 2013). For a simulation-based method of computing reliability, the approach is to generate simulated examinees with known characteristics, simulate test data using calibrated-model parameters, score the test data using calibrated-model parameters, and finally compare estimated examinee characteristics with those characteristics known to be true in the simulation. For DLM assessments, the known characteristics of the simulated examinees are the set of linkage levels the examinee has mastered and not mastered.

The use of simulation is necessitated by two factors: the assessment blueprint and the classification-based results that such administrations give. The method provides results consistent with classical reliability metrics in that perfect reliability is evidenced by consistency in classification, and zero reliability is evidenced by a lack of classification consistency. In the end, reliability simulation replicates DLM versions of scores from actual examinees based upon the actual set of items each examinee has taken. Therefore, this simulation provides a replication of the administered items for the examinees. Because the simulation is based on a replication of the exact same items that were administered to examinees, the two administrations are perfectly parallel. However, the use of simulation produces approximate estimates of reliability, which are contingent on the accuracy of the current scoring model.

#### **VIII.1.A.i. Reliability Sampling Procedure**

The simulation design that was used to obtain reliability estimates developed a resampling design to mirror the trends existing in the DLM assessment data. In accordance with Standard 2.1, the sampling design used the entire set of operational testing data to generate simulated examinees. 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<sup>26</sup> 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 (described in Chapter V),<sup>27</sup> 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 41 shows steps 1-4 of the simulation process as a flow chart.

---

<sup>26</sup> Calibrated-model parameters were treated as true and fixed values for the simulation.

<sup>27</sup> 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.

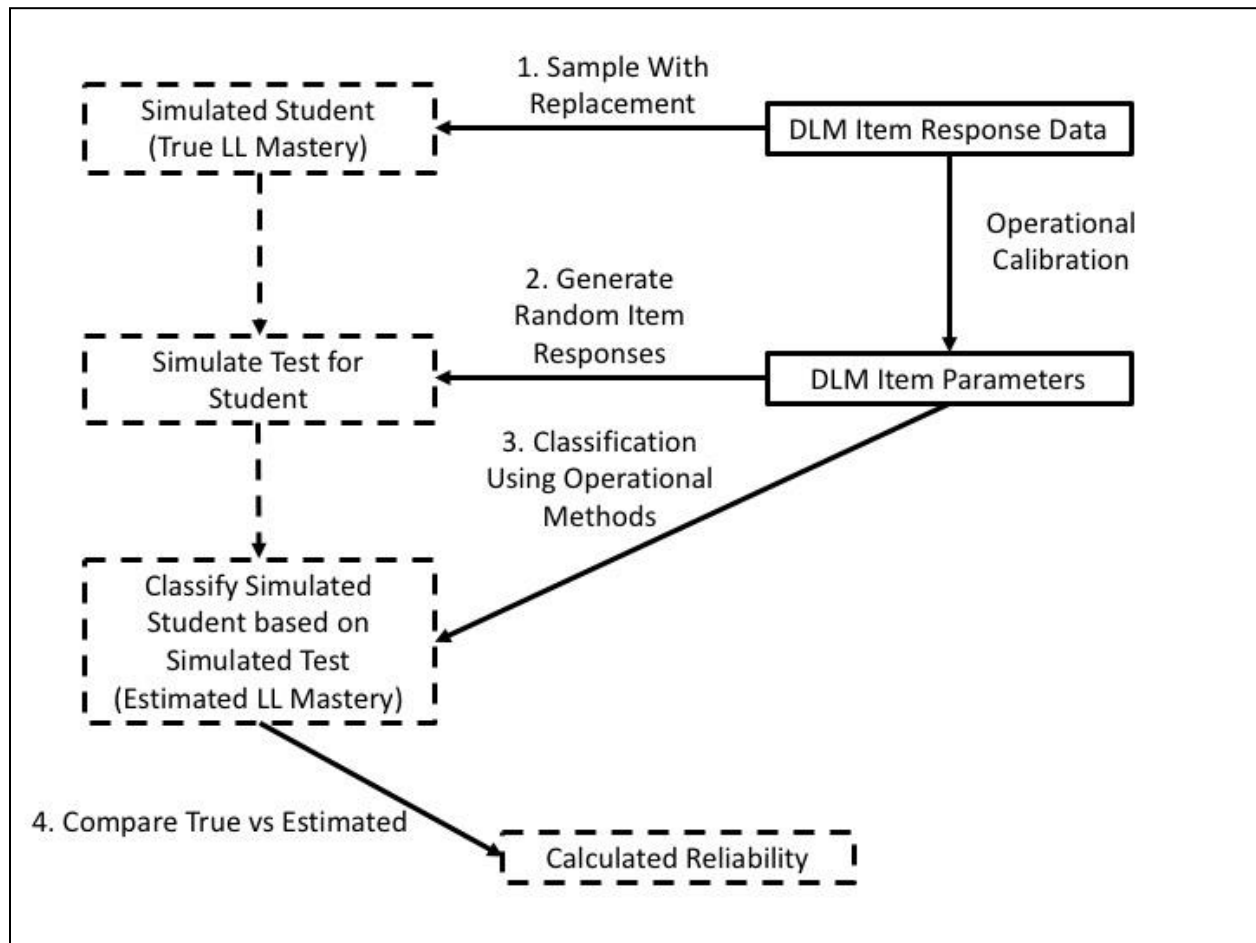


Figure 41. Simulation process for creating reliability evidence.

Note: LL = linkage level.

## VIII.2. 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, each important in the DLM testing design: (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, a total of 102 analyses were conducted to summarize reliability. Due to 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 conducted in accordance with Standard 2.2. Additionally, providing reliability evidence for each of the six levels ensures that these reliability-estimation procedures meet Standard 2.5.

### ***VIII.2.A. PERFORMANCE LEVEL RELIABILITY EVIDENCE***

Results from DLM assessments are reported using four performance levels. The total linkage levels mastered 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 at each 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 impact 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) for a given content area. Three statistics are included to provide a comprehensive summary of results. The specific metrics were chosen due to their interpretability.

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

Table 65 shows this information across all grades. Polychoric correlations between true and estimated performance levels range from .949 to .975. Correct classification rates range from 0.785 to 0.888 and Cohen’s Kappa values are between 0.840 and 0.914. 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 65. Summary of Performance Level Reliability Evidence

Grade/Course	Polychoric Correlation	Correct Classification Rate	Cohen's Kappa
3	0.970	0.888	0.892
4	0.965	0.834	0.882
5	0.969	0.852	0.881
6	0.949	0.800	0.849
7	0.960	0.841	0.857
8	0.951	0.800	0.841
9	0.954	0.828	0.840
10	0.955	0.823	0.845
11	0.956	0.830	0.844
12	0.975	0.885	0.884
Biology	0.975	0.785	0.914

### ***VIII.2.B. 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 tests to describe content-area performance. That is, the number of linkage levels mastered within a content area can be thought of as being 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 number 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 number 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 66 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 66 also meet Standard 2.19.

Table 66. Summary of Content Area Reliability Evidence

<b>Grade/Course</b>	<b>Linkage Levels Mastered Correlation</b>	<b>Average Student Correct Classification</b>	<b>Average Student Cohen's Kappa</b>
3	0.948	0.986	0.973
4	0.948	0.981	0.960
5	0.952	0.978	0.954
6	0.939	0.978	0.957
7	0.951	0.981	0.963
8	0.940	0.976	0.952
9	0.944	0.984	0.969
10	0.944	0.982	0.965
11	0.945	0.984	0.969
12	0.953	0.987	0.975
Biology	0.961	0.972	0.939

From the table, it is evident that content-area reliability, as demonstrated by the correlation between true and estimated number of linkage levels mastered, ranges from .939 to .961. These values indicate the DLM scoring procedure of reporting the number of linkage levels mastered provides reliable results of student performance.

### ***VIII.2.C. DOMAIN RELIABILITY EVIDENCE***

Within the content area of science, students are assessed on EEs within three domains. Because individual student score reports summarize the number and percentage of linkage levels students mastered for each science domain (see Chapter VII for more information), reliability evidence is provided for each.

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 are the same as described for content-area reliability.

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

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

Table 67 shows the three summary values for each domain, by grade and course. Values range from 0.604<sup>28</sup> to 0.999, indicating that overall the DLM method of reporting the total and percentage of linkage levels mastered by domain results in values that can be reliably reproduced.

Table 67. Summary of Science Domain Reliability Evidence

<b>Grade/Course</b>	<b>Domain</b>	<b>Linkage Levels Mastered Correlation</b>	<b>Average Student Correct Classification</b>	<b>Average Student Cohen's Kappa</b>
3	ESS	0.791	0.996	0.995
3	LS	0.645	0.997	0.996
3	PS	0.918	0.994	0.990
4	ESS	0.798	0.995	0.994
4	LS	0.614	0.996	0.994
4	PS	0.919	0.993	0.989
5	ESS	0.804	0.994	0.992
5	LS	0.604	0.995	0.994
5	PS	0.922	0.993	0.989
6	ESS	0.845	0.995	0.994
6	LS	0.757	0.993	0.989
6	PS	0.879	0.995	0.993
7	ESS	0.861	0.996	0.995
7	LS	0.776	0.993	0.990

<sup>28</sup> The EEs in the 0.6 range were all within the life science domain at the elementary grade band. The test development team will evaluate the items to determine if changes are needed.

Grade/Course	Domain	Linkage Levels Mastered Correlation	Average Student Correct Classification	Average Student Cohen's Kappa
7	PS	0.871	0.995	0.993
8	ESS	0.842	0.995	0.993
8	LS	0.757	0.992	0.988
8	PS	0.878	0.995	0.993
9	ESS	0.821	0.994	0.992
9	LS	0.782	0.994	0.992
9	PS	0.859	0.996	0.995
10	ESS	0.826	0.994	0.992
10	LS	0.787	0.994	0.992
10	PS	0.871	0.995	0.994
11	ESS	0.832	0.994	0.992
11	LS	0.797	0.994	0.992
11	PS	0.863	0.996	0.994
12	ESS	0.824	0.995	0.994
12	LS	0.834	0.996	0.994
12	PS	0.898	0.996	0.995
Biology	SCI.LS1.A	0.836	0.992	0.989
Biology	SCI.LS1.B	0.999	0.999	0.999
Biology	SCI.LS2.A	0.740	0.996	0.995
Biology	SCI.LS3.B	0.999	0.999	0.999
Biology	SCI.LS4.C	0.892	0.995	0.994

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

#### ***VIII.2.D. 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, the highest linkage level mastered per EE is examined, rather than the whole content area. 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.

The following 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 form. Table 68 and Figure 42 provide proportions and the number of EEs, respectively, falling within pre-specified 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 68. Reliability Summaries Across All 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	0.000	0.000	0.000	0.088	0.147	0.118	0.382	0.147	0.118
Correct Classification Rate	0.000	0.000	0.000	0.000	0.000	0.294	0.559	0.118	0.029
Kappa	0.000	0.118	0.088	0.176	0.206	0.176	0.088	0.147	0.000

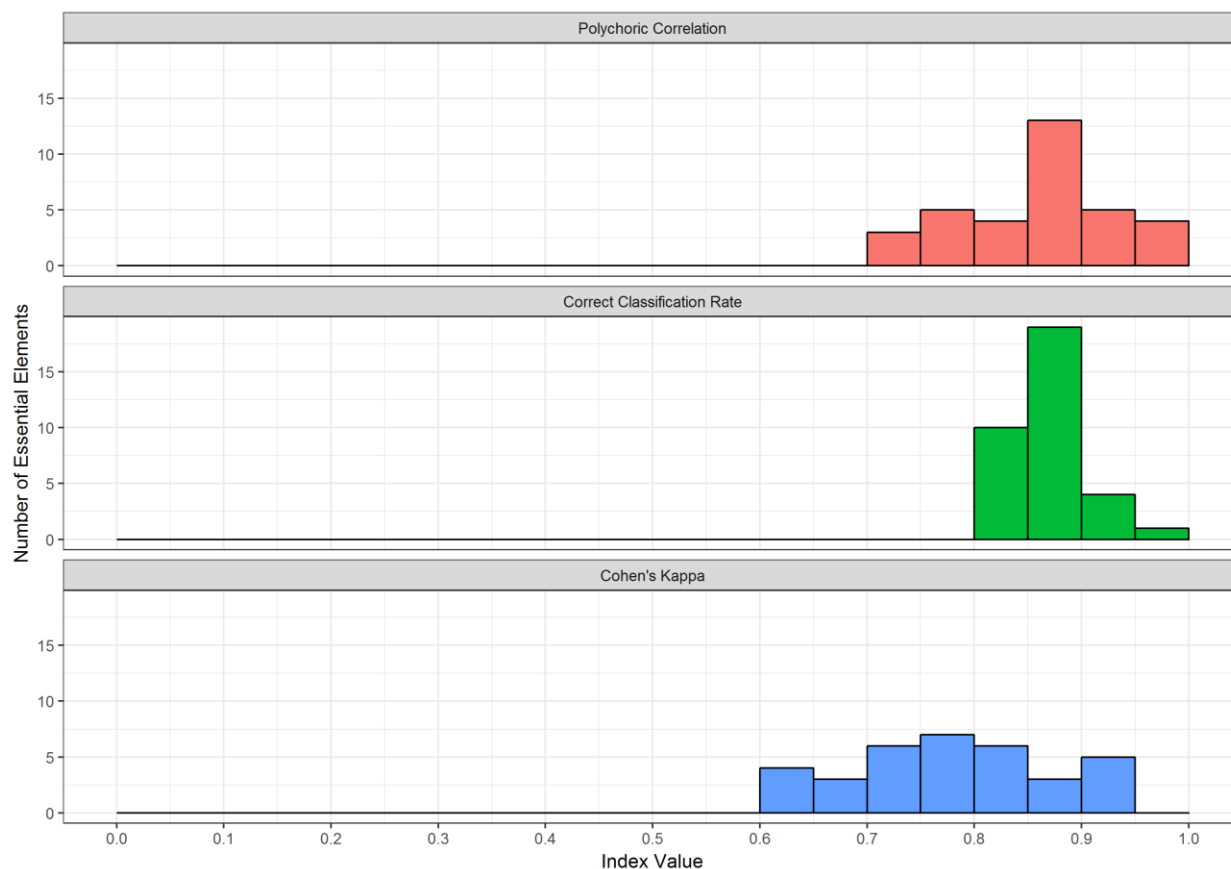


Figure 42. Number of linkage levels mastered within EE reliability summaries.

### VIII.2.E. 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.<sup>29</sup> 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 one example, Table 69 shows an example table of simulated results for one linkage level.

<sup>29</sup> The linkage level reliability evidence presented here focuses on consistency of measurement given student responses to items. For more information on how students were assigned linkage levels during assessment, see Chapter 4 – Adaptive Delivery.

Table 69. Example of True and Estimated Mastery Status from Reliability Simulation

		<u>Estimated Mastery Status</u>	
		<u>Nonmaster</u>	<u>Master</u>
True Mastery Status	Nonmaster	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 status.
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.

*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 70 and Figure 43 provide proportions and number of linkage levels, respectively, that fall within pre-specified ranges of values for the three reliability summary statistics (i.e., correct classification rate, kappa, and correlation).*

The correlations and correct classification rates show reliability evidence for the classification of mastery at the linkage level. There were 13 linkage levels that had Kappa values below 0.6. Four linkage levels were at the elementary grade band, six were in middle school and three were in high school; all were either in the life science or Earth and space science domains. The test development team will evaluate the items at these linkage levels to determine if changes are needed.



Table 70. 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	0.000	0.000	0.010	0.039	0.020	0.049	0.127	0.196	0.559
Correct Classification Rate	0.000	0.000	0.000	0.000	0.000	0.010	0.235	0.520	0.235
Kappa	0.127	0.088	0.088	0.118	0.108	0.137	0.186	0.088	0.059

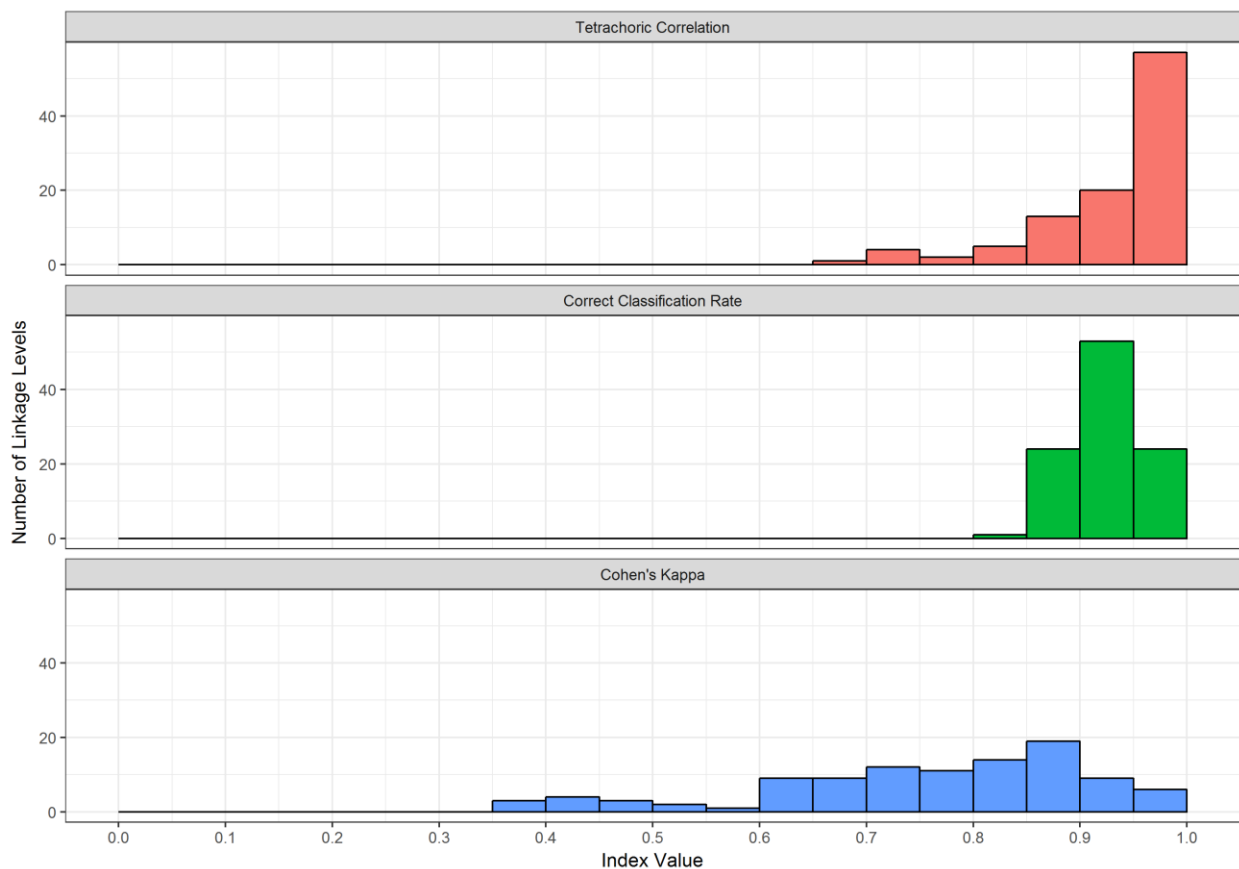


Figure 43. Linkage-level reliability summaries.

### ***VIII.2.F. 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 knowledge, skills, and understandings as defined by the three linkage levels, evidence of reliability can be summarized for each linkage level to approximate conditional evidence over all EEs, similar to a conditional standard error of measurement for a total score.

Conditional reliability evidence by linkage level is based on the true and estimated mastery statuses for each linkage level, 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 44 provides the number of linkage levels that fall within pre-specified ranges of values for the reliability summary statistics. The correlations and correct classification rates generally indicate that all three linkage levels provide reliable classifications of student mastery. Results showed that for the 13 linkage levels described in the previous section, all of the Kappa values below 0.6 were at the P and T levels. Again, the test development team will evaluate the items at these linkage levels to determine if changes are needed.

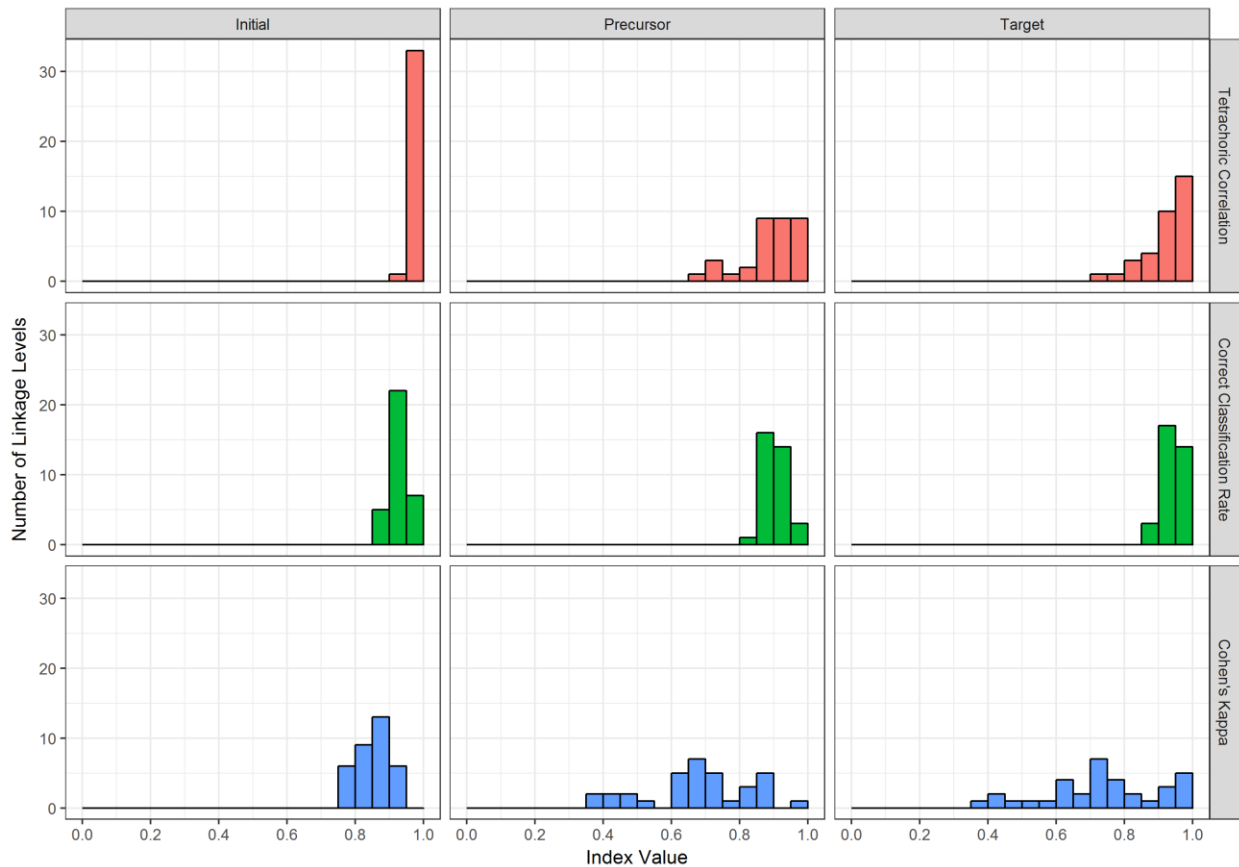


Figure 44. Conditional reliability evidence summarized by linkage level.

### 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 diagnostic classification modeling 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 would also impact 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 used in this chapter provides a replication of the exact same test items (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 provide evidence in support of the overall validity argument for scores produced by the Dynamic Learning Maps (DLM) Science Alternate Assessment System. Chapter IX presents additional evidence. The special studies presented here were conducted throughout the assessment development, administration, and evaluation processes. These studies address four of the critical sources of evidence as described in *Standards for Educational and Psychological Testing* (AERA et al., 2014): evidence based on (a) test content, (b) response processes, (c) internal structure, and (d) consequences of testing. Each study addresses assumptions related to the theory of action, specifically, related to the four propositions for score interpretation and use. These propositions and score purposes are discussed in depth in the Evaluation Summary section of Chapter XI, where the overall validity framework for the DLM Alternate Assessment System is laid out alongside evidence sources.

### 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., p. 14). The interpretation and use of DLM scores depends on the validity of the model of learning and cognition underlying the system and of the correspondence between student learning standards and items and full tests. The validity studies presented in this section focus on the alignment of test content to content standards via the DLM content maps (which underlie the assessment system) and preliminary evidence of student opportunity to learn the assessed content.

#### IX.1.A. EXTERNAL ALIGNMENT STUDY

HumRRO conducted an external alignment study on the 2015–2016 DLM science operational assessment system (Nemeth & Purl, 2017). The purpose of the study was to investigate the relationships between the content structures in the DLM Science Alternate Assessment System and assessment items. The alignment study focused on the following three relationships (see Figure 45):

1. EEs to general education content standards;
2. the vertical articulation of linkage levels for each Essential Element (EE); and
3. testlets and testlet items to linkage levels.

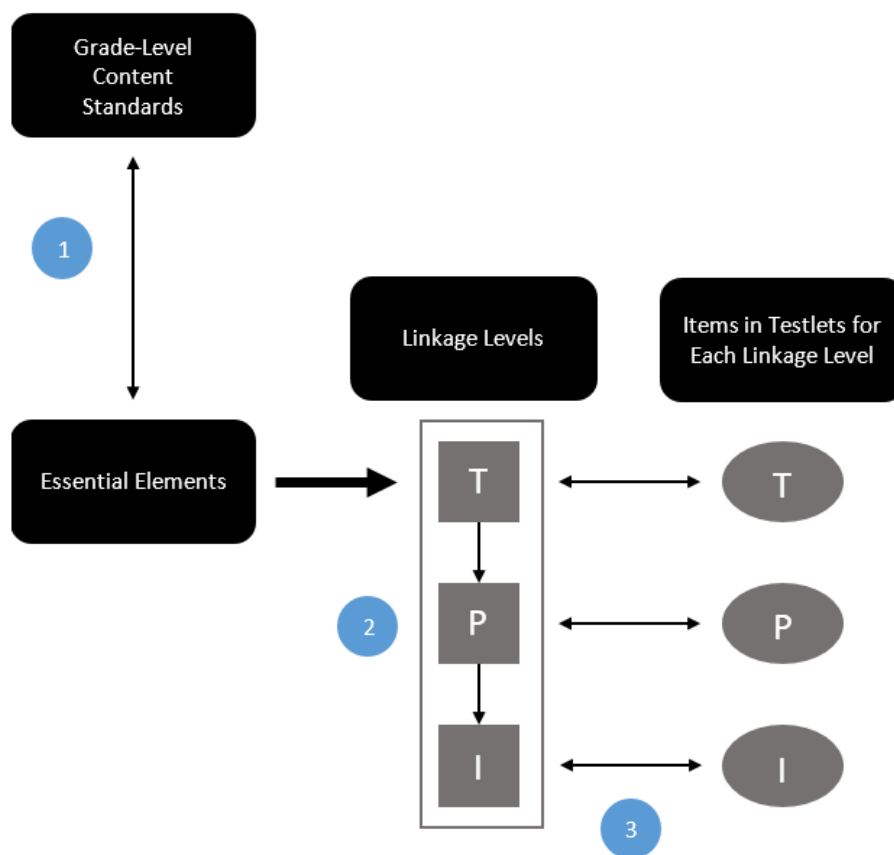


Figure 45. Design of the DLM science assessment.

*Note:* Linkage levels are Target (T), Precursor (P), and Initial (I).

Each of the three focal areas for alignment were evaluated against a set of criteria. Within the first focal area, EEs were rated as not aligned, partially aligned, or fully aligned to intended content, categories, and complexity. For the second focal area, linkage levels within each EE were rated as non-progressing or progressing in skills/knowledge and/or cognitive demand across adjacent linkage levels. Finally, within the third focal area, assessment items were rated as not aligned, partially aligned, or fully aligned to intended content, categories and complexity. In cases where panelists rated elements as not or partially aligned or non-progressing, they were asked to provide rationales for their ratings and/or suggested changes to improve alignment. For each criterion, HumRRO established a threshold for acceptable alignment.

All of the 2015-2016 operational assessment content for the DLM science assessment was examined in the study, including 31 EEs (93 linkage levels) and 288 items across 94 testlets. This includes 9 testlets (28 items) assessing 3 EEs that overlap the high school and biology end of instruction blueprints.

The following sections provide a brief summary of findings from the external alignment study. Full results are provided in the separate technical report (Nemeth & Purl, 2017). Follow-up analyses conducted by the Center for Educational Testing and Evaluation (CETE) are described after the HumRRO findings. Plans for subsequent steps are summarized in the *CETE Response to External Evaluation of DLM Science Alternate Assessment System Alignment* (Appendix G).

### **IX.1.A.i. Alignment of EEs to General Education Content Standards**

To evaluate this relationship, panelists reviewed the 34 EEs in multiple ways: (1) they evaluated the content alignment (Criterion 1) between the EEs and the Next Generation Science Standards (NGSS) content (disciplinary core ideas [DCIs] and science and engineering practices [SEP]); (2) using panelists' ratings from Criterion 1, the EEs were evaluated regarding a match to the Domain, DCI, and Topic of the corresponding NGSS; (3) panelists determined consensus cognitive process dimensions first for the NGSS and then for the EEs, independently, allowing for a comparison of the cognitive process dimensions between the standards.

All EEs identified in the test blueprints were included in these analyses. The rules for the criterion applied to the alignment between EEs and the NGSS were as follows:

- Criterion 1: 90% or more of the EE ratings were 'partially' or 'fully' aligned.
- Criterion 2: EEs match the Domain, DCI, and Topic of the corresponding NGSS.
- Criterion 3: 75% or more of the EE ratings were at the same or lower cognitive process dimension as the NGSS.

The results of applying these rules to the ratings are shown in Table 71. In general, the EE ratings reflected strong linkage with the grade-level standards ratings. Specifically, the EE ratings across all grade bands measured the intended content (Criterion 1), and represented content from the reporting categories as expected (Criterion 2). The High School Unique EE ratings aligned with associated SEP fell just below the 90% criterion; however, when the HS & Biology Common EE ratings were included, the high school EEs met the 90% criterion. More than 75% of the elementary, high school, and biology EE ratings were found to assess the same or lower cognitive process dimension as the NGSS; however, 33% of the middle school EE ratings (3 EEs) reflected a higher cognitive process dimension than the NGSS rating (Criterion 3).

Table 71. Percentage of Essential Element Ratings Which Met Each Criterion

	Criterion 1		Criterion 2	Criterion 3
	Essential Element Alignment		Represent Intended Categories	Essential Element Complexity
	Are EE ratings aligned with associated DCIs?	Are EE ratings aligned with associated SEP?	Do EEs adequately represent reporting categories?	Are EE ratings at same or lower cognitive process dimension as NGSS?
Elementary	100%	100%	100%	78%
Middle School	100%	100%	100%	67%
High School Unique	100%	87%	100%	100%
HS & Biology Common	100%	100%	100%	100%
Biology Unique	100%	100%	100%	100%

### IX.1.A.ii. Vertical Articulation of Linkage Levels for each Essential Element

The criterion applied to the vertical articulation of linkage levels was that 90% or more of the linkage level transition ratings are “progressing.” The results are provided in Table 72. Across grade band pools, the linkage level transition ratings met the criterion for ‘progressing’ between initial to precursor and precursor to target. In high school biology, 89% of the initial to precursor and 84% of the precursor to target linkage level transition ratings were ‘progressing’. However, when combined with the common EEs, the high school biology pool ratings met the 90% threshold.



Table 72. Percentage of Linkage Level Transition Ratings Which Met the Criterion

	Vertical Articulation	
	Do initial to precursor linkage level transition ratings indicate progression?	Do precursor to target linkage level transition ratings indicate progression?
Elementary	97%	100%
Middle School	94%	94%
High School Unique	100%	100%
HS & Biology Common	100%	100%
Biology Unique	89%	84%

### IX.1.A.iii. Alignment of Testlets and Items to Linkage Levels

This relationship was evaluated in multiple ways. Panelists evaluated the content alignment (Criterion 1) between the items and the testlets to the associated linkage levels. Using panelists' ratings from Criterion 1, the items and testlets were evaluated regarding a match to the EE Domain, DCI, and Topic of the corresponding linkage levels. Panelists verified the cognitive process dimensions of the items. Finally, the cognitive process dimensions of target linkage level items and the corresponding EE were compared. The rules for the criterion applied to the alignment between items or testlets and linkage levels are as follows:

- **Criterion 1: Items/Testlets Represent Intended Content**
  - 90% or more of the item ratings were 'partially' or 'fully' aligned to EE linkage level DCI
  - 90% or more of the item ratings were 'partially' or 'fully' aligned to EE linkage level SEP
  - 90% or more of testlet ratings were 'partially' or 'fully' covering the assigned EE linkage level content
- **Criterion 2: Items/Testlets Represent Intended Categories**
  - Testlets match the EE linkage level Domain, DCI, and Topic of the assigned linkage level

- **Criterion 3: Items Represent Intended Complexity**
  - 90% or more of the assigned cognitive process dimensions are confirmed by panelists' ratings for items
  - 75% or more of the target linkage level item ratings were at the same or lower cognitive process dimension as the EE

Table 73 provides a summary of conclusions on focus 3. In general, the item ratings indicated good overall alignment with the linkage levels. Panelists rated the assessment items for all grade bands as measuring the intended EE linkage level DCI, even though not all item ratings aligned with the EE linkage level SEP in middle school and high school unique (Criterion 1). Overall, testlet ratings were greater than the 90% criterion level, indicating adequate linkage level coverage across items within a testlet. Panelists found items and testlets for all grade levels to closely match the expected Domain, DCI, and Topic associated with the EE (Criterion 2). There were mixed results on Criterion 3. In all groups, panelist ratings showed agreement with more than 90% of the cognitive process dimensions assigned to items within +1/-1 cognitive process dimension. For 65% of the high school and biology common items, the cognitive process dimension was higher for the item than the associated EE.

Table 73. Percentage of Testlet Items Which Met Each Criterion

	Criterion 1			Criterion 2	Criterion 3	
	Item Alignment			Represent Intended Categories	Item Complexity	
	Are item ratings aligned with EE linkage level DCI?	Are item ratings aligned with EE linkage level SEP?	Do testlet ratings fully cover EE linkage level content?	Do testlets adequately represent intended categories?	Do panelist ratings agree with all linkage level items' cognitive process dimensions within +1/-1?	Do target linkage level items reflect lower or same cognitive process dimensions as the EEs?
Elementary	100%	90%	99%	100%	100%	89%
Middle School	100%	81%	93%	100%	97%	94%
High School Unique	100%	88%	99%	100%	98%	83%
HS & Biology Common	99%	90%	98%	100%	96%	35%

	Criterion 1			Criterion 2	Criterion 3	
	Item Alignment			Represent Intended Categories	Item Complexity	
	Are item ratings aligned with EE linkage level DCI?	Are item ratings aligned with EE linkage level SEP?	Do testlet ratings fully cover EE linkage level content?	Do testlets adequately represent intended categories?	Do panelist ratings agree with all linkage level items' cognitive process dimensions within +1/-1?	Do target linkage level items reflect lower or same cognitive process dimensions as the EEs?
Biology Unique	100%	94%	100%	100%	96%	84%

#### IX.1.A.iv. Follow-Up Analyses

Traditionally, alignment study results yield statistics about elements or relationships within an assessment system, and judgments of adequacy based on those statistics. The HumRRO external alignment report did not report conventional alignment statistics. Instead, most results were calculated and reported with individual ratings as the unit of analysis and reporting. While the HumRRO report provides useful information and indicates a high degree of alignment, the statistics incorporate rater disagreement within panels and do not provide final judgments directly about the units in the assessment system itself.

Using HumRRO's original ratings, DLM staff applied a majority rule within panels to render a final per-panel judgment on each element or relationship being evaluated. This process yielded more traditional alignment statistics, which were then evaluated against HumRRO's thresholds.

- For focus 1, all pools met all criteria, with the exception of middle school/criterion 3.
- For focus 2, all pools met the criterion.
- For focus 3, all pools met all criteria with the exception of middle school/criterion 1 SEP alignment; and elementary, high school, and biology/criterion 3.

The procedure and full results are provided in Appendix G.

Overall, the external alignment study provides evidence of the DLM Science Alternate Assessment System components that connect the general education science content standards to the assessment items, via EEs and linkage levels. The external alignment study provides substantial content-related evidence to support the DLM Consortium's claims about what students know and can do in science. Areas for further investigation and action based on the findings are addressed in Chapter XI.

### ***IX.1.B. OPPORTUNITY TO LEARN***

As part of the fall field test administration (see Chapter III for details), a survey was administered to educators in order to obtain feedback on their students' opportunity to learn science content during the 2015-2016 school year. Students were randomly selected and enrolled to participate in the survey. If a student was enrolled in the survey, the educator who was responsible for administering the science assessment would also complete the survey questions about that student. Of the 2,037 students that were enrolled in the survey, 837 had completed surveys, for a response rate of approximately 41%.

Educators were asked to indicate the average number of hours they either spent on instruction or planned for instruction on science content during the 2015-16 school year. Table 74 below presents the number and percentage of educators by average number of hours spent instructing or hours planned to instruct students on science content within ten topics. Table 5 displays the number and percentage of educators who either spent time instructing their students or planned to instruct their students in science practices during science instruction. Educators could select more than one science practice.

Overall, the majority of educators spent, on average, between one and ten hours of instruction on most science topics during the 2015-2016 school year. Approximately 40% of educators did not spend any instructional time on the topics of heredity or biological evolution. The science practice that educators engaged their students in most frequently was to ask questions and define problems, while the least frequently used practice was to engage in argument from evidence.

The Opportunity to Learn survey results suggest that there is a significant need for improvement with respect to providing students with the most significant cognitive disabilities access to science curriculum aligned with the *Framework*. With increased opportunities to learn science content and engage in scientific practices, it is anticipated that these students will be better able to demonstrate science academic skills (Andersen, Bechard, & Merriweather, 2016).

Table 74. Average Number of Hours Spent Instructing Science Topics

Science Topic	None		1-10 hours		11-20 hours		21-30 hours		More than 30 hours		Missing	
	<u>n</u>	<u>%</u>	<u>n</u>	<u>%</u>	<u>n</u>	<u>%</u>	<u>n</u>	<u>%</u>	<u>n</u>	<u>%</u>	<u>n</u>	<u>%</u>
Matter and its Interactions	166	19.8	481	57.5	119	14.2	21	2.5	37	4.4	13	1.6
Motion and Stability: Forces and Interactions	202	24.1	475	56.8	106	12.7	21	2.5	21	2.5	12	1.4
Energy	162	19.4	495	59.1	116	13.9	28	3.4	23	2.8	13	1.6
From Molecules to Organisms: Structure and Processes	239	28.6	433	51.7	112	13.4	20	2.4	19	2.3	14	1.7
Ecosystems: Interactions, Energy, and Dynamics	214	25.6	423	50.5	133	15.9	40	4.8	14	1.7	13	1.6
Heredity: Inheritance and Variation of Traits	359	42.9	366	43.7	75	9.0	13	1.6	12	1.4	12	1.4
Biological Evolution: Unity and Diversity	333	39.8	387	46.2	76	9.1	11	1.3	15	1.8	15	1.8
Earth's Place in the Universe	167	20.0	460	55.0	135	16.1	42	5.0	20	2.4	13	1.6
Earth's Systems	107	12.8	475	56.8	160	19.1	50	6.0	32	3.8	13	1.6
Earth and Human Activity	160	19.1	482	57.6	126	15.1	38	4.5	18	2.2	13	1.6

Table 75. Science Practices in Which the Student Was Instructed (N=837)

<b>Science Practice</b>	<i>n</i>	%
Asking questions and defining problems	680	81.2
Planning and carrying out investigations	497	59.4
Analyzing and interpreting data	480	57.4
Obtaining, evaluating, and communicating information	477	57.0
Developing and using models	465	55.6
Using mathematics and computational thinking	348	41.6
Constructing explanations and designing solutions	241	28.8
Engaging in argument from evidence	160	19.1

*Note.* Educators were allowed to select multiple responses.

## **IX.2. EVIDENCE BASED ON RESPONSE PROCESSES**

The study of the response processes of test takers provides evidence regarding the fit between the test construct and the nature of how students actually experience test content (AERA et al., 2014). Both theoretical and empirical evidence is appropriate and should come from both the individual test taker and external observation. The interpretation and use of DLM scores depends in part on the validation of whether the cognitive processes that students are engaged in when taking the test match the claims made about the test construct. This category of evidence includes studies on student and test administrator behaviors during testlet administration. Because testlets must be administered with fidelity in order to support the ability of students to respond based on their knowledge of the construct, evidence of fidelity is included in this section. Furthermore, plans for obtaining test administrator feedback on students' knowledge, skills, and understandings to respond to testlets during the spring 2017 administration are provided in this section.

### ***IX.2.A. EVALUATION OF TEST ADMINISTRATION***

One study was conducted and one is planned to better understand response processes and test administration procedures. Data for the first study were collected during the 2015-2016 academic year using similar procedures used for ELA and mathematics assessments. Specifically, it includes the use of a protocol to gather standardized observational evidence of test administrations, including both response processes and fidelity of administration. The second method includes gathering feedback from test administrators.

### IX.2.A.i. Observations of Test Administration

The DLM Consortium uses a test administration observation protocol to gather information about how educators in the consortium states deliver testlets to students with the most significant cognitive disabilities. This protocol gives observers a standardized way to describe the way a DLM testlet was administered—no matter their role or experience with DLM assessments. The test administration observation protocol captures data about student actions (navigation, responding, etc.), educator assistance, variations from standard administration, engagement, and barriers to engagement. Test administration observations are collected by DLM project staff, as well as state education agency and local education agency staff. The observations protocol is only used for descriptive purposes; it is not used to evaluate or coach the educator or to monitor student performance. Most items are a direct report of what was observed, for instance, how the test administrator set up for the assessment, and what the test administrator and student said and did. One section asks observers to make judgments about the student’s engagement during the session.

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

For science, 37 test administration observations were collected in 2015-16. There was one observation of a science testlet administration from a field test event in the fall of 2015 and 36 observations collected in the spring of 2016. The number of observations collected by state are shown in Table 76.

Table 76. Teacher Observations by State (N = 37)

State	<i>n</i>	%
Missouri	22	59.5
Oklahoma	15	40.5

Of the 37 science test administration observations, 29 (78%) were of computer-delivered testlets and 8 (22%) were of teacher-administered testlets. Of the 37 administrations observed, 13 (35%) were in a small room used for testing, and 24 (65%) were in the students’ typical classroom.

To investigate the assumptions that underlie the claims of the validity argument, several parts of the test-administration observation protocol corresponded to 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 77, with behaviors identified as supporting, neutral, or non-supporting. For example, clarifying directions (34.5% of observations) removes student confusion over the task demands as a source



of construct-irrelevant variance and supports the student’s meaningful, construct-related engagement with the item. In contrast, reducing the number of choices available to the student is a clear indicator that the teacher directly influenced the student’s answer choice.

Table 77. Test Administrator Actions During Computer-Delivered Testlets (N = 29)

Evidence	Action	n	%
Supporting	Used verbal prompts to direct the student’s attention	10	34.5
	Clarified directions	10	34.5
Neutral	Navigated one or more screens for the student	19	65.5
	Repeated question(s) before student responded	16	55.2
	Defined vocabulary used in the testlet	1	3.4
	Repeated question(s) after student responded	4	13.8
	Asked the student to clarify one or more responses	0	NA
Non-supporting	Used physical prompts	5	17.2
	Reduced number of choices available to student	0	NA

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 65.5% of the observations is not necessarily an indication that 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 would be counter to the assumption that students are able to interact with the system as intended. The observation protocol did not capture the reason the test administrator chose to navigate, and the reason was not always obviously inferred just 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 (see Table 78). Independent response selection was observed in 41.4% of the cases and the use of eye gaze (one unique form of independent selection that was recorded separately) was seen in 13.8% of the observations. Verbal prompts for navigation and response selection are strategies that are within the realm of allowable flexibility during test administration. These strategies, which are commonly used during direct instruction for students with the most significant cognitive disabilities, would be 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 78. Student Actions during Computer-Delivered Testlets (N = 29)

Action	n	%
Navigated the screens independently	10	34.5
Navigated the screens with verbal prompts	5	17.2
Selected answers independently	12	41.4
Selected answers with verbal prompts	12	41.4
Indicated answers using eye gaze	4	13.8
Indicated answers using materials outside of KITE Client	4	13.8
Skipped one or more items	2	6.9
Used manipulatives	2	6.9

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

Another assumption in the validity argument is that students are able to respond to tasks irrespective of a sensory, mobility, health, communication, or behavioral constraint. 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 8 observations of teacher-administered testlets, observers did not note that the student had difficulty with accessibility. For computer-delivered testlets, evidence to evaluate this assumption was observed by noting students' knowledge, skills, and understandings to indicate responses to items using multiple response modes such as sign language, eye gaze, and using manipulatives or materials outside of KITE Client. Table 78 presents a summary of the frequencies of these behaviors. Additional evidence for this assumption was gathered by observing whether students were able to complete testlets. Of the 37 test-administration observations collected, in 36 cases (97%) students completed the testlet.

Another assumption underlying the validity argument is that test administrators enter student responses with fidelity. Observers rated whether test administrators accurately captured student responses. In order to record student responses with fidelity, test administrators needed to observe multiple modes of communication, such as verbal, gesture, and eye gaze. Table 79 summarizes students' response modes for teacher-administered testlets.

Table 79. Primary Response Mode for Teacher-Administered Testlet (N = 8)

Response mode	<i>n</i>	%
Verbal	3	37.5
Gesture	3	37.5
Eye gaze	4	50
Other	0	NA
No response	0	NA

Note. Respondent could select multiple responses to this question.

Across both computer-delivered and teacher-administered observations and all student response modes, test administrators recorded responses for the student in 17 cases (46%). In all (100%) of those 17 cases, observers noted that the entered response matched the student's response. This evidence supports the assumption that test administrators entered student responses with fidelity.

Plans for collecting data in 2017 include using the revised version of the test administration protocol that contains science as a subject area and recruiting DLM state partners to use the protocol themselves and distribute it to district staff for their own observations.

### IX.2.A.ii. Test Administrator Feedback Studies

Test administrators provided feedback after administering the spring operational DLM assessments. Survey data that inform evaluations of assumptions regarding response processes include test administrator perceptions of student ability to respond as intended, free of barriers, and test administrator perceptions of the ease of administering teacher-administered testlets. Perceptions of student response come from the spring 2016 test administrator survey.<sup>30</sup>

The spring 2015 test administrator survey included three items about students' ability to respond. Test administrators were asked to rate statements from *strongly disagree* to *strongly agree*. Results are presented in Table 80.

The majority of test administrators agreed or strongly agreed that their students (1) responded to items to the best of their knowledge ability, (2) were able to respond regardless of disability, behavior, or health concerns, and (3) had access to all necessary supports to participate.

---

<sup>30</sup> Recruitment and response information for this survey was provided in Chapter IV.

Table 80. Test Administrator Perceptions of Student Experience with Assessments, Spring 2016

Statement	Strongly Disagree		Disagree		Agree		Strongly Agree	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Student responded to items to the best of his/her knowledge and ability	237	10.4	298	13.1	1171	51.4	570	25.0
Student was able to respond regardless of his/her disability, behavior, or health concerns	400	17.6	402	17.7	1113	49.1	352	15.5
Student had access to all necessary supports to participate	125	5.5	183	8.1	1315	58.0	644	28.4

### IX.3. EVIDENCE BASED ON INTERNAL STRUCTURE

Analyses that address the internal structure of an assessment 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., p. 16). Given the heterogeneous nature of the student population, statistical analyses can examine whether particular items function differently for specific subgroups of students.

#### 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). Studies that use DIF analyses can uncover internal inconsistency if particular items are functioning differently in a systematic way for identifiable subgroups of students (AERA et al., 2014). While DIF does not always indicate a weakness in the test item, it can help point to construct-irrelevant variance or unexpected multidimensionality, thereby contributing to an overall argument for validity and fairness.

##### IX.3.A.i. Method

The initial DIF analysis for items in the DLM science alternate assessment was conducted using data collected during the spring 2016 administration and procedures previously developed and applied to evaluate DIF in DLM ELA and mathematics assessments. Because 2015-2016 was the first operational year of DLM science assessments and DIF analyses were dependent upon the amount of data collected for each item, the initial DIF analyses examined only performance for male and female subgroups. As additional data is collected in subsequent operational years, the

scope of DIF analyses will be expanded to include additional items, subgroups (e.g., expressive communication skills), and approaches to detecting DIF.

Items were selected for inclusion in the initial DIF analyses based on minimum sample size requirements for the two groups. Jodoin & Gierl examined Type I error and power rates in a simulation study examining DIF detection using a logistic regression approach (2001). Two of their conditions featured a 1:2 ratio of sample size between the focal and reference groups. 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. However, it should also be noted that Type I error rates are not necessarily problematic in the DLM operational context given that DIF detection triggers content team review of items, rather than an automatic decision to eliminate the item from the operational pool.

Within the DLM population, the number of female students responding to items is smaller than the number of male students by a ratio of approximately 1:2; therefore, a threshold for item inclusion was imposed whereby the female group must have at least 100 students responding to the item. The threshold of 100 was selected to balance the need for a sufficient sample size in the focal group with the relatively low number of students responding to many DLM items. Only operational content meeting sample size thresholds was included in the initial DIF analyses.

Using the above criteria for inclusion, 300 items (95%) were selected for inclusion in the analysis. Eighty-two items were evaluated for evidence of DIF in the elementary and middle school grade bands, and 136 items were evaluated in the high school grade band.<sup>31</sup> Sample sizes were between 276 and 4,134 per item.

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 comprised of the student's total linkage levels mastered in the content area of the item and a group membership variable, with females coded zero as the focal group and males coded one as the reference group. An interaction term was included to evaluate whether non-uniform DIF was present for each item (Swaminathan & Rogers, 1990), which, when present, is indicative that the item functions differently as a result of the interaction between total linkage levels mastered and gender. Said another way, when non-uniform DIF is present, the gender group with the highest probability of a correct response to the item differs along the range of total linkage levels mastered, whereby one group is favored at the low end of the spectrum and the other group is favored at the high end of the spectrum.

---

<sup>31</sup> Because biology was administered in only two states with small populations, none of the biology items met threshold for DIF evaluation in 2016.

Three logistic regression models were fitted for each item:

$$M_0: \text{logit}(\pi_i) = \alpha + \beta X + \gamma_i + \delta_i X$$

$$M_1: \text{logit}(\pi_i) = \alpha + \beta X + \gamma_i$$

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

Where  $\pi_i$  is the probability of a correct response to the item for group  $i$ ,  $X$  is the matching criterion,  $\alpha$  is the intercept,  $\beta$  is the slope,  $\gamma_i$  is the group-specific parameter, and  $\delta_i X$  is the interaction term.

Due to the number of items being 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 the 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 impact of the addition of the group and interaction terms to the equation. All effect-size values are reported using both the Zumbo & Thomas (1997) and Jodoin & Gierl (2001) indices for reflecting a negligible, moderate, or large effect. The Zumbo & Thomas thresholds for classifying DIF effect size are based off of Cohen's (1992) guidelines for identifying a small, medium, or large effect, with corresponding thresholds of 0.13 and 0.26 for distinguishing negligible, moderate, and large effects. The Jodoin & Gierl approach expanded on the Zumbo & 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 non-uniform DIF and makes use of classification guidelines that are based on the widely accepted ETS Mantel-Haenszel classification guidelines. The Jodoin & Gierl threshold values for distinguishing negligible, moderate, and large DIF are 0.035 and 0.07, whereby items with an effect size less than 0.035 are classified as having negligible DIF, and so on. Similar to the ETS method, negligible effect is classified with an A, moderate effect with a B, and large effect with a C for both methods.

### IX.3.A.ii. Results

**Uniform DIF Model.** A total of 34 items were flagged for evidence of uniform DIF when comparing  $M_1$  to  $M_2$ . Table 81 summarizes the number of items flagged for evidence of uniform DIF by grade band. The percent flagged for each grade band ranged from 10 to 13.



Table 81. Items Flagged for Evidence of Uniform DIF

Grade Band	Items Flagged	Total Items	% Flagged	Number Moderate or Large Effect Size
Elementary	9	82	11	0
Middle	11	82	13	0
High	14	136	10	0

Using the Zumbo and Thomas (1997) effect-size classification criteria, all 34 items were found to have a negligible change in effect size after adding the gender term to the regression equation.

Using the Jodoin & Gierl (2001) effect-size classification criteria, all 34 items were found to have a negligible change in effect size after adding the gender term to the regression equation.

**Combined Model.** A total of 34 items were flagged for evidence of DIF when both the gender and interaction terms were included in the regression equation. Table 82 reviews the number of items flagged for either uniform or non-uniform DIF by grade band.

Table 82. Items Flagged for Evidence of DIF for the Combined Model

Grade Band	Items Flagged	Total Items	% Flagged	Number Moderate or Large Effect Size
Elementary	10	82	12	0
Middle	14	82	17	0
High	10	136	7	0

Using the Zumbo and Thomas (1997) effect-size classification criteria, all 34 items were found to have a negligible change in effect size after adding the gender term to the regression equation.

Using the Jodoin & Gierl (2001) effect-size classification criteria, all 34 items were found to have a negligible change in effect size after adding the gender term to the regression equation.

Overall, results from the uniform and non-uniform DIF analyses across all pools of content had low flagging rates and all flagged items had negligible effect sizes.

While not found in the 2015-2016 administration, the process for items flagged for evidence of DIF with either a moderate or large effect size dictates additional are reviewed by content 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 results of flagging alone.

#### IX.4. EVIDENCE BASED ON RELATIONS TO OTHER VARIABLES

According to *Standards for Educational and Psychological Testing*, “analyses of the relationship of test scores to variables external to the test provide another important source of validity



evidence” (AERA et al., p. 16). For the first operational testing year in science, external validity evidence was evaluated using two types of correlational analyses. First, inter-correlations were calculated between DLM content areas for students assessed in English language arts, mathematics and science using total number of linkage levels mastered. Relationships across content areas can provide an indication of how consistently students perform across the different constructs of interest. However, since these constructs are inherently different (and therefore assessed separately), only moderate relationships are expected. Second, correlations between student demographic characteristics and assessment results were calculated for students assessed in science. Relationships between student characteristics and assessment results can provide a form of discriminant validity evidence when correlations are close to zero. In other words, how a student performs on the test should be unrelated to demographic characteristics such as gender and race. Variables were selected for inclusion based on the amount of data that was available (e.g. 97% of the data for English language learner participation was missing and was therefore, not included in this analysis).

Table 83 below displays the correlation coefficients between science total linkage levels mastered and ELA and mathematics. Overall, relationships were moderate as expected.

Table 83. Correlations of Total Linkage Levels Mastered in Science with English Language Arts and Mathematics

<b>Content Area</b>	<b>Correlation</b>
English Language Arts	0.57
Mathematics	0.59

Table 84 shows the Pearson correlation coefficients between science total linkage levels mastered and selected student demographic characteristics. All coefficients were close to zero suggesting that student performance is unrelated to the student characteristics of gender, race and Hispanic ethnicity as expected.

Table 84. Correlations of Total Linkage Levels Mastered with Selected Demographic Characteristics

<b>Characteristic</b>	<b>Correlation</b>
Gender	0.03
Race	0.03
Hispanic Ethnicity	0.04

Overall, the evidence available after the first operational administration in science supports the validity claim that DLM assessment results are related to other measures of student achievement and unrelated to student characteristics that should not impact academic achievement.

## **IX.5. 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., p. 19). In order to establish sound score interpretations and delimit score use, score reports must be useful and provide relevant information for teachers that informs instructional choices and goal setting. Teachers must use horizontal and vertical recommendations to plan subsequent instruction, and scores can only be interpreted and used for purposes called out in the theory of action as part of the validity argument. Chapter VII provides evidence that the DLM Consortium developed score reports and interpretive resources to support intended uses and interpretations.

As educators and students become familiar with a new assessment during the first operational year, there is limited potential for consequential evidence. Two sources of evidence are discussed for 2015-2016. Results are presented on a multi-stage research effort on DLM score report design and interpretation along with plans for a longitudinal test administrator survey.

### ***IX.5.A. DLM SCORE REPORT DESIGN AND USE***

During the development of the assessments for ELA and mathematics, the DLM Consortium embarked on a series of studies to inform the development of and evaluate the effectiveness of individual student score reports. The resulting score report template was also adopted for science with small changes to accommodate the differences between subjects. The studies that informed the development of the initial template are summarized below.

First, focus groups were conducted in five states with parents of children with disabilities to learn about parent perceptions of alternate assessment based on alternate achievement standards (AA-AAS) and parent need for information about student performance (Nitsch, 2013). When asked to rate their knowledge of alternate assessments on a scale of 1 to 10, with 1 being uninformed and 10 being very informed, parents rated themselves as having relatively little knowledge of AA-AAS, and some indicated they had not received AA-AAS score reports from their schools. Parents tended to perceive the purpose of AA-AAS as to fulfill a legislative mandate and to drive decisions about the school (including educator evaluation and determination of resources) rather than to provide information about their child or measure things relevant to their child’s learning. Concerns about the information parents received on AA-AAS results included lack of understanding of how scores were determined or how the content was related to academic content standards, unfamiliar terminology, a focus on deficits more so than progress, and lack of information about how results could be used to change instruction or provide different supports to their child.

In 2014, additional focus groups were conducted with parents, advocates, and educators (Clark, et al., 2015). Participants evaluated prototype score reports. Prototypes were refined between

waves of feedback, with the goal of maximizing the clarity of the contents and supporting accurate interpretations. Preliminary evidence supported educators' ability to interpret the reports' contents. Parents appreciated the emphasis on strengths rather than deficits but expressed concern about educators' ability to communicate about the contents. Participant feedback led to many of the features seen in the 2014–2015 ELA and mathematics score reports, including narrative statements and linkage level descriptors for every EE (see DLM System Design below for more information about report contents).

Building on the previous research that informed score report design (Nitsch, 2013) and refinement (Clark et al., 2015), the purpose of this study was to evaluate educators' interpretations and use of DLM individual student score reports. Specific research questions included the following:

1. How do participants read and interpret the information in reports?
2. How do participants explain results to parents?
3. What resources do participants use to support their interpretation and use of report contents?
4. How do participants use report contents for educational planning and instruction?

#### **IX.5.A.i. Methods**

As the study was based on the report templates initially developed for ELA and mathematics, the differences for science are described in footnotes in this section. Appendix F shows an example of the science individual student report. As described in Chapter VII, the Performance Profile aggregates linkage level mastery information for reporting on each conceptual area<sup>32</sup> and for the subject overall. The Learning Profile shows rows for each EE and columns that correspond to the five linkage levels<sup>33</sup> (Initial Precursor, Distal Precursor, Proximal Precursor, Target, and Successor). Table 85 summarizes the components of the Performance Profile and Learning Profile that make up the individual student score report. These components were part of the coding scheme used for data analysis and are referred to by number throughout the results section.

---

<sup>32</sup> For science the aggregation occurs for each science domain: life science, physical science and earth and space science.

<sup>33</sup> For science there are three linkage levels: Initial, Precursor, and Target.

Table 85. Components of the DLM Individual Student Score Report

Performance Profile	Learning Profile
1) Overall performance level: <ul style="list-style-type: none"> <li>a) Narrative</li> <li>b) Graphic</li> <li>c) Performance level descriptors</li> </ul> 2) Conceptual areas: bar graphs with subtitles           3) Mastery list: <ul style="list-style-type: none"> <li>a) Conceptual area headings</li> <li>b) Introductory statement</li> <li>c) Bulleted statements</li> </ul>	4) Learning Profile narrative           5) Conceptual area and Essential Element codes           6) Mastery information: <ul style="list-style-type: none"> <li>a) Mastered (green)</li> <li>b) No evidence of mastery (blue)</li> <li>c) Untested (no shading)</li> </ul>

Results were based on individual interviews and paired interviews conducted with teachers in one state. Protocols were slightly different for individual and paired interviews, but both versions were semi-structured.

The individual interview protocol began with general questions about the participant’s background with DLM assessments and previous experience with the score reports. Then the participant was presented with the first score report and asked what it said about the student. Participants were asked to think aloud while they read the contents. Probes were used for clarification of responses and to ensure participants attended to each part of the report (e.g., to point them back to a section they skipped). After interpreting each section of the report (i.e., Performance Profile and Learning Profile), the participant was asked how they might say things differently when explaining the report to a parent. The same process (initial interpretation and reinterpretation for a parent) was followed for a second, contrasting report. The interview concluded with an opportunity for the participant to make recommendations about resources that other teachers would need to support their interpretation and use of DLM score reports.

The paired interview began with the same general background questions as the individual interview but also included a question about the participants’ history of collaboration. The pair was then presented with a score report and asked to talk aloud about their interpretation of its contents. The primary focus of the interview was the use of the report to plan for instruction, including long-term educational planning and mid-year adjustments to instruction. Participants engaged in unstructured dialog about the contents and probes were used during the dialog as needed for clarification and elaboration to cover both major categories of use (instruction and Individualized Education Program planning). After repeating the process with a second, contrasting report, the interview concluded with an opportunity for recommendations about resources to support score report interpretation and use.

Both types of interviews used 2014–2015 score reports with realistic student results but fictitious student identifiers. Sample score reports were prepared in both subjects (ELA and math) and

across elementary, middle, and high school grades. Samples were also selected within each subject and grade band to provide contrasting patterns of student performance.

Each interview incorporated two sample reports. The choice of specific reports for each interview were based on the participant's familiarity with the grade band and subject. For example, a middle school educator who was responsible for both ELA and mathematics might be presented with an ELA grade 6 report for a high-achieving student and a mathematics grade 7 report for a low-achieving student. There was no intentional sequence in which report was presented first.

Interview participants included 12 teachers from two states and two parent advocates from one state. In the first state, eight teachers taught in a school that exclusively served students with intellectual and multiple disabilities from sixth grade through age 21. Teacher participants in the first state taught in secondary grades (grades 6-8, grades 9-10, or grades 11-12). Two of the teachers in the second state taught students with intellectual and multiple disabilities at a regional high school. The remaining two teachers taught student with disabilities at two elementary schools in the same district. Their years of teaching experience ranged from 1 to 26 years. Eight teachers participated in individual interviews and four more participated in two paired interviews.

Individual interviews were coded using a two-step process. First, the researcher reviewed each transcript to mark responses related to the primary research questions (i.e., reading and interpretation, explanation to parents, resources to support interpretation, and uses of report contents). During the second step, the researcher added codes to identify the part of the report the participant was referring to. Thematic codes were also used to identify processes or elements associated with the primary codes. For example, within responses coded as reading and interpretation, statements were also coded to indicate the types of behaviors (e.g., paraphrase, question about contents, misinterpretation). A tentative list of codes was developed prior to analysis, based on review of the literature. Codes were added and refined as new ideas emerged from the data. Paired interviews relied on many of the same codes as individual interviews, but the emphasis was primarily on uses of the contents rather than interpretation. Since the results presented in this manual are preliminary, they are descriptive with regard to the themes, not quantified for dominant patterns.

### **IX.5.A.ii. Results**

**Reading and Interpretation.** Participants varied in the parts of the report that they tended to rely on for information. Results are described with numeric references back to the report component listed in Table 85.

Since the interview imposed minimal structure on the order in which participants reviewed the report and the emphasis they placed on each section, each participant's preferences for information were clear in the think-aloud portion of the interview, even before discussing the report contents. The following examples illustrate a few scenarios:

- Anna<sup>34</sup> walked systematically through each major section of the entire report, starting with the Performance Profile narrative (1a) to characterize the student's overall performance, describing conceptual areas (2) as general strengths and weaknesses, and using the mastery list (3) to reflect on skills seen during the assessment. In the Learning Profile, she emphasized the mastery information (6) and did not use the narrative (5).
- Liz briefly mentioned the numbers in the Performance Profile narrative (1a), and spoke briefly about all parts of the Performance Profile, but had a strong preference for the mastery information (6) in the Learning Profile.
- Margaret primarily relied on the conceptual areas (2) and looked to the mastery list bullets (3c) to identify examples of the skills in each area, especially when talking to parents. When thinking about instruction, she gravitated to the mastery information (6) in the Learning Profile.

In general, participants paid little attention to narrative statements (1a, 4), and only one briefly mentioned the performance level graphic (2). The Performance Profile mastery statements (3) and Learning Profile mastery table (6) were emphasized the most. More detail about interpretation of the Learning Profile is provided in the Report Use section below.

As participants talked through the report contents, most of their comments were verbatim or near verbatim language from the report. Minimal paraphrasing was occasionally used when interpreting results for parents:

*I basically sort of explained the [performance] levels first . . . so I said emergent is they're just starting out with this skill. They may not have a good understanding. And then I said approaching Target, they have some understanding. And then I said Target is right where we want them.*

Statements about report contents were also evaluated for signs of misinterpretation or misunderstanding. Since most statements were verbatim or near verbatim, there were few opportunities for misinterpretation. One type of misinterpretation came from inappropriately applying terms from one part of the report to results in other sections. For example, in one case, a student was described as “emerging” (a performance level descriptor) in one of the conceptual areas although there are no performance levels assigned to conceptual areas. In another case, the student was described as having “mastered” a conceptual area although mastery judgments are only made at the linkage level. Both of these misstatements were attempts to give a qualitative label to a percentage of skills mastered in a conceptual area.

One participant misinterpreted the percent values reported for conceptual areas when talking to parents. Instead of describing percentage of skills mastered, she interpreted percent as it is often used in monitoring instruction and setting instructional goals for students with the most significant cognitive disabilities: percent accuracy or percent correct over repeated trials.

---

<sup>34</sup> All names are pseudonyms.



*So it's like constructs understanding [conceptual area]—he can identify concrete details in an informational text [linkage level]. But reminding the parent that that was only like a 20 percent. . . . But it seems that oh, my child can identify that. Then you're like, well, but if we look back here, again, remember, that was one out of five times. So it's still only with 20 percent accuracy, which is—you want 80 percent. So definitely make sure they understand that like a Target child, that goal is about 80 percent for their classmates.*

The most extreme misconception was seen for one participant who asked many questions that reflected his confusion. Some of his challenge was in relating the score report contents to the assessment design and administration. He could not recall how testlets were assigned or the relationship between the linkage level tested and where mastery would be reported. He also wanted to see information in the Performance Profile (i.e., which skills were not mastered) without realizing it was in the Learning Profile. He reported using the Performance Profile bulleted mastery list with parents and the Learning Profile to think about instruction.

#### ***IX.5.A.ii.a Interpreting Reports for Parents***

Each participant indicated that they were selective about the parts of the report they chose to discuss with parents. Most commonly mentioned were the conceptual area (CA) bar graphs<sup>35</sup> (2), bulleted mastery list statements (3a), and the entire Learning Profile. For example, one teacher used the CA bar graphs to explain the student's general strengths and weaknesses before discussing more specific skills from the bulleted list as examples from specific CAs. Those who preferred to discuss the Learning Profile with parents pointed out that it allowed them to focus on current mastery as well as areas for instruction, whether that be to reteach something that was not mastered or move to another skill after mastering a previous one. The participant who reported less discussion of the report with parents said she focused only on the CA bar graphs and referenced a couple of skills from the Learning Profile. Her rationale was that parents' best level of understanding was in the CAs. She sent the report home with them and invited them to ask her questions after they looked it over on their own.

Although the mastery list (3) and the Learning Profile (6) contained very similar information, some teachers preferred one over the other. Those who preferred the bulleted mastery list tied the CA headings (3a) back to the bar graphs to help anchor their conversation with the parent. When discussing results that did not resonate with parents (e.g., the student demonstrated mastery of a skill the parent thought was implausible or did not demonstrate mastery of a skill the parent believed the student possessed), another strategy was to refer to the introductory statements (3b) to remind the parent that the report was explaining evidence of mastery from the DLM assessments and that there were multiple ways the student might demonstrate the skill.

As participants described the ways in which they talked with parents about report contents, it became clear that they added contextual information to support parents' understanding. For

---

<sup>35</sup> For science, the bar graphs are provided for each domain: life science, physical science and earth and space science.



example, one teacher drew connections to the reports for the general education assessments and content standards, since many parents were familiar with those for other children in their family. Another strategy was to explain why the assessment was challenging that year (e.g., that the assessment was still relatively new, or that they expected the student to improve after becoming more familiar with working in a computer-based environment).

When discussing specific mastery statements or linkage levels from the Learning Profile, another contextualizing strategy was to describe what the skill looked like for that student, either during assessment or during instruction. One participant modeled how she would talk to a parent about an EE that had no evidence of mastery on the Learning Profile:

*I even have parents with some intellectual needs. I would actually say it to them that your student—you see these highlighted areas right here in the blue? These areas were the areas where they're struggling, right here, and these areas are the areas that they did really well, and we want to focus on those areas where they were struggling, and right here—so understanding function of the objects—okay, what does that mean? So let's say, we need [the student] to understand that when she goes over and turns that light on—so understanding what that means, we're going to work on that.*

Yet describing skills to parents was difficult when teachers themselves did not understand the linkage level statement. Two types of challenges were noted. First, academic vocabulary was seen as a barrier to talking with parents about the report. One participant commented on the word "subitizing" in a linkage level descriptor:

*I had that word and we were like what does that mean? We had to get on our phone and look it up to see what it meant, and it was like I can't even teach it if I don't know what it means, and how does a parent understand it if we don't know what it means?*

A second challenge occurred when two similar linkage level statements were difficult to distinguish from one another. One participant illustrated this challenge as she talked through her understanding of "match pictures with representations of real objects" and "match pictures with real objects":

*That says matching pictures with representation of real objects. That's interesting. Match a picture with a real object. . . . I might have a parent ask me why did they do well here and they didn't do well here? Why did they not do well there and they did well here? . . . So, these are two different areas. This one is in the—I'm going to get this wrong. One is in reading . . . reading, and yes, and this one is . . . reading information, right. Okay, yes. I know, but I'm missing it, but okay, yes, yes. So this is in the story itself. This is in the story itself. So when she's reading the story and understanding, she's getting that information. Okay. She's able to match pictures with, yes, okay. And this is just absolute picture, just like, identifying. Okay. All right.*

### **IX.5.B. TEACHER RESOURCES**

All teachers in this preliminary study were from the same campus. The campus had an instructional facilitator and built-in time for both structured professional development sessions

and professional learning community meetings. All of the participants credited those resources with helping them interpret and use the score reports. For example, they had a one-hour professional development session on how to read the score reports. In the professional learning community meetings, they planned for assessment, shared materials and resources, and helped one another with interpretation of linkage levels. Several participants mentioned talking with the student's teacher from the previous year (whether from within their school or at another school) to better understand how a student was demonstrating a skill that was listed as mastered on the score report.

### **IX.5.B.i. Report Use for Planning Instruction**

Participants described a range of uses of the report contents beyond sharing the results with parents. For this manual, uses are roughly grouped into planning for instruction and Individualized Education Program (IEP) development.

**Planning for Instruction.** A consistent finding across interviews was teachers' use of the Learning Profile to guide instruction. This included looking to the next linkage level beyond the highest level mastered for a given EE and planning to instruct next on that level. However, where students were assessed and did not show mastery, or where teachers thought the student's mastery was limited, teachers indicated they would reteach a skill that the student had already mastered.

Some participants provided evidence of more sophisticated evaluation and planning, particularly by looking at connections across linkage levels and EEs to think about larger instructional goals.

*Because he's mastered the Level 3, which is the Precursor—so we want him to get up to the Target, so I would start teaching for the Target for the student, tying it back into the Precursor stuff that he can do, so that we're not working on stuff that he already knows.*

*So if we can connect those two Elements there, we know that we can start up here with them on this one, and I'd have to explain that to a parent, and then I would want to know where he's at with this. Once we teach him how to do that, how fast is he going to pick that up to doing the real-world problems with numbers, and if he can do real-world problems up here with numbers, can he do it the same way here? This is adding and subtracting—so this is multiplying, so it would be different, but how is it different there and the same there?*

Sometimes an apparently inconsistent or unusual pattern of performance raised questions for the teacher. The typical response was a desire to assess further using their routine classroom methods to understand possible reasons for the inconsistency:

*He can combine and partition sets, which should lead to multiplying. I don't understand why he can do multiplying in one but not combining in another. I guess I would want to take a look at that one and see how those lead to each other, because combining and portioning are the same I guess for both multiplication and adding and subtracting.*

When planning for instruction in an area the student had not mastered, the teacher sometimes relied on understanding of the DLM assessment content. One common instructional strategy for students with the most significant cognitive disabilities is to first teach a skill in a familiar context and then work on transferring the skill to novel situations. One participant describing instruction on "identify the end of a familiar routine" offered this example related to a reading testlet:

*What type of routine for it? I know that on the assessments that was really hard for me to think of what type of routine are we using . . . because the example has you doing stuff out of a book, and that's the routine is what's in the book but then how do you end that routine? . . . Well what do we do at the end of math? It all depends on the day. . . . Okay when we are getting ready to go on the bus, what's the last thing that you do? You buckle yourself in. Okay. That type of thing for familiarity.*

There were a few other ways in which teachers mentioned using the report to plan for instruction, but none of them were described in depth. Examples included using the Learning Profile to develop lesson plans and creating instructional groupings when students working on different skills were being taught together.

**IEP Planning.** Participants described using score report contents primarily for two parts of IEP development: statements on the student's present levels of performance and annual goals. The tendency was to use the performance level narrative (1a) and mastery skill list (3c) nearly verbatim in statements of present levels of performance:

*I'd take this whole thing and say use this. So say over the assessment is covering fifty skills, for ten Essential Elements, Hunter mastered 37 skills during the year and overall his mastery fell on to at Target. And then I would say specifically what he has mastered. And then, if he didn't show skills: however, Hunter was tested, did not show these skills or he struggled with these skills, and then we'd say what he struggled with.*

The Learning Profile, and specifically the next skills that had not been mastered, were one source of information participants reported using to develop IEP goals. However, the expectation in their school was that the Learning Profile be considered along with other assessments and school-developed checklists in order to identify goals for the student in reading, writing, and math. The contents of IEP goals spanned multiple EEs, and the objectives associated with the goals were based on teacher estimates of reasonable instructional targets:

*We look at all of the elements that are being assessed. We say where they're starting . . . We would look at where they're starting, either where they were assessed at or like this year we talked about they were at the Initial [Precursor] level. Most of our students are. And we created some scales, but we would look at where we felt like they could achieve within a year, and we kind of made it into a percentage. So this is where they're starting. These are the things that we would like to see them get to this year and so create a percentage within that.*

Besides these two uses of score reports to guide IEP development, one teacher pointed to another possible use of the information for IEP teams. When reviewing a sample score report that showed a student whose overall performance was at the highest performance level, she questioned that student's placement and eligibility for an alternate assessment. Both educational setting and assessment eligibility would be determined by an IEP team.

### ***IX.5.C. BASELINE TEST ADMINISTRATOR SURVEY RESPONSES***

As mentioned earlier in this chapter, a survey is planned for the 2017 spring administration of the science assessment and will serve as one source of data for consequential validity evidence. This survey will be distributed to test administrators and will include questions regarding their perceptions of the assessment contents; specifically, whether or not they agree that the content of the test measures important academic skills and reflects high expectations for the student. These questions will be repeated annually for longitudinal data collection, and test administrators will be asked to complete the survey for each student to whom they administered a DLM assessment. Results from the 2017 administration of the survey will be included in the 2016-2017 update to the technical manual.

### **IX.6. CONCLUSION**

This chapter presents additional studies to support the overall validity argument for the DLM Alternate Assessment System. The studies are organized into categories (content, response process, internal structure, relations to other 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. More specifically, validity evidence based on test content was provided through an external alignment study as well as survey data on students' opportunity to learn science content. Existing validity evidence based on response process and plans for future data collection were outlined and item-level bias was evaluated as part of the evidence provided on the internal structure of the assessment. Plans for evaluating the relationships between assessment outcomes and external variables are provided. Finally, the chapter presented evidence based on consequences of testing is provided through score report design and use studies and additional plans for a longitudinal test administrator survey.

The final chapter of this technical manual, Chapter XI, references evidence presented throughout the technical manual, including this chapter, and expands the discussion of the overall validity argument. The concluding chapter also provides areas for further inquiry and ongoing evaluation of the DLM Science Alternate Assessment System.

## **X. TRAINING AND INSTRUCTIONAL ACTIVITIES**

Chapter X describes the training that was provided in 2015-2016, which included webinars for state and local education agency staff, four required training modules and one optional science module for test administrators, and several optional instructional activities. Required test administrator training ensured that test administrators had both the context and practical knowledge of the assessment system design, administration, and security practices to administer the test with fidelity. All required test administrator training was aligned with the *Test Administration Manual 2015-2016* (Dynamic Learning Maps, 2016b). See Chapter IV for a thorough discussion of test administration.

### **X.1. TRAINING FOR STATE EDUCATION AGENCY STAFF**

State education agency (SEA) staff are integral to the implementation of the DLM alternate assessment system. In 2015-2016, webinars were provided for state and local agency staff. The webinars were targeted to various roles and by model. In late fall, the webinars were for Assessment Coordinators, Technical Liaison, and Data Stewards. These webinars were live presentations using Skype for Business with time allotted for question and answers, both by audio and through the chat window. The webinars were recorded and made available on each state's DLM website. The webinars identified changes in Educator Portal from 2014-15 for staff returning to their past roles. At the same time, the webinars also focused on staff who were new to the roles for the DLM alternate assessment. DLM staff also published the frequently asked questions from all webinars.

#### **X.1.A. TRAINING FOR LOCAL EDUCATION AGENCY STAFF**

Three main roles support implementation of the assessment system. These roles are normally held by one or more district-level staff members, but in some cases are fulfilled at the building level.

- The Assessment Coordinator oversees the assessment process, including managing staff roles and responsibilities, developing and implementing a comprehensive training plan, developing a schedule for test implementation, monitoring and supporting test preparations and administration, and developing a plan to facilitate communication with parents/guardians and staff.
- The Data Steward manages educator, student, and roster data.
- The Technical Liaison verifies that the network and testing devices are prepared for test administration.

Webinars were held prior to the opening of the spring assessment window for district and building staff who were responsible for overseeing test administration. The purposes of these webinars were to provide reminders about the assessment administration process and describe strategies for monitoring assessment administration. The content of the monitoring webinar is included in Appendix D.

## **X.2. REQUIRED TRAINING FOR TEST ADMINISTRATORS**

Training is required annually for educators who serve as test administrators and administer the DLM alternate assessments. In 2015-2016, training was available in two formats: facilitated training (in-person training with post-tests in Moodle) and self-directed training (all content and post-tests within Moodle).

All new test administrators were required to successfully complete four modules and pass all four post-tests with a score of 80% or higher before delivering assessments; they were not allowed access to their students' log-in information for the student Kansas Interactive Testing Engine (KITE) platform until successfully completing their training. Test administrators were able to retake post-tests as many times as needed in order to pass all parts of the training.

Returning test administrators had to successfully complete a single combined module with a score of 80% on each of four post-tests before being allowed access to their students' log-in information. Training time was estimated at less than one hour. If a returning test administrator did not successfully complete the module post-test on the first attempt, they were required to take additional training. This training could take an additional 30 minutes to 4 hours, depending on the areas in which the test administrator was not successful on the first attempt.

Educators in each state had access to both facilitated and self-directed training options for new test administrator training. Participants chose the correct version according to their state's guidelines. Figure 46 illustrates the differences between the two training formats. Training for returning test administrators was only available in self-directed format.



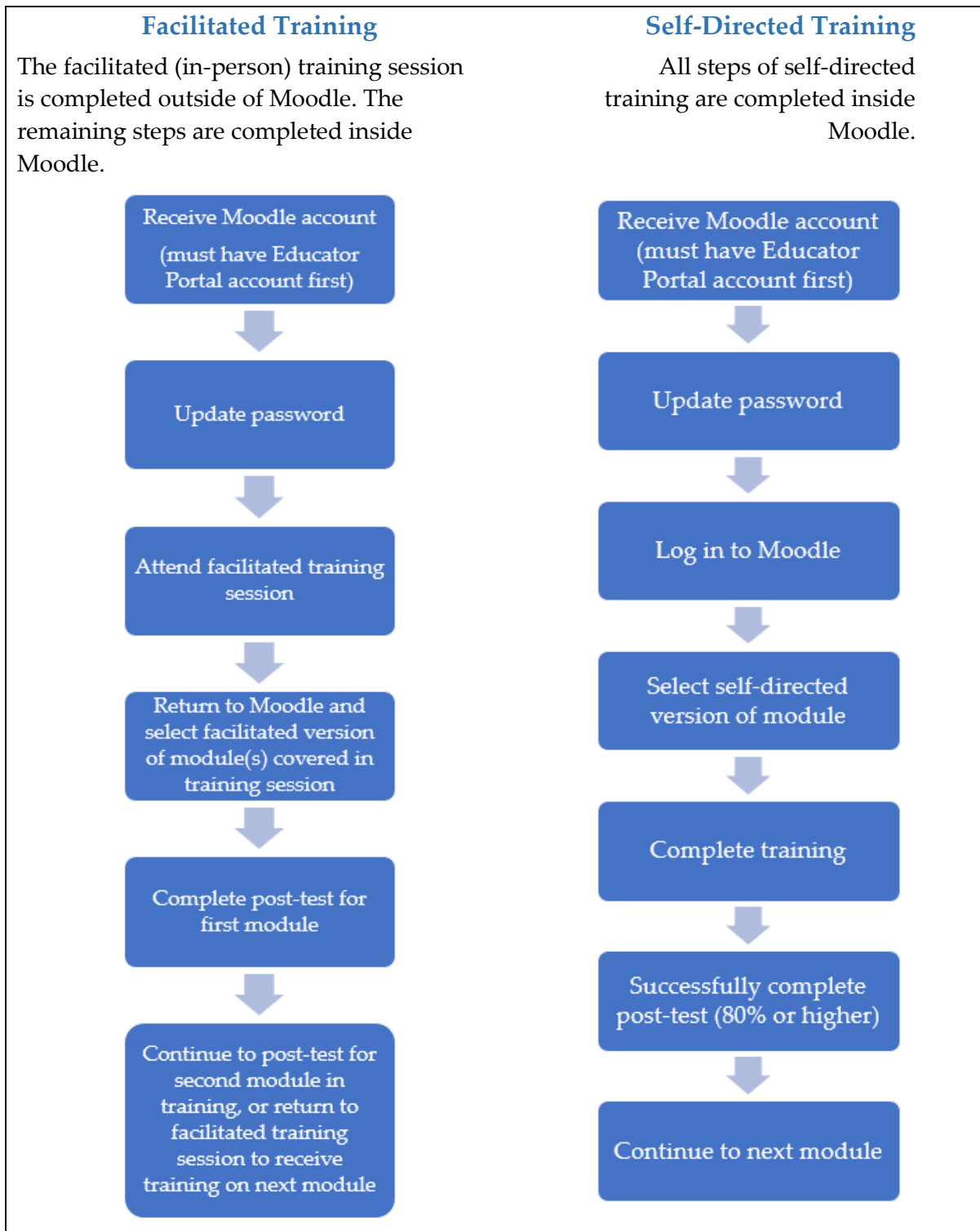


Figure 46. Required training processes flows for facilitated and self-directed training.



### ***X.2.A. FACILITATED TRAINING***

The facilitated modules are intended for use with groups. This version of the modules was designed to meet the need for face-to-face training without requiring a train-the-trainers approach or requiring the facilitator to have deep expertise in the subject matter. Each state determined its own policy guidance regarding who served as facilitators. Examples of individuals who served as facilitators included district- and building-level test coordinators, district special education coordinators, instructional coaches, lead educators, SEA staff, and trainers from regional education agencies that are responsible for professional development.

Facilitators were provided an agenda, a detailed guide, handouts, videos, and other supports required to facilitate a meaningful, face-to-face training. Facilitators showed the DLM staff-produced videos and implemented learning activities as described in the facilitator guide. Facilitators who wished to add to the training contents or deliver the content themselves rather than via video also had access to the PowerPoint slides and scripts. Appendix H includes the complete set of training materials for all four Required Test Administrator Training modules used in 2015-16.

Facilitators were encouraged to prepare themselves by reviewing all videos and all sections of the *Test Administration Manual 2015-2016* (Dynamic Learning Maps, 2016b) addressed in the training. States also recommended that facilitators complete the training requirements themselves. Facilitators who were also test administrators were required to pass the post-tests. Facilitators were asked to ensure that participants had Educator Portal accounts and Moodle accounts and had accessed them prior to the facilitated training session. Facilitator responsibilities included setting up the training area with equipment, delivering the facilitated training modules, and directing users to return all equipment. Finally, facilitators directed test administrators to take each module post-test in Moodle with support from the *Guide to DLM Required Test Administrator Training* (Dynamic Learning Maps, 2014b) for detail and access procedures. Facilitated training was flexibly structured so post-tests could be taken onsite during training sessions (e.g., in a computer lab) or independently after the training session was complete. Whether during the facilitated training or afterwards, facilitators were to direct all test administrators to take the post-tests independently and never as teams or as a group activity.

### ***X.2.B. SELF-DIRECTED TRAINING***

The self-directed modules were designed to meet the needs of educators who were unable to attend facilitated sessions and needed access to on-demand training. Self-directed modules combine videos, text, and online learning activities to engage educators with a range of content, strategies, and supports, as well as the opportunity to reflect upon and apply what they are learning. The videos are identical to those used in facilitated training. Each module ends with a post-test.

In 2015-16, the self-directed training was completed entirely in Moodle with support from the *Guide to DLM Required Test Administrator Training* (Dynamic Learning Maps, 2014b) for detail

and access procedures, including the review of all module slides and procedures for completing all post-tests.

### ***X.2.C. TRAINING CONTENT***

Training content included four required modules about the DLM assessment system in general, including ELA and mathematics; and one optional science module. Since all of the states participating in DLM science in 2015-16 also participated in ELA and mathematics, test administrators only needed additional information regarding the ways in which science differed from ELA and mathematics. The science module did not have an accompanying post-test. Contents of all modules are provided in Appendix H.

#### **X.2.C.i. Module 1: About the DLM System**

Module 1 of the test administrator training provided an overview of the DLM system components and DLM test security. Topics included illustration and discussion of the DLM maps, Claims and Conceptual Areas, Essential Elements (EEs), testlets, linkage levels, and the security demands of the DLM system. Participants were expected to demonstrate an understanding of the DLM maps, including the academic nature of the knowledge, skills, and understandings described within them. They were also expected to develop a working definition of the EEs and differentiate them from functional skills. Participants were to be able to define Claims and place them within the context of instructional practice. Finally, educators were expected to practice the security guidelines for assessments as outlined in Module 1.

Module 1 explained the development of DLM testlets. It also emphasized the fact that Target level testlets are aligned directly to the Essential Element being tested, while explaining that testlets at other linkage levels are developed using the DLM map nodes that build up to, and extend from, the target node(s). In addition, participants were taught about the adaptive nature of the assessment, explaining that students could potentially see all five levels of testlets (Initial Precursor, Distal Precursor, Proximal Precursor, Target, and Successor) in their assessment, whether ELA or mathematics. They were introduced to mini-maps that specifically detail the nodes assessed at each linkage level.

After viewing Module 1, participants were expected to know all the DLM security standards. These standards apply to anyone working with the DLM assessment. The standards are meant to ensure that assessment content is not compromised, and they include not reproducing or storing testlets, not sharing testlets via email, social media, or file sharing, and not reproducing testlets by any means, except in clearly specified situations (e.g., braille forms of the testlets).

Participants agreed to uphold the DLM security expectations by signing an annual agreement document and committing to integrity. In addition, participants were instructed to follow their own state's additional policies that govern test security.

### **X.2.C.ii. Module 2: Accessibility by Design**

Module 2 of the required training focused on accessibility. Participants were shown the characteristics of the DLM system that were designed to be optimally accessible to diverse learners, as well as the six steps for customizing supports for specific student needs, as described in detail in the *DLM Accessibility Manual*.

The training emphasized how Universal Design for Learning was used to ensure that test content was optimally accessible. The technology platform used to deliver assessments, KITE Client, was introduced, along with an explanation of its accessibility features, including guidelines for selecting features for the Personal Needs and Preferences Profile (PNP).

Participants were expected to demonstrate understanding of accessibility features, their purpose, student eligibility, and appropriate practice. In addition, participants were shown how to complete the PNP and how the PNP and First Contact survey responses combined to develop a personal learning profile to guide administration decisions for each student.

Module 2 also demonstrated how to actualize all accessibility features for an individual student, both within KITE Client and through external supports, in conjunction with Testlet Information Pages (TIPs).

Module 2 addressed flexibility in the ways that students access the items and materials, including what is considered appropriate flexibility (e.g., test administrator adapts the physical arrangement of the response options) and what is not (e.g., test administrator reduces the number of response options).

Finally, participants were taught how accessibility supports must be consistent with those that students receive in routine instruction and how those supports may extend beyond testing accommodations that are specifically mentioned in the student's IEP.

### **X.2.C.iii. Module 3: Understanding and Delivering Testlets in the DLM System**

Module 3 focused on participants' understanding and delivery of content through testlets within KITE Client. Topics included testlet structure, item types, completing testlets, standard test administration process, allowable practices, and practices to avoid.

The third module provided participants with focused information on how the assessments are delivered via computer. Content included the testlet structures used in the assessment system, the various item types used (e.g., single-select multiple choice, matching, sorting, drag and drop), how to navigate and complete testlets, and what to do on test day.

Module 3 also addressed teacher-administered testlets, including the specific structures used and the processes for completing testlets by administering them outside KITE Client. The module also covered how the test administrator enters responses into KITE Client. The training emphasized the importance of educator directions provided within the testlet and specific directions to each content area (i.e., reading, mathematics, and writing). This module also included details on standard administration processes, allowable practices, and practices to avoid.

#### **X.2.C.iv. Module 4: Preparing to Administer the Assessment**

Module 4 prepared participants in their role as test administrators. They learned to check data, complete the First Contact survey, use practice activities and released testlets, and plan and schedule assessment administration.

Participants reviewed the test administrators' role in completing data management requirements in Educator Portal, supported by full instructions in the *Test Administration Manual 2015-2016* (Dynamic Learning Maps, 2014b). Participants reviewed the DLM assessment components, which are accessed through Educator Portal (e.g., First Contact survey) and where student information is entered. Participants learned about students' required activities during operational testing as opposed to opportunities to practice through released testlets or practice activities available in KITE Client.

The training specifically addressed the First Contact survey, which is completed before testing begins. It uses test administrator responses to questions about student communication and academic skills to determine which linkage level is best to start students at the first time they encounter the DLM assessments. The First Contact survey is completed online, but test administrators also have access to all the questions in advance in an appendix to the *Test Administration Manual 2015-2016*. The First Contact survey includes questions regarding special education services and primary disability categorizations as well as sensory and motor capabilities, communication abilities, academic skill, attention and computer access.

The module also addressed planning and scheduling the assessments. Prior to the assessments, test administrators were directed to allow their students taking the assessments to complete practice activities to expose them to the KITE system. Test administrators were advised to retrieve TIPs, determine the appropriate length of each assessment session, and to consider the schedules according to their states' requirements. Test administrators were also instructed to arrange a space for assessments that is quiet, clear from distractions, and able to accommodate students' accessibility needs.

#### **X.2.C.v. Optional Science Module**

In addition to the four required modules designed for all test administrators in all DLM states, a supplemental science module was available (but not required) for test administrators in states administering science. The main focus of the video compared the DLM content areas, pointing out the differences between the science framework, testlet delivery, and design to that of ELA and mathematics. The training included comparison charts and information relevant to test administrators who deliver tests to students in all three content areas. The science module was approximately ten minutes long, but unlike the other modules, the science module did not follow with a post-test.

Like ELA and mathematics, the DLM content standards for science are called Essential Elements. While the ELA and mathematics EEs were written to all college and career readiness standards, the science EEs were written for a selected set of science standards.

Unlike ELA and mathematics, science was tested in grade bands: elementary, middle and high school instead of by grade level. The science EEs are specific statements of knowledge and skills, including science and engineering practices, linked to the grade-level expectation identified in the National Research Council's Framework for K-12 Science Education and the NGSS.

The module described how the current DLM Science framework is different from that of ELA and mathematics. Understanding the science framework involves understanding the relationship among all of the elements within the system. These elements include the domains of life science, physical science, and earth and space science. Another difference is that science has only three linkage levels, with the highest being the Target level and aligning to the content of the EE. The Precursor and Initial levels are less complex than the Target and provide access to the Target level at a reduced depth, breadth, and complexity level. Initial testlets are usually administered by the test administrator, who observes the student's behavior as directed by the system and then records responses in the system. Testlets at the Precursor level allow students to develop the knowledge and skills needed to reach the Target.

The module also explained that most of the supports available in the system for ELA and mathematics were also available for science. The science TIPs presented pictures for the Initial testlets and test administrators were strongly encouraged to print them in color. Testlets for science begin with an engagement activity to provide a context or science story. Students participating in science from both the Integrated Model states and Year-End Model states were administered nine science testlets during the spring assessment window.

### **X.3. INSTRUCTIONAL ACTIVITIES**

Science instructional activities were developed to support educators who were beginning to use the DLM Science EEs. Eight science instructional activities were made available to teachers on the DLM science resources webpage ([http://www.dynamiclearningmaps.org/sci\\_resources](http://www.dynamiclearningmaps.org/sci_resources)). While activities were not developed for every Essential Element, they did cover each science domain (Earth and space science, life science, physical science) and grade band (elementary, middle, and high school). The activities that were available in 2015-16 included:

EE.5.ESS1-2: The Daylight Hours  
EE.5.LS2-1: Food Cycles  
EE.5.PS3-1: Energy from the Sun  
EE.MS.ESS2-6: Weather Watchers  
EE.MS.LS2-2: What Animals Eat  
EE.MS.PS1-2: Chemical Changes  
EE.HS.ESS3-3: Conserving Natural Resources  
EE.HS.LS1-2: Respiratory System

Each science instructional activity provides an examples of how to teach a science lesson that addresses one Essential Element and differentiates instruction for students who access the content at three different linkage levels.

DLM staff collaborated with educators from states in the DLM science consortium to develop the science instructional activities. Educators drafted these activities during the January 2015 item writer workshop, and drafts were reviewed by special educators and science educators before they were published.



## XI. CONCLUSION AND DISCUSSION

The Dynamic Learning Maps 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 completed its first operational administration year in 2015-2016. This technical manual provides evidence 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 (Chapter I, Figure 2). The contents of this manual address the information summarized in Table 86.

Table 86. Review of Technical Manual Contents

Chapter(s)	Contents
I, II	Reviews the foundations of the assessment system, including the development of the theory of action to guide each subsequent step and the development of the Essential Elements and linkage levels for science.
III, IV, X	Provides procedural evidence of test content development and administration, accessibility features and procedures, security protocols, and test administrator training.
V	Describes the statistical model used to produce scores <sup>36</sup> based on student responses.
VI	Provides a description of how cut points were developed to interpret results via performance levels.
VII, VIII	Describes results and analysis of the first operational administration's data, evaluating how students performed on the assessment, the distributions of those scores, aggregated and disaggregated results, and analysis of the internal consistency of student responses.
IX	Provides additional studies focused on specific topics related to validity and in support of the score propositions and purposes.

This chapter reviews the evidence provided in this technical manual and places it within a validity framework in order to assess the program's overall success at producing scores that

---

<sup>36</sup> The term "results" is typically used in place of "scores" to highlight the fact that DLM assessment results are not based on scale scores. For ease of reading, the term "score" is used in this chapter.



mean what they are intended to mean. In addition, future research studies are discussed as part of ongoing and iterative processes of program responsiveness, validation, and evaluation.

## **XI.1. VALIDITY FRAMEWORK**

The *Standards for Educational and Psychological Testing* (AERA et al., 2014) are the professional standards used broadly to evaluate educational assessments; the DLM Alternate Assessment System is no exception. The *Standards* define validity as “the degree to which evidence and theory support the interpretations of test scores for proposed uses of the test” (p. 11) and assert that validity is the “most fundamental consideration in developing tests and evaluating tests” (p. 11). Using the *Standards* as a baseline for the evaluation of the DLM assessments, this manual’s primary purpose is to provide evidence and theory to support the propositions laid out in the DLM theory of action (see Chapter I). The four propositions serve as an organizing framework for the summary and evaluation of validity evidence in this chapter. To this end, Chapter X looks back at the previously presented evidence in support of the score purposes and their proposed interpretations and uses.

All aspects of the validity argument must be carefully evaluated (Lissitz, 2009; Sireci, 2009). The purpose of the assessment with its resultant scores is critical to the overall validity argument as it is indicative of the model from which the assessment was originally designed (Mislevy, 2009). It follows, then, that the evidence collected throughout the entire development process should point to a clear and persuasive link between the original assessment purpose and the uses and interpretations of the results. Clarity between what can be observed (e.g., student responses to assessment tasks) and what must be inferred (e.g., student ability in the content area) must inform the validity and interpretative arguments (Kane, 2006). In addition, the dimensions and organization of the overall validity argument matter, as they include not only the content sampled and procedural bases of the assessment, but also evidence for the underlying construct to be assessed, what may be included on the assessment that is irrelevant to the construct, and the relative importance of the consequences of the resulting scores (Messick, 1989; Linn, 2009).

Validation is the process of evaluating the evidence and theory presented in the overall validity argument. Using the *Standards* as our foundation, the DLM System began the validation process “with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use” (AERA et al., 2014, p. 11). These propositions<sup>37</sup> then informed the development of the theory of action (as described in Chapter I, Figure 2), which focused overall on combining high expectations for students with the most significant cognitive disabilities with appropriate educational supports for teachers, to result in improved academic experiences and outcomes for students.

---

<sup>37</sup> The term “proposition” is used here to mean a claim within the overall validity argument. The term “claim” is reserved in this technical manual for use specific to content claims (see Chapter III).

## **XI.2. PROPOSITIONS FOR SCORE INTERPRETATION AND USE**

The DLM Consortium developed an argument-based approach to validity that established four propositions to support the intended uses and interpretations of DLM scores. These propositions are laid out within a context of precursors, assessment design assumptions, and ultimate goals for the program within the theory of action (Chapter I, Figure 2). The propositions relate directly to the ultimate program goals and specific score purposes, providing the framework within which validity evidence can be judged. The four propositions are as follow:

1. Scores represent what students know and can do.
2. Achievement level descriptors provide useful information about student achievement.
3. Inferences regarding student achievement, progress and growth can be drawn at the domain level.
4. Assessment scores provide useful information to guide instructional decisions.

Summative scores from the DLM assessments are intended for use for several purposes:

1. Reporting achievement within the taught content aligned to grade-level content standards to a variety of audiences including educators and parents
2. Inclusion in state accountability models to evaluate school and district performance
3. Planning instructional priorities and program improvements for the following school year

Appropriate interpretations and uses of DLM scores support the overall goals of the DLM Alternate Assessment System:

1. Students with the most significant cognitive disabilities are able to demonstrate what they know and can do.
2. Teachers make sound instructional decisions based on data.
3. Parents, teachers, and students have high expectations for students' academic achievement.
4. The trajectory of student growth in academic knowledge and skills improves.

Holding high expectations for students with the most significant cognitive disabilities and providing appropriate educational supports for teachers will lead to improved academic experiences and outcomes for students.

## **XI.3. SUMMARY AND EVALUATION OF VALIDITY EVIDENCE**

To build the validity argument, the examination of the proposed score interpretations and purposes necessarily points back to evidence previously presented in this technical manual. This validation review was conducted by examining evidence associated with each proposition, organized by categories of evidence as presented in the *Standards* (AERA et al., 2014). These

categories are (a) test content, (b) response processes, (c) internal structure, (d) other variables, and (e) consequences of testing.

Within each category, we describe related evidence. Although some evidence supports more than one proposition, for the sake of conciseness it is only described with one proposition. Table 87 in the Evaluation Summary section of this chapter summarizes the sources of validity evidence as organized by the propositions and each evidence category.

### ***XI.3.A. PROPOSITION 1: SCORES REPRESENT WHAT STUDENTS KNOW AND CAN DO***

#### **XI.3.A.i. Evidence Based on 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 DLM Alternate Assessment System is intended to support claims about what students know and can do in science.

The interpretation and use of DLM scores depends on evidence of the relationships among the content components of the assessment system. Assumptions related to test content focus on whether the DLM Essential Elements, grade-level expectations for students with the most significant cognitive disabilities, must address the content domains with fidelity and be adequately linked to standards, in this case the *Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (National Research Council, 2012) and the Next Generation Science Standards (2013). Coverage of content, as specified by test blueprints, provides evidence of representation of the target domain overall. Essential Elements and linkage levels are identified for assessment. Thus, items within testlets are aligned to the Essential Elements via the associated linkage levels. Finally, teachers must have instructed the student on the content prior to assessment in order for students to have had the necessary opportunity to learn.

Content-related evidence to support this proposition is described primarily in terms of the goal of alignment. Alignment is “at the heart of the process” of content-oriented evidence of validation and involves evaluating the degree to which test content corresponds to student learning standards (AERA et al., 2014, p.15), which are the Essential Elements in the DLM system. Alignment was considered across the design, development, and operational stages. A second source of content-related evidence in the development phase was the use of procedures to ensure that items and testlets maximize construct-relevant and minimize construct-irrelevant features.

#### ***XI.3.A.i.a Design Phase***

Chapter II describes procedural evidence that supports the representation of the content domains. Through an iterative process and with expert and educator feedback, teams developed Essential Elements for science and linkage levels to describe the target skill at reduced levels of depth, breadth or complexity.

Essential Elements convey the grade-level expectations for students with the most significant cognitive disabilities. As described in Chapter II, the Essential Elements were carefully developed to align to the disciplinary core ideas (DCIs, the content), and science and engineering practices (SEPs), in each grade band, representing high expectations for students so they would be prepared for college, career, and citizenship. The development of the test blueprint demonstrates how content was sampled to cover the content with coverage defined by the science domains.

### ***XI.3.A.i.b Development Phase***

Using a variant of evidence-centered design, the consortium developed Essential Element Concept Maps (EECM) to support assessment development. As described in Chapter III, EECMs are graphic organizers for each Essential Element that define science content specifications for assessment. They link the Essential Elements (content standards) to the test content itself, including descriptions of each linkage level, key vocabulary, misconceptions and definitions, prerequisite and requisite skills, and accessibility requirements.

Testlet development procedures (Chapter III) followed guidance in the *Standards* (AERA et al., 2014). Item writers were recruited from multiple states in the science consortium and were selected based on their qualifications in academic content areas and/or experience teaching students with the most significant cognitive disabilities. Item writers received comprehensive training and had opportunities for guided practice and feedback throughout the item writing session. Training focused on accessibility, Universal Design for Learning, content development, and bias and sensitivity. The DLM testlets were designed to be accessible to all students in the target population, starting from the first delivered testlets. Item writers were taught to use DLM core vocabulary to minimize unnecessary barriers to student demonstrations of conceptual understanding that might be introduced by using excessively complex vocabulary in items. The vast majority of item writers evaluated the process and their products positively.

Testlets were reviewed (see Chapter III) for content, accessibility, instructional relevance, and bias and sensitivity at multiple points before pilot and field testing. Internal reviews for content and accessibility preceded external reviews by educators from across the consortium. The DLM test development staff considered feedback from all panelists when deciding whether to reject items or revise them before pilot or field testing. External reviews looked at item-level content criteria (alignment, depth of knowledge, quality and appropriateness, accuracy), accessibility (instructional relevance, clarity and appropriateness of images and graphics, minimizing barriers to students with specific needs), and bias/sensitivity (identifying items that require prior knowledge outside the bounds of the targeted content, ensuring fair representation of diversity, avoiding stereotypes and negative naming, removing language that affects a student's demonstration of their knowledge on the measurement target, and removing any language that is likely to cause strong emotional response). The percentage of science items or testlets rated as "accept" ranged across grades and rounds of review from 82% to 91%. The rate at which content was recommended for rejection ranged from approximately 1% to 3% across grades and rounds of review.

The final step of the development phase—pilot and field testing—provided additional content-related evidence (Chapter III). DLM staff used item flagging rules that allowed them to check for the reasonableness of the fungibility assumption that would later be applied in the diagnostic classification model used for scoring (Chapter V). A total of 112 items (21.1% of total) were flagged as needing review by content teams. The procedural evidence presented about the construction of the DLM assessments provides strong evidence of alignment between the definition of the constructs as represented in the Essential Elements and the content of the testlets developed using principles of Universal Design for Learning and evidence centered design.

### ***XI.3.A.i.c Operational Phase***

Chapter IX provides the results of an external alignment study. Overall, the external alignment study provided strong evidence of relationships among the content structures within the DLM assessment system: science standards to Essential Elements, vertical progressions of linkage levels associated with each Essential Element, and item-linkage level relationships. Across all foci, criteria and pools, in 53 of 60 cases (88%) the HumRRO-established criterion was met. The study indicates that students with the most significant cognitive disabilities have access to challenging academic content at each grade level. Areas for improvement include reviewing cognitive process dimension ratings for Target level items that panelists rated as higher than the associated EE, and evaluating overall item alignment when considering both intended dimensions of items (DCI and science and engineering practice). A full written response to the findings is provided in the *CETE Response to Alignment Study* (Appendix G).

### ***XI.3.A.i.d Curriculum Alignment***

Implicit in the intended uses of the DLM results is that the outcomes reflect content the student has had an opportunity to learn. Evidence that students have received instruction in the grade-level Essential Elements supports the use of results for accountability and school evaluation purposes.

Preliminary evidence of students' opportunity to learn the assessed content came from spring 2016 surveys in which teachers estimated the average number of hours they had spent on instruction or hours planned to instruct students on science content within ten topics (see Chapter IX). The majority of educators spent on average between one and ten hours of instruction on most science topics during the 2015-16 school year. The least amount of instructional time was spent on the topics of heredity and biological evolution as well as the scientific practice of engaging in argument from evidence. Overall, evidence suggested that there is an opportunity in the field of science education to provide students with the most significant cognitive disabilities more and better access to science curriculum in their classrooms. With increased opportunities to learn science content and engage in scientific practices, it is anticipated that these students will be better able to demonstrate science academic skills.



Future science development work includes a plan to expand beyond the instructional activities (see Chapter X) to create modules that support teaching the DLM science Essential Elements.

### **XI.3.A.ii. Evidence Based on Response Process**

The interpretation and use of DLM scores depends in part on the validation of whether the cognitive processes that students are engaged in when taking the test match the claims made about the test construct. Evidence is needed to analyze the response processes of test takers in order to determine the fit between the test construct and how students actually experience test content (AERA et al., 2014). Both theoretical and empirical evidence is appropriate and should come from the individual test taker and external observation. Given the cognitive and communication challenges of students with the most significant cognitive disabilities, this category includes procedural evidence as well as empirical evidence that relies on teacher feedback, and, to a lesser extent, student verbalization.

#### ***XI.3.A.ii.a Assessment Design and Development***

Along with procedures and evidence described earlier regarding test content, several aspects of the assessment development process were intended to minimize response barriers and promote construct-relevant interactions with items. For example, as described in Chapter III, the item writing process began with assignment of an Essential Element and EECM and featured training and practice activities that included discussion of how a student might demonstrate the knowledge, skills, or understanding in the nodes included on the EECM. Similarly, item writers were provided with guidance and feedback during the item writing process to promote the production of testlets accessible to the largest number of students possible. Strategies to maximize accessibility of the assessment content and avoid barriers to meaningful student interaction with items included using the DLM core vocabulary, avoiding terminology that could advantage or disadvantage particular students, and consideration of issues that could cause potential barriers for students at every step of the item writing process. Item writers and external reviewers were from diverse backgrounds and different states within the science consortium. Having diverse perspectives represented by external reviewers minimized the chance that students would be disadvantaged due to the inclusion of unnecessary regional or cultural content in testlets. External review panels evaluated items and testlets for accessibility of graphics, clear use of language that minimized the need for inference or prior knowledge, and instructional relevance for students. Additionally, reviewers were asked to judge testlets to be reasonably free of barriers for students with limited working memory, communication disorders, and/or limited implicit understanding of the intentions and emotions of others. The application of these criteria supported the development of content designed to allow all students to interact meaningfully with the assessments.

#### ***XI.3.A.ii.b Fidelity of Administration***

The DLM assessments are intended to be administered with as much standardization as possible and with the expectation that test administrators maintain fidelity to the important aspects of the administration process where flexibility is needed. This balance of

standardization and flexibility is necessary given the heterogeneity of students with the most significant cognitive disabilities. General guidance is provided on these practices through multiple manuals and required test administrator training (see Chapters IV and X). Testlet Information Pages (TIPs; see Chapter IV) support teachers' readiness to deliver specific testlets to specific students with integrity. The majority of respondents to a spring 2016 survey indicated they had confidence in their ability to deliver computer-delivered and teacher-administered testlets (Chapter IV). They also evaluated the Kansas Interactive Testing Engine (KITE) as easy to use to administer testlets.

Test administration observations (Chapter IX) were conducted to further understand response processes for students. Observations were designed to understand whether students were able to interact with the system as intended and to respond to items irrespective of a sensory, mobility, health, communication, or behavioral constraint. The observations provided information on student interaction with testlet contents (e.g., images, figures, engagement activities) and the teacher's actions during administration. Results provided evidence that students were able to communicate their responses through various means and that test administrators accurately captured student responses.

In limited cases during the spring 2016 administration, constancy was compromised by an interruption in the adaptive delivery algorithm (see Chapter IV). The impact of these incidents on score interpretations and inferences was mitigated in most cases by having students revert to the last correctly assigned testlet and resume testing. To support appropriate uses of results for impacted students, the state was provided an incident file (Chapter VII) to assist them in making decisions about how to treat those students' scores within the context of their accountability systems.

### ***XI.3.A.ii.c Accessibility***

Accessibility must be evaluated to identify evidence that the delivery of items and testlets are accessible and appropriate for the full range of students with the most significant cognitive disabilities. Student and test administrator interaction with the KITE system must be evaluated to see if the system provides the necessary supports. Procedures for determining each student's personal needs and executing the correct system features to meet those needs must be in place.

Test administrators recorded accessibility supports in the student's Personal Needs and Preferences profile. To support test administrators in making appropriate decisions about those supports, accessibility was addressed through manuals (Chapter IV), required test administrator training (Chapter X) and additional resources, such as access to released testlets with several simulated students (Chapter IV). Test administration observations revealed that students were able to respond to a task using multiple response modes including verbal, gesture, and eye-gaze. Evidence in support of accessibility was collected by having observers note difficulty with accessibility supports during observations of teacher-administered testlets. Observations of teacher-administered testlets revealed no difficulties with administration.

Surveys of teachers at the end of 2015-2016 test administration provided feedback related to assumptions about accessibility during the assessment process. Three-fourths of teachers



indicated they knew how to use accessibility supports and allowable practices (Chapter IV). Evidence of the effectiveness of these supports was mixed. While 86% agreed that students had access to all needed supports, 76% indicated the student responded to the best of his or her ability, and 65% agreed that the student was able to respond regardless of health, behavior, or disability concerns (Chapter IX). This pattern suggests some students still encounter barriers during the assessment process. It is not known whether those barriers are due to gaps between students' accessibility needs and existing supports in the DLM assessment system, whether students were assessed outside of optimal times (e.g., during behavioral difficulties), or due to other issues.

Where accessibility gaps may be identified due to limited compatibility between types of assistive devices and the KITE system, technology enhancements will be scheduled to improve accessibility. The DLM Consortium has already partnered with the Assistive Technology Industry Association to collect input from manufacturers on compatibility of their devices with KITE Client, and this partnership is expected to continue. More research will be necessary to determine whether students have more opportunities to use those features during instruction in the future, or whether differences may remain because of variations in delivery mode (i.e., instruction delivered directly by the test administrator versus the DLM assessments administered online).

### **XI.3.A.iii. Evidence Based on Internal Structure**

Analyses to support evaluation of evidence based on 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). In this category of evidence, the Essential Elements with three linkage levels provide multi-dimensional representations of content in the academic domains. Reliability analyses describe the consistency of measurement at the linkage level, Essential Element, and overall content area. Additionally, given the heterogeneous nature of the student population and the various and interrelated subgroup categories (e.g., communication mode), differential item functioning (DIF) analyses examine whether particular items function differently for specific subgroups.

#### ***XI.3.A.iii.a Linkage Levels and Statistical Modeling***

The architecture of the DLM Science Alternate Assessment System are the Essential Elements and linkage levels, which are sets of learning targets at varying degrees of complexity aligned to the grade-level expectation. Through input from experts and educators, the Essential Elements and linkage level statements were developed and intended to reflect within EE progression of content in terms of cognitive depth, breadth or complexity as well as across vertical grade band progression of content. The external alignment study (Chapter IX) confirmed that linkage levels within EEs were progressing. Empirical evaluation of the difficulty of testlets administered in the 2015 fall field test also confirmed that higher linkage levels were more difficult than lower linkage levels (Chapter III).

Consistent with the assessment system design, diagnostic classification models are used for statistical modeling. Chapter V provides evidence for the appropriateness of the statistical model, and the score reporting approach used by the DLM system. In addition, evidence provided in Chapter V illustrates how linkage levels can describe mastery at appropriate levels of specificity and are distinct from one another.

### ***XI.3.A.iii.b Reliability***

“[T]he general notion of reliability/precision is defined in terms of consistency over replications of the testing procedure” (AERA et al., 2014, p. 35). Evidence of reliability must show “appropriate evidence of reliability/precision” (AERA et al., 2014, p. 42). Because the DLM Alternate Assessment System uses non-traditional psychometric models (diagnostic classification models) to produce student score reports, evidence for the reliability of scores is based on methods that are commensurate with the models used to produce score reports.

Reliability evidence for the DLM assessments must address the assumption of internal consistency, including decision consistency and accuracy. For the DLM assessments, reliability is provided at multiple levels:<sup>38</sup> (a) performance level reliability; (b) the number of linkage levels mastered within a science domain; (c) the number of total linkage levels mastered; (d) the number of linkage levels mastered within each EE; (e) the mastery status of each of the 102 linkage levels across all EEs; and (f) conditional evidence for each of the three linkage levels. Reliability estimates are provided for three overall metrics: correct classification rate, classification kappa, and correlation between true and estimated values.

As described in Chapter VIII, the reliability summaries for the number of linkage levels mastered within an EE presented reasonable levels of reliability (100% of EEs with polychoric correlations  $\geq .70$ ). All of the classification accuracy rates were  $\geq .80$  and 100% of the kappa values were  $\geq 0.6$ . Similarly, the reliability summaries for mastery classification status of each linkage level showed reasonable levels of reliability (93% of linkage levels with tetrachoric correlations  $\geq .80$ ). While approximately 13% of linkage level kappa values fell below 0.6, all of the classification accuracy rates were  $\geq .80$ . Overall, reliability measures for the DLM Science Alternate Assessment System address the *Standards* (AERA et al., 2014), using methods that were consistent with assumptions of the diagnostic classification model. The analyses yielded evidence to support the argument for internal consistency of the program.

### ***XI.3.A.iii.c Evaluation of Item-Level Bias***

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). Studies that use DIF analyses can uncover internal inconsistency if particular items are functioning differently and systematically for identifiable subgroups of students (AERA et al., 2014). While DIF does not always indicate a weakness in the test items, it can help point to construct-irrelevant variance or

---

<sup>38</sup> Evidence for reliability of results in the content area is presented with proposition #2.

unexpected multidimensionality, thereby contributing to an overall argument for validity and fairness.

As described in Chapter IX, both uniform and a combined model analysis of gender DIF yielded flags for between 7 and 13% of items by grade band, with no flagged items having moderate to large effect sizes. While no items were flagged for moderate to large DIF, the existence of DIF would not necessarily indicate a flaw in the assessment; rather, results serve to inform future steps in the development cycle. For example, items flagged for DIF would be inspected and could be revised or eliminated by content developers. The limited existence of DIF in the current analysis provides additional evidence of strong internal structure.

#### **XI.3.A.iv. Evidence Based on Relationships to Other Variables**

To date, evidence based on relationships to other variables is limited to correlations between student performance in science and the other DLM content areas as well as student performance in science and selected demographic characteristics (see Chapter IX). Overall, the correlational evidence supports the validity claim that DLM assessment results are related to other measures of student achievement and unrelated to student characteristics that should not impact academic achievement.

Additional studies for evaluating external validity evidence are planned for the 2017-18 school year.

#### **XI.3.A.v. Evidence for Consequences of Assessment**

Consequential evidence may be limited in the first year of an operational assessment system as the system has not yet had an opportunity to have an effect. As described in Chapter IX, a spring 2017 survey is planned to assess educators' perceptions of the academic content in the DLM science assessment. The DLM assessments represent a departure from many of the states' previous alternate assessments in the breadth of academic skills assessed. The Essential Elements reflect challenging learning targets for students, while the alternate academic achievement standards set high expectations for achievement; fewer students reached the At Target and Advanced performance levels (see Chapter VII) than on the states' previous alternate assessments. It is expected that educators' survey responses will provide evidence of their awareness that the DLM assessments contain challenging content and reflect high expectations for students.

#### ***XI.3.B. PROPOSITION 2: ACHIEVEMENT LEVEL DESCRIPTORS PROVIDE USEFUL INFORMATION ABOUT STUDENT ACHIEVEMENT***

The DLM approach to standard setting relied on mastery profiles to anchor panelists' content-based judgments to arrive at performance level cut points based on multiple rounds of range finding and pinpointing. Cut points were set to distinguish four performance levels describing student achievement. Grade-level specific performance level descriptors (PLDs) were not used during the standard setting workshop. Instead, they emerged based on the final cut points and were completed after standard setting in 2016.

### **XI.3.B.i. Evidence Based on Content**

Cut points for the four performance levels were determined during the standard setting workshop as described in Chapter VI. Well-qualified panelists fully engaged in a process by which they made use of mastery profiles that summarized linkage level mastery by EE to specify cuts for the total number of linkage levels a student must master to be classified in a performance level. Panelists also relied on content-based evidence when classifying profiles to performance levels, including node description booklets, example items and testlets, and assessment blueprints.

Following specification of cut points for the four performance levels, PLDs were also created for each grade-level or band that cut points were set for during standard setting (Chapter VI). Beginning at the standard setting workshop, and continuing with DLM staff content team development, the specific content being assessed at each linkage level was used to guide the development of the grade-level and band specific PLDs.

Standard setting panelists began the process by drafting lists of skills and understandings that they determined were characteristic of specific performance levels, after establishing cut points. These skills were used as a starting point for the DLM content teams as they developed language for grade-level or band specific descriptions for each performance level. The content team reviewed the EEs, EECMs, and linkage level descriptors on the profiles to determine skills and understandings assessed at the grade level or band. Using multiple sources of information, all anchored in the EEs and the structure of the linkage levels, the content team evaluated the placement of skills into each of the four performance levels. These sources of evidence provide support for the claim that achievement level descriptors provide useful information about student achievement, describing grade-level or band specific content expectations.

### **XI.3.B.ii. Evidence Based on Internal Structure**

As presented in Chapter VIII, performance level reliability indicates consistency of measurement for the assessment as a whole. These statistics are analogous to total score reliability in assessments that use classical or IRT-based models. Reliability evidence was demonstrated by the correlation between true and estimated number of linkage levels mastered, which ranged from .939 to .961. These values indicate that measurement is generally consistent and reveal low measurement error in the total number of linkage levels a student is determined to have mastered, which translates to greater accuracy in assigning students to performance levels. As such, the descriptions of knowledge, skills, and ability typical of students in each performance level has a high likelihood of describing individual students classified to the particular performance level, increasing their utility for meaningful interpretative use by educators and parents.

### **XI.3.B.iii. Evidence for Consequences of Assessment**

In order to establish sound score interpretations and delimit score use, score reports must be useful and provide relevant information for teachers to inform instructional choices and goal setting. Teachers must use results to plan subsequent instruction, and scores can only be

interpreted and used for purposes called out in the theory of action as part of the validity argument.

Assessment results (Chapter VII) were provided to all DLM member states to be reported to parents and to educators at state and local education agencies. Individual reports were provided to teachers and parents. State users received a general research file, which included the student's overall performance level. Individual student score reports also included performance level and a summary of skills the student mastered, resulting in the assignment of the performance level. In addition, aggregated reports were provided to state and local education agencies summarizing student achievement by performance level (Chapter VII). Score reports for the 2016-2017 academic year will include the grade-specific PLDs in place of the bulleted list of skills mastered by domain.

Evidence of intended use of performance level information in score reports is summarized in the research to inform DLM score reports (Chapter IX). Teachers indicated they used the overall performance level when discussing the student's achievement with parents or guardians, but referred to other parts of the score report when planning for instruction. Future research will include usability studies to determine how educators use the overall performance level and the grade/content PLDs, which describe what students in a performance level typically know and can do to inform instructional choices and goal setting.

### ***XI.3.C. PROPOSITION 3: INFERENCES REGARDING STUDENT ACHIEVEMENT, PROGRESS AND GROWTH CAN BE DRAWN AT THE DOMAIN LEVEL***

Individual student score reports (Chapter VII) support interpretation and score use by providing information about student achievement at the domain level. The individual student score report is comprised of two parts: the Performance Profile and the Learning Profile.<sup>39</sup> The Performance Profile, a summary report of individual student results, includes bar graphs indicating the percentage of skills mastered within each domain, as well as a bulleted list of the specific skills mastered in each domain. The Learning Profile, a more fine-grained summary of student mastery of specific knowledge, skills and understandings, includes linkage level mastery reported within each Essential Element. While this proposition also refers to measures of student progress or growth, the consortium has not yet determined whether or how to calculate growth for individual students. The 2015-2016 evidence is delimited to student achievement.

#### **XI.3.C.i. Evidence Based on Content**

Domains organize groups of EEs to support understandings of how students make progress in the content of the domain. The DLM science test blueprints, which specify the EEs assessed, ensure student results reflect performance adequately across the three domains. Specifying

---

<sup>39</sup> Only provided to states participating in the DLM English language arts and mathematics integrated model of assessment.



blueprint requirements at the domain level ensures representation and supports inferences at this level.

### **XI.3.C.ii. Evidence Based on Internal Structure**

As student results are reported at the domain level, it is important to evaluate the reliability evidence for results by domain in order to support inferences regarding student achievement at that level. Reliability evidence for 2015-2016 was calculated at the overall content area, the domain level and at the Essential Element level. The reliability summaries for the number of linkage levels mastered within a science domain showed acceptable levels of reliability (91% of EEs with Pearson correlations  $\geq .70$ ). The classification accuracy values and kappa values were all  $\geq .80$ .

### **XI.3.C.iii. Evidence for Consequences of Assessment**

Validity evidence is necessary to support the assumptions that teachers use score reports to inform instructional choices and goal setting and that score reports are useful and provide relevant information for teachers. Preliminary evidence from score report usability studies described in Chapter IX indicate that teachers refer to the Performance Profile results regarding conceptual areas<sup>40</sup> when explaining reports to parents and when identifying patterns of strength and areas for improvement. Future studies will include usability studies to gain information as to how educators use score report information at that level to guide instruction.

### ***XI.3.D. PROPOSITION 4: ASSESSMENT SCORES PROVIDE USEFUL INFORMATION TO GUIDE INSTRUCTIONAL DECISIONS***

This proposition is especially intended to support the intended use of results to plan instructional priorities and program improvements (use #3). Guiding instructional decisions may be conceptualized as individual student level decisions (i.e., those that teachers might make after receiving a student score report from the previous year) or school/program decisions (e.g., decisions about strategic priorities or curricular changes based on aggregated information). Evidence came from the original design of score reports and interpretive materials, and studies on score report design and interpretation. To support this proposition, there must be evidence that scores are interpreted and used only for their intended purposes, and that teachers can use score reports to inform instructional choices and goal setting. While consequential evidence presented for earlier propositions also supports proposition 4, evidence for this proposition specifically addresses interpretation and use of report contents.

---

<sup>40</sup> The conceptual areas used in ELA and mathematics are used to organize EEs in the same way that domains are used in science to organize the EEs. As described in Chapter IX, the usability studies were originally conducted for ELA and mathematics. As science score reports followed the same template as the ELA and mathematics reports, the findings are presumed to transfer to science score reports. Future score report research will incorporate all three subjects.

### **XI.3.D.i. Evidence for Consequences of Assessment**

As described in Chapter VII, various guiding documents and supporting resources were created to help key stakeholders interpret assessment results as intended. *The Parent Interpretive Guide* provided a sample Individual Student Report to explain how the assessment measures student performance on alternate achievement standards for students with the most significant cognitive disabilities (Dynamic Learning Maps, 2015b). Explanatory letter templates were developed to be used by teachers and state superintendents to introduce the student reports. These letters provide context for the reports including what the DLM assessment is, when it was administered, and what results tell about student performance. A teacher interpretive guide was provided for all those who would discuss results with parents or other stakeholders. The *Scoring and Reporting Guide for Administrators* was designed for principals and district administrators. It covered each type of report provided for the DLM assessments, presented suggestions for how to interpret each report, and suggested uses for the information (Dynamic Learning Maps, 2015c).

As described in Chapter IX, research that informed the development of score reports included qualitative data collection and analysis to understand (1) parents' needs for information in score reports, (2) how stakeholders read and interpret score reports, and (3) how teachers would use assessment results to plan for individual and group instruction. Prototype score reports were developed based on parent perceptions of the challenges with previous alternate assessment score reports. Prototypes were reviewed and refined after multiple rounds of input from parents, educators, and parent advocates. The summative reports contain Performance Profiles and Learning Profiles.

There is preliminary evidence from stakeholder focus groups, teacher interviews, and paired discovery activities (see DLM Score Report Design and Use section in Chapter IX) that stakeholders can read the reports accurately and find them useful. In teacher interviews, the Learning Profile portion of the individual score report was most useful for the purpose of planning instruction, including re-teaching skills. Participants described using score report contents primarily for two parts of Individualized Education Program development: statements on the student's present levels of performance and annual goals. Teachers also tended to use the performance level narrative and mastery skill list nearly verbatim in statements of present levels of performance.

Considering the newness of the DLM assessment system and the length and complexity of information in the individual student score reports, this line of score report research offers strong evidence in support of the proposition that scores provide information that can be used for instructional decision-making. Follow-up studies are planned on teacher decision-making and how score report interpretation translates into actual instructional change, within and across years. Evidence is still needed on score report interpretation by other stakeholder groups, including parents from diverse backgrounds and school administrators, and on the interpretation and use of aggregated reports for decision-making at the school and program levels. To date, this research has been limited to stakeholder interpretation of score reports,



without the use of interpretive resources. Future research will also evaluate the extent to which these resources support appropriate interpretations and uses.

### ***XI.3.E. EVALUATION SUMMARY***

The accumulated evidence available by the end of the 2015-16 year provides preliminary support for the validity argument, particularly at a level that would be expected by the end of the first operational year of an assessment system. Each proposition is addressed by evidence in one or more of the categories of validity evidence, as summarized in Table 87. While many sources of evidence support multiple propositions, Table 87 lists the primary associations. For example, proposition 4 is indirectly supported by content-related evidence described for propositions 1 through 3. Table 88 shows the titles and sections for the chapters cited in Table 87.

Table 87. Dynamic Learning Maps Science Alternate Assessment System Propositions and Sources of Related Evidence for 2015-16

Proposition	Sources of Evidence*				
	Test Content	Response Processes	Internal Structure	Relations with Other Variables	Consequences of Testing
Scores represent what students know and can do.	2, 3, 4, 5, 6, 7, 8, 21, 27	4, 5, 6, 9, 10, 22, 26	1, 2, 7, 11, 20, 21, 23	24	
Achievement level descriptors provide useful information about student achievement.	1, 12, 13, 14, 15		20		16, 17, 18, 25
Inferences regarding student achievement, progress, and growth can be drawn at the domain level.	1, 3, 17		20		25
Assessment scores provide useful information to guide instructional decisions.					19, 25

Note: \* See Table 88 for a list of evidence sources. Only direct sources of evidence are listed. Some propositions are also supported indirectly by evidence presented for other propositions.

Table 88. Evidence Sources Cited in Previous Table

<b>Evidence #</b>	<b>Chapter</b>	<b>Section</b>
1	I	System Components
2	II	Development of the Essential Elements
3	II	Test Blueprints
4	III	Essential Element Concept Maps for Test Development
5	III	Item Writing
6	III	External Reviews
7	III	Pilot Administration
8	III	Field Testing
9	IV	Test Administration Resources and Materials
10	IV	Implementation Evidence from 2015-16 Test Administration
11	V	All
12	VI	Standard Setting Approach
13	VI	Panelists
14	VI	Meeting Procedures
15	VI	Grade Level Performance Level Descriptors
16	VII	Student Performance
17	VII	Score Reports
18	VII	Data Files
19	VII	Score Report Interpretation Resources
20	VIII	Reliability Evidence
21	IX	Evidence Based on Test Content
22	IX	Evidence Based on Response Process
23	IX	Evidence Based on Internal Structure
24	IX	Evidence Based on Relations to Other Variables
25	IX	Evidence Based on Consequences of Testing
26	X	Required Training for Test Administrators
27	X	Instructional Activities for Educators

The overall evaluation of the extent to which each proposition is supported by the evidence collected by 2015-16 is summarized in Table 89.

Table 89. Evaluation of Evidence for Each Proposition

<b>Proposition</b>	<b>Overall Evaluation</b>
1. Scores represent what students know and can do.	There is strong procedural evidence for content representation and response process. Alignment evidence for the operational assessment system is generally strong, although areas for improvement are noted. Evidence of internal structure is strong for this stage of the assessment program; future statistical modeling with additional data will provide stronger evidence.
2. Achievement level descriptors provide useful information about student achievement.	In 2015-16, the policy-level PLDs were reported. Grade-specific PLDs were developed for first use in 2016-17. Procedural evidence supports PLD relationship to the content and structure of the academic content standards. Additional evidence will be needed to evaluate the actual use of the descriptors.
3. Inferences regarding student achievement, progress, and growth can be drawn at the domain level.	There is procedural and empirical evidence to support the structure of the domains and the reporting of achievement in these areas. The consortium is not yet ready to define progress and growth, but will collect and report evidence as appropriate in future technical documentation.
4. Assessment scores provide useful information that can guide instructional decisions.	Overall evidence is strong for the first year of the science program. Interpretive resources support appropriate uses of assessment scores. Based on evidence collected for the ELA and mathematics score report templates (which were adopted for science), stakeholders can interpret report contents and teachers can describe their use for instructional decision-making. Additional evidence is needed as the assessment program matures, including evidence of score use in school and program decision-making.

## **XI.4. CONTINUOUS IMPROVEMENT**

### **XI.4.A. OPERATIONAL ASSESSMENT**

As noted previously in this manual, 2015-16 was the first year the DLM Alternate Assessment System was operational for science. While the 2015-16 assessments were carried out in a manner that supports the validity of the proposed uses of the DLM information for the intended purposes, the consortium is committed to continuous improvement of assessments, teacher and student experiences, and technological delivery of the assessment system. Through evaluation of the DLM English language arts and mathematics assessment administration in prior years, the science assessment benefited from system-wide improvements that were implemented during the 2015-16 administration. This section describes examples of those improvements in test development, administration, and training.

Improvements to test development procedures focus on ensuring accurate, high quality assessment content. The guidelines and procedures for item writing are reviewed annually using multiple sources of information from the field and research findings and data collected throughout the school year. Finally, five-item testlets have been developed and are currently being field tested to replace the three-item testlets. The longer testlets will provide greater support for the interpretation of student results made at the finer-grained linkage level.

Improvements to the 2015-16 test administration procedures focused on ensuring accessibility, accurate delivery of testlet assignments and a high-quality assessment experience for teachers and students. Improvements to synthetic audio were made, with a significant number of testlets receiving updated audio files to support student use of the spoken audio accessibility support. These updates made synthetic audio more consistent across testlets and improved the quality of read-alouds. The quality of color contrast was also enhanced. The *Accessibility Manual* was updated to include improved explanations of supports and the use of accessibility features. Case examples of students with complex needs were included to assist educators with decision-making for students who require a combination of supports and other allowable practices. Information included on the 2015-16 TIPs was also revised based on input from the field. Changes focused on increased usability, logical ordering and specific instructions for educators on how materials are to be used in teacher-administered testlets that require them.

Significant improvements were also made to the 2015-16 required training for test administrators. Project staff and an ad-hoc committee of state partners reviewed the content of the required training. As a result, the training content was streamlined, and DLM staff created differentiated versions for new and returning DLM test administrators. Module post-tests were also improved and a new learning platform was selected, allowing better course design and management features for training modules.

### ***XI.4.B. FUTURE RESEARCH***

The continuous improvement process leads to future directions for research to inform and improve the DLM Science Alternate Assessment System in 2017-18 and beyond. Some areas for investigation have been described earlier in this chapter and throughout the manual.

Over the next few years, we have planned several research studies and analyses. For instance, several initiatives and studies are scheduled or currently underway to provide additional support for the current linkage level scoring model. This includes model fit analyses that are planned to evaluate how well the response data from the DLM science assessment fit the selected latent class statistical model. Model fit will be evaluated using both relative and absolute fit indices. Plans are being developed to flag items for evidence of misfit that the test development team will use to make decisions about operational items and test development priorities.

Other research is also anticipated as sample sizes increase across the second and subsequent years of operational delivery. For example, DIF analyses, which were limited in 2015-2016, may be replicated with different focal and reference groups after the 2016-2017 administration. 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.

In the near future we also anticipate working with states to collect additional, state-level validity evidence. For example, states may collect data (e.g., online progress monitoring) that would be appropriate for use to evaluate the relationship of student responses on DLM assessments to other variables. Since states are responsible for making policy decisions and setting expectations regarding the use of assessment data, they are also well-positioned to provide additional procedural evidence on uses of DLM results for various purposes.

Two additional studies are underway to support the collection of validity evidence. A score report interpretation study is currently in progress to collect information about how teachers read and interpret DLM score report information. The planned study provides an online on-demand tutorial for teachers to view to aid in understanding report contents and their instructional uses. Teacher survey data is also planned for collection during spring 2017 to provide additional data collection for longitudinal survey items specific to subject area as further validity evidence.

Long-term research and development plans are also outlined to support the assessment system and ongoing data collection efforts. For example, professional development modules to support instruction in the science EEs and additional instructional activities will be available in future years, and we will seek stakeholder feedback on those resources. The longitudinal teacher survey beginning in 2017 will also provide evidence of instructional time and alignment of curriculum and instruction with the EEs.

Longitudinal data collection is ongoing as part of the regular operations of the assessment system. As previously mentioned, an annually-administered teacher survey will provide a

source of data from which to investigate changes over time in some of the key assumptions of the validity argument, such as the relationship between accessibility features used during assessment and instruction. Additionally, the survey will provide a means of investigating the long term effects of the assessment system for students and educators. Project staff are planning more intensive studies to collect evidence related to consequences of the assessment system including the extent to which overall system goals are met and negative consequences are avoided.

All future studies will be guided by advice from the DLM Technical Advisory Committee and the state partners, using processes established over the life of the consortium.