Georgia Milestones Assessment System

2019 Operational Technical Report

Georgia Department of Education

DATA RECOGNITION CORPORATION

December 31, 2019
Release Date: July 2020

Developed and published under contract with the Georgia Department of Education by Data Recognition Corporation, 13490 Bass Lake Road, Maple Grove, MN, 55311. Copyright ©2019 by the Georgia Department of Education. Based on a template copyright © by Data Recognition Corporation. All rights reserved. No part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of the Georgia Department of Education and the publisher.

FOREWORD

The *Georgia Milestones Assessment System 2019 Operational Technical Report* documents the processes and procedures that took place to support the implementation of the Georgia Milestones Assessment System (Georgia Milestones) by Data Recognition Corporation (DRC) under the supervision of the Georgia Department of Education (GaDOE).

The technical information herein is intended for use by those who evaluate tests, interpret scores, or use test results in making educational decisions. It is assumed that the reader has technical knowledge of test construction and measurement procedures, as stated 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).

Secure information has been redacted from this report and the related appendices.

TABLE OF CONTENTS

FOREWORD ..... **iii**

TABLE OF CONTENTS ..... **iv**

LIST OF TABLES ..... **vi**

LIST OF FIGURES ..... **x**

EXECUTIVE SUMMARY **1**

CHAPTER 1: INTRODUCTION **3**

1.1 Background on the Georgia Milestones Assessment System 3
1.2 Purpose of the Georgia Milestones Assessments 4
1.3 Design of the Georgia Milestones Assessments 4

CHAPTER 2: ITEM AND TEST DEVELOPMENT **6**

2.1 Operational Development 6
2.2 Field Test Development 18
2.3 Materials Developed to Inform the Public about the Testing Program 30

CHAPTER 3: STANDARDS, STANDARD SETTING, AND STANDARDS VALIDATION **57**

3.1 August 2015 Georgia Milestones Standard Setting Overview 57
3.2 Achievement Level Descriptors 58
3.3 Standard Setting Workshop 58
3.4 Cut Scores for Reading 59
3.5 Policy Review Committee 59
3.6 2017–2018 Georgia Milestones Standards Validation Overview 60
3.7 Updated Achievement Level Descriptors for Science and Social Studies 61
3.8 Content-Based Standards Validation Workshops for Science and Social Studies 61
3.9 Review of the Recommendations from the Content-Based Standards Validation Workshops 62
3.10 Final Reporting Scale for All Tests 64

CHAPTER 4: TEST ADMINISTRATION **66**

4.1 Training of Districts 66
4.2 Ancillary Materials 71
4.3 Test Security Measures 78
4.4 Test Administration 80

CHAPTER 5: PERFORMANCE SCORING **90**

5.1 Scoring of Multiple-Choice Items 90
5.2 Scoring of Constructed-Response Items 90
5.3 Pre-rangefinding 90
5.4 Rangefinding 92
5.5 Field Test Scoring 97
5.6 Operational Scoring 98
5.7 Operational Training Materials 100
5.8 Operational Training Process 100
5.9 Handscoring Process 101
5.10 Handscoring Validity Process 103
5.11 Quality Control 104

CHAPTER 6: OPERATIONAL ANALYSES: KEY CHECKING, CALIBRATION, AND SCALING 106

6.1 Data Collection Design 106
6.2 Operational Analyses 109
6.3 Scaling Methods 127
6.4 Scoring Table Production 129
6.5 Post-equating Verification Results 129

CHAPTER 7: TEST RESULTS 141

7.1 Current Administration Data 141
7.2 Reports 143
7.3 Student Data Files 149
7.4 Score Conversion Tables between Georgia Milestones and Lexile Measures 149
7.5 Interpretive Guide for Score Reports 149

CHAPTER 8: RELIABILITY AND VALIDITY EVIDENCE 150

8.1 Reliability 150
8.2 Validity 169

CHAPTER 9: CONCLUSION AND DISCUSSION 189

9.1 Summary of the Technical Report 189
9.2 Conclusions about Reliability and Validity 192

REFERENCES 196

APPENDIX A: ITEM AND TEST DEVELOPMENT

APPENDIX B: STANDARDS AND STANDARD SETTING

APPENDIX C: TEST ADMINISTRATION

APPENDIX D: END-OF-GRADE OPERATIONAL ANALYSIS: CALIBRATION, SCALING, AND EQUATING

APPENDIX E: END-OF-COURSE OPERATIONAL ANALYSIS: CALIBRATION, SCALING, AND EQUATING

APPENDIX F: END-OF-GRADE SAMPLE REPORTS AND DATA FILE LAYOUTS

APPENDIX G: END-OF-COURSE SAMPLE REPORTS AND DATA FILE LAYOUTS

APPENDIX H: END-OF-GRADE SUMMER OPERATIONAL ANALYSIS: CALIBRATION, SCALING, AND EQUATING

APPENDIX I: END-OF-COURSE SUMMER OPERATIONAL ANALYSIS: CALIBRATION, SCALING, AND EQUATING

APPENDIX J: END-OF-GRADE DOMAIN MASTERY CUT TABLES

APPENDIX K: END-OF-COURSE DOMAIN MASTERY CUT TABLES

APPENDIX L: GEORGIA MILESTONES PERFORMANCE SCORING

LIST OF TABLES

Table 2.1: Test Blueprint for Approximate Percentage of Points by Domain 10

Table 2.2: Elements of Universal Design 14

Table 2.3: Number of Criterion-Reference Anchor Points by Assessment for Spring 2019 EOG and EOC 16

Table 2.4: Items Omitted from Braille Forms within EOG and EOC 2018–2019 Test Administrations 17

Table 2.5: Item Development Scope for the Spring 2019 Field Test 19

Table 2.6: July 2018 Content and Bias Review Participant Count by Grade/Course and Content Area 23

Table 2.7: July 2018 Content and Bias Review Participant Count by Content Area and Region 23

Table 2.8: July 2018 Content and Bias Review Participant Count by Content Area and Ethnicity 24

Table 2.9: July 2018 Content and Bias Review Participant Count by Content Area and Representation of Special Populations (English Learner, Special Education, and Gifted) and by Gender 24

Table 2.10: July 2018 Content and Bias Review Results 25

Table 2.11: Item Data Flagging Criteria 29

Table 2.12: July 2018 Item Data Review Committee Results 30

Table 3.1: Approved Cut Scores 65

Table 3.2: Approved Reading Cut Scores 65

Table 4.1: Schedule for System Test Coordinator Conference 67

Table 4.2: Schedule for Georgia Milestones Pre-administration Workshops 69

Table 4.3: End-of-Course Winter 2018–Summer 2019 Test Administration Windows 81

Table 5.1: Number of Items by Item Type in the 2019 Georgia Milestones Embedded Field Test 91

Table 5.2: Approximate Number of Responses Rangefound and Days Required for the 2019 Georgia Milestone Embedded Field Test 93

Table 5.3: Demographics of the Rangefinding Committee for the 2019 Georgia Milestones EOG Embedded Field Test 94

Table 5.4: Demographics of the Rangefinding Committee for the 2019 Georgia Milestones EOC Embedded Field Test 95

Table 5.5: 2018–2019 Georgia Milestones Operational Scoring Parameters 102

Table 6.1: Structure of the 2018–2019 Georgia Milestones, English Language Arts 107

Table 6.2: Structure of the 2018–2019 Georgia Milestones, Mathematics 108

Table 6.3: Structure of the 2018–2019 Georgia Milestones, Science 108

Table 6.4: Structure of the 2018–2019 Georgia Milestones, Social Studies 109

Table 6.5: 2018–2019 Georgia Milestones Item Difficulty Statistics, English Language Arts 110

Table 6.6: 2018–2019 Georgia Milestones Item Difficulty Statistics, Mathematics 111

Table 6.7: 2018–2019 Georgia Milestones Item Difficulty Statistics, Science 111

Table 6.8: 2018–2019 Georgia Milestones Item Difficulty Statistics, Social Studies 111

Table 6.9: 2018–2019 Georgia Milestones Item Discrimination Statistics, English Language Arts 113

Table 6.10: 2018–2019 Georgia Milestones Item Discrimination Statistics, Mathematics 113

Table 6.11: 2018–2019 Georgia Milestones Item Discrimination Statistics, Science 113

Table 6.12: 2018–2019 Georgia Milestones Item Discrimination Statistics, Social Studies 114

Table 6.13: 2018–2019 Georgia Milestones Item Difficulty Parameter Estimates, English Language Arts 115

Table 6.14: 2018–2019 Georgia Milestones Item Difficulty Parameter Estimates, Mathematics 115

Table 6.15: 2018–2019 Georgia Milestones Item Difficulty Parameter Estimates, Science 116

Table 6.16: 2018–2019 Georgia Milestones Item Difficulty Parameter Estimates, Social Studies .. 116

Table 6.17: 2018–2019 Georgia Milestones Principal Component Analysis Summary, English Language Arts 117

Table 6.18: 2018–2019 Georgia Milestones Principal Component Analysis Summary, Mathematics 117

Table 6.19: 2017–2018 Georgia Milestones Principal Component Analysis Summary, Science 118

Table 6.20: 2017–2018 Georgia Milestones Principal Component Analysis Summary, Social Studies 118

Table 6.21: Summary of Item Residual Correlations for 2018–2019 Georgia Milestones Assessments, English Language Arts 120

Table 6.22: Summary of Item Residual Correlations for 2018–2019 Georgia Milestones Assessments, Mathematics 120

Table 6.23: Summary of Item Residual Correlations for 2018–2019 Georgia Milestones Assessments, Science 121

Table 6.24: Summary of Item Residual Correlations for 2018–2019 Georgia Milestones Assessments, Social Studies 121

Table 6.25: 2019 EOG Item Pairs with Large Residual Correlations (> 0.20) 122

Table 6.26: 2018–2019 EOC Item Pairs with Large Residual Correlations (> 0.20) 123

Table 6.27: 2018–2019 Georgia Milestones Summary of Infit and Outfit Mean Square Statistics, English Language Arts 124

Table 6.28: 2018–2019 Georgia Milestones Summary of Infit and Outfit Mean Square Statistics, Mathematics 124

Table 6.29: 2018–2019 Georgia Milestones Summary of Infit and Outfit Mean Square Statistics, Science 125

Table 6.30: 2018–2019 Georgia Milestones Summary of Infit and Outfit Mean Square Statistics, Social Studies 125

Table 6.31: Classical Item Analysis Flagging Criteria 130

Table 6.32: Sizes for the EOG Calibration for Early Return Student Sample 130

Table 6.33: Sizes for the Winter EOC Calibration for Early Return Student Sample 130

Table 6.34: Sizes for the Spring EOC Calibration for Early Return Student Sample 131

Table 6.35: Reliability Estimates Based on the EOG Calibration for Early Return Student Sample 131

Table 6.36: Reliability Estimates Based on the Winter EOC Calibration for Early Return Student Sample 131

Table 6.37: Reliability Estimates Based on the Spring EOC Calibration Early Return Student Sample 132

Table 6.38: EOG Pre- and Post-equating Achievement Level Summary/Comparison 133

Table 6.39: Winter EOC Pre- and Post-equating Achievement Level Summary/Comparison 133

Table 6.40: Spring EOC Pre- and Post-equating Achievement Level Summary/Comparison..... 133

Table 6.41: 2018–2019 Georgia Milestones Test-Level Descriptive Statistics, English Language Arts 134

Table 6.42: 2018–2019 Georgia Milestones Test-Level Descriptive Statistics, Mathematics 135

Table 6.43: 2018–2019 Georgia Milestones Test-Level Descriptive Statistics, Science..... 136

Table 6.44: 2018–2019 Georgia Milestones Test-Level Descriptive Statistics, Social Studies..... 136

Table 6.45: Content Area and Grade-Level of TerraNova Items Embedded within 2018–2019 Georgia Milestones..... 138

Table 7.1: 2019 Georgia Milestones Summary Statistics for EOG English Language Arts 142

Table 7.2: 2019 Georgia Milestones Summary Statistics for EOG Mathematics 142

Table 7.3: 2019 Georgia Milestones Summary Statistics for EOG Science..... 142

Table 7.4: 2019 Georgia Milestones Summary Statistics for EOG Social Studies..... 142

Table 7.5: Georgia Milestones Summary Statistics for EOC Winter 2018..... 143

Table 7.6: Georgia Milestones Summary Statistics for EOC Spring 2019 143

Table 7.7: Summary of 2018–2019 Georgia Milestones Test Report Types and Delivery Methods to Systems..... 144

Table 8.1: 2018–2019 Georgia Milestones Reliability Statistics by Subgroup, English Language Arts 153

Table 8.2: 2018–2019 Georgia Milestones Reliability Statistics by Subgroup, Mathematics 154

Table 8.3: 2018–2019 Georgia Milestones Reliability Statistics by Subgroup, Science..... 155

Table 8.4: 2018–2019 Georgia Milestones Reliability Statistics by Subgroup, Social Studies..... 156

Table 8.5: 2018–2019 Georgia Milestones SEM by Subgroup, English Language Arts..... 157

Table 8.6: 2018–2019 Georgia Milestones SEM by Subgroup, Mathematics..... 158

Table 8.7: 2018–2019 Georgia Milestones SEM by Subgroup, Science 159

Table 8.8: 2018–2019 Georgia Milestones SEM by Subgroup, Social Studies 160

Table 8.9: 2018–2019 Georgia Milestones CSEM at Cut Scores for English Language Arts Forms 162

Table 8.10: 2018–2019 Georgia Milestones CSEM at Cut Scores for Mathematics Forms 163

Table 8.11: 2018–2019 Georgia Milestones CSEM at Cut Scores for Science Forms 164

Table 8.12: 2018–2019 Georgia Milestones CSEM at Cut Scores for Social Studies Forms..... 164

Table 8.13: Example of Contingency Table with Three Cut Scores..... 165

Table 8.14: 2018–2019 Georgia Milestones Classification Consistency and Accuracy, English Language Arts Forms 167

Table 8.15: 2018–2019 Georgia Milestones Classification Consistency and Accuracy, Mathematics Forms 168

Table 8.16: 2018–2019 Georgia Milestones Classification Consistency and Accuracy, Science Forms 169

Table 8.17: 2018–2019 Georgia Milestones Classification Consistency and Accuracy, Social Studies Forms 169

Table 8.18: DIF Categories for Multiple-Choice Items 173

Table 8.19: DIF Categories for Constructed-Response Items..... 173

Table 8.20: 2019 Georgia Milestones Operational EOG Number of Strong (C) DIF Flags..... 174

Table 8.21: 2018–2019 Georgia Milestones Operational EOC Number of Strong (C) DIF Flags... 175

Table 8.22: 2019 Georgia Milestones EOG List of Operational DIF Flags 176

Table 8.23: 2018–2019 Georgia Milestones EOC List of Operational DIF Flags 177

Table 8.24: 2018–2019 Principal Component Analyses of Georgia Milestones Items—Percentage of Variation Accounted for by First and Second Dimensions 179

Table 8.25: 2019 Correlation between Georgia Milestones EOG Scale Scores and TerraNova Scale Scores 181

Table 8.26: 2018–2019 Correlation between Georgia Milestones EOC Scale Scores and TerraNova Scale Scores 182

LIST OF FIGURES

Figure 2.1: Excerpt from an EOG Georgia Milestones Assessment Guide 32
Figure 2.2: Excerpt from an EOC Georgia Milestones Assessment Guide 33
Figure 2.3: Excerpt of the EOG Content Weights for the 2018–2019 School Year 35
Figure 2.4: Excerpt of the EOC Content Weights for the 2018–2019 School Year 36
Figure 2.5: Sample Table of Contents from an EOG Study/Resource Guide 38
Figure 2.6: Sample Table of Contents from an EOC Study/Resource Guide 39
Figure 2.7: Sample Page with Suggested Instructional Activities from an EOG Study/Resource Guide 40
Figure 2.8: Sample Page with Sample Test Questions from an EOC Study/Resource Guide 41
Figure 2.9: Sample Page of the Parent Brochure 43
Figure 2.10: Table of Contents from the EOG *Interpretive Guide for Score Reports* 45
Figure 2.11: Table of Contents from the EOC *Interpretive Guide for Score Reports* 47
Figure 2.12: Table of Contents from the *Student Assessment Handbook* 50
Figure 2.13: Table of Contents from the *Accommodations Manual* 55
Figure 4.1: Cover Pages of the *Test Administration Manuals* 72
Figure 4.2: Sample Script of the EOG *Test Administration Manual* 74
Figure 4.3: Sample Script of the EOC *Test Administration Manual* 75
Figure 4.4: Notice for Secure Test Materials 79
Figure 4.5: Georgia Milestones EOG Suggested Administration Schedule 82
Figure 4.6: Georgia Milestones EOC Suggested Administration Schedule 84
Figure 4.7: Georgia Milestones Accommodations 87

EXECUTIVE SUMMARY

The *Georgia Milestones Assessment System 2019 Operational Technical Report* documents the processes and procedures that took place to support the implementation of the Georgia Milestones Assessment System (Georgia Milestones) by Data Recognition Corporation (DRC) under the supervision of the Georgia Department of Education (GaDOE). This technical report shows how the processes, procedures, and results of Georgia Milestones relate to issues of validity and reliability and to the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Some principal information presented in this technical report is summarized below.

Purpose: The Georgia Milestones assessments focus on how well students have acquired the knowledge, skills, and abilities specifically described in the Georgia state-mandated content standards for English language arts (ELA), mathematics, science, and social studies. Georgia Milestones includes End-of-Grade (EOG) assessments in ELA and mathematics for grades 3 through 8 and in science and social studies for grades 5 and 8. Georgia Milestones also includes End-of-Course (EOC) assessments for 10 courses offered to high school students. The assessments yield information on academic achievement at the student, class, school, system, and state levels and are used to gauge the overall quality of education throughout Georgia. Georgia Milestones assessments also include a set of embedded TerraNova items that provide national norm-referenced scores in addition to Georgia Milestones scores.

Test Development: The operational items for Georgia Milestones were derived from three sources: items from the previous operational administrations of Georgia Milestones, items from embedded field testing within operational Georgia Milestones administrations or standalone field test administrations, and items from the TerraNova item bank. All items selected for operational administration have been reviewed for content and bias issues by Georgia educators to support the subsequent operational administrations. TerraNova items were also reviewed by Georgia educators, and only the items that were aligned to the Georgia state-mandated content standards counted toward students' Georgia Milestones operational scores.

Administration: The administration of the 2018–2019 Georgia Milestones assessments was conducted throughout the 2018–2019 school year. The fall 2018 EOC mid-month administrations took place in August, September, October, and November. The winter 2018 EOC administration took place between November 26, 2018, and January 4, 2019. The spring 2019 EOC mid-month administrations took place in January, February, and March. The spring 2019 Georgia Milestones administration for the EOG assessments took place between April 8 and May 17, 2019. The spring 2019 Georgia Milestones administration for the EOC assessments took place between April 22 and May 31, 2019. The EOG retest took place between May 13 and July 19, 2019, and the EOC summer administration took place between June 17 and July 19, 2019. A high level of security and control was maintained for all examination materials. Delivery of testing materials was handled by the system and school test coordinators. Two core operational forms were administered online to the student population. Braille and large-print editions were made available to those students who required either of these accommodations.

Operational Analyses: The Georgia Milestones program is based on the application of pre-equating to support the rapid reporting of student results. Pre-equating involves using existing item parameter estimates to construct conversions between the raw scores and the Georgia Milestones scale scores prior to the actual administration of the assessment. The parameters used to build the pre-equated score conversions are derived from field testing and are updated using the state population after the spring operational administration.

A series of post-equating verification analyses were carried out to verify that pre-equated scoring tables could be used to produce student scores. Based on the results of the post-equating verification analyses, the use of the pre-equated scoring tables was determined to be appropriate for all test forms except grade 8 ELA, 9th Grade Literature & Composition, and American Literature & Composition. Post-equated scoring tables were used to report scores on grade 8 ELA, 9th Grade Literature & Composition, and American Literature & Composition and pre-equated scoring tables were used for all other 2019 Georgia Milestones forms.

Results: Raw-score-to-scale-score conversion tables were used to score student results. Electronic Georgia Milestones reports were delivered to systems throughout Georgia.

The GaDOE maintains a Technical Advisory Committee (TAC) to provide guidance on technical assessment matters pertaining to psychometrics, validity and reliability, accuracy, and fairness. Members of the TAC are highly regarded national psychometric experts who have been widely published in their fields. Areas of expertise include assessment design, test equating, computer-based testing, assessment accommodations, and reliability and validity. The GaDOE met regularly with the TAC throughout the design, implementation, administration, and analysis and reporting of the Georgia Milestones assessments.

This technical report provides detailed information related to the items listed above and demonstrates that the processes and procedures followed by Georgia Milestones adhered to appropriate standards and practices of educational assessment. Ultimately, this technical report serves to document evidence that valid inferences about Georgia student performance can be derived from the Georgia Milestones assessments.

CHAPTER 1:
INTRODUCTION

Chapter 1 of this technical report serves to describe the purpose and design of the Georgia Milestones Assessment System (Georgia Milestones) as well as the time frame and highlights of major activities for the current report.

1.1 Background on the Georgia Milestones Assessment System

The Georgia Milestones assessments are criterion-referenced assessments that are designed to measure how well students have acquired the skills and knowledge described in the state-mandated content standards. The specific goals associated with the Georgia Milestones testing program include the following:

- Measuring how well students have acquired the knowledge and skills across the full achievement continuum as described in the Georgia-mandated content standards
- Providing a consistent and coherent signal about student preparedness for the next level, be it the next grade, the next course, college, or a career
- Informing state and federal accountability, including educator effectiveness, at the school, district, and state levels
- Providing a consistent and coherent signal about student achievement both within the system (across grades and courses) and within external measures (NAEP, PSAT, SAT, ACT)
- Ensuring that the assessments are fair for all students, including those with disabilities or limited English proficiency, at all levels of achievement

Georgia Milestones include a series of criterion-referenced assessments in English language arts (ELA), mathematics, science, and social studies, organized in two major components. The End-of-Grade (EOG) component refers to the assessments administered in grades 3 through 8 for ELA and mathematics and in grades 5 and 8 for science and social studies. The End-of-Course (EOC) component refers to 10 course-specific assessments administered to high school students. The EOC tests have been developed for the following courses:

- 9th Grade Literature & Composition (9LCO)
- American Literature & Composition (AMLC)
- Coordinate Algebra (CALG)
- Analytic Geometry (AGEO)
- Algebra I (ALG1)
- Geometry (GEOM)
- Biology (BIOL)
- Physical Science (PHSC)
- United States History (HIST)

- Economics (ECON)

Key features of the Georgia Milestones assessments include the following:

- Integration of reading, language arts, and writing within a single assessment of ELA
- Inclusion of constructed-response items, technology-enhanced items, and multiple-choice items in the ELA and mathematics assessments
- Inclusion of technology-enhanced items and multiple-choice items in the science and social studies assessments
- Inclusion of writing tasks (in response to text)
- Integration of norm-referenced items to provide national norm-referenced comparisons for each student in addition to state-specific criterion-referenced inferences
- Online test administration

Performance on the Georgia Milestones assessments is reported on a scale of measurement specific to each grade/content area or course. Performance on each Georgia Milestones assessment is also classified into one of four achievement levels: *Beginning Learner*, *Developing Learner*, *Proficient Learner*, and *Distinguished Learner*. Scores on the Georgia Milestones EOC assessments are incorporated as final exams that count for 20% of course grades (as defined by State Board of Education Rule 160-4-2-13).

1.2 Purpose of the Georgia Milestones Assessments

The Georgia Milestones assessments are criterion-referenced assessments that are designed to measure how well students have acquired the knowledge and skills across the full achievement continuum as described in the Georgia-mandated content standards. They are intended to provide a consistent and coherent signal about student preparedness for the next level, be it the next grade, the next course, college, or a career. The Georgia Milestones assessments are designed to inform state and federal accountability, including educator effectiveness, at the school, district, and state levels. The assessments provide a consistent and coherent signal about student achievement both within the system (across grades and courses) and within external measures (NAEP, PSAT, SAT, ACT). The assessments are fair for all students, including those with disabilities or limited English proficiency, at all levels of achievement and are used to gauge the overall quality of education throughout Georgia.

1.3 Design of the Georgia Milestones Assessments

The administrations for the Georgia Milestones assessments were held throughout the 2018–2019 school year. One main test administration was held for the EOG assessments in spring 2019. An opportunity to retake the ELA assessment was provided for students in grades 3, 5, and 8, and an opportunity to retake the mathematics assessment was provided for students in grades 5 and 8. Three main test administrations were held for the EOC assessments: winter 2018, spring 2019, and summer 2019. In addition, EOC mid-month administration opportunities were provided for students who needed to test outside of the three main administration windows.

Each main administration of Georgia Milestones included two operational forms administered online. Test designs for the main administrations are in Tables 1.1 through 1.4 below.

Each Georgia Milestones assessment included a subset of 20 items from TerraNova, a norm-referenced achievement test. The incorporation of standardized assessment items produced a test score for each student that can be compared to a specified national reference group, usually of other students of the same grade and age. All 20 TerraNova items contributed to a student's norm-referenced score. Typically, about 10 of these TerraNova items align to state-mandated content standards and serve a dual purpose by also contributing to the Georgia Milestones score. Note that the field test items embedded in the Georgia Milestones assessment did not contribute to a student's total raw score.

One form for each grade/content area or course was brailled to enable students who are visually impaired to participate in the Georgia Milestones administration. In some content areas and at some grade-levels, it was necessary to drop items from the assessment due to difficulties associated with the braille translation. More details about braille versions are included in Section 2.1.6. Note that the use of item response theory (IRT) models to construct Georgia Milestones assessments means that it is possible to drop items from the assessment and still provide scores of comparable quality to the full Georgia Milestones form. More details about the use of IRT in building the Georgia Milestones assessments are included in Chapter 6.

One form for each grade/content area or course was provided as a Video Sign Language (VSL) form for students with an American Sign Language accommodation. The VSL accommodation allowed students to launch a video player within the DRC INSIGHT system for items and passages (if applicable). To receive the VSL accommodation, students needed to have it listed in their Individualized Education Plan.

CHAPTER 2:
ITEM AND TEST DEVELOPMENT

Chapter 2 of this technical report provides a summary of the major test development activities that occurred to create the 2018–2019 operational test forms and the materials developed to inform the public about the testing program. As each major event is presented and discussed, the role of the event in contributing to the validity of the use of the test results is highlighted.

According to the most recent edition of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), "validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). Essential validation evidence supporting the Georgia Milestones program is produced during the test development process. Content-related validation evidence supports inferences from a sample of observations (i.e., the test) to a domain of observations (i.e., the content area). A substantial source of this validation evidence is gathered from expert judgment of whether the test tasks are an adequate and representative sample of the domains being measured. Content-related validity evidence can support interpretations of test scores in terms of performance within and across performance domains. If the content domain is specified clearly and a representative sample of performance tasks is drawn from the domain, then inferences about expected performance over the domain based on observed performances should be legitimate. While validity evidence is necessary to support inferences from test scores, responsibility for the validity of the actual use of the test scores lies with the person or agency using the test scores.

Chapter 2 of this technical report shows the involvement of Georgia educators in the item and test development process, demonstrating adherence to AERA, APA, and NCME (2014) Standards 1.1, 1.11, 4.0, 4.1, 4.2, 4.12, 7.2, 8.4, 12.4, and 12.8. It also shows how test specification documents that were derived from earlier developmental activities guided the final phases of test development and ultimately yielded the test booklets that students used, addressing Standards 3.1, 3.2, 3.5, 4.6, 4.7, 4.8, and 4.10.

This chapter discusses the development activities associated with the Georgia Milestones assessments administered throughout the 2018–2019 school year.

2.1 Operational Development

2.1.1 2018–2019 Test Design

The 2018–2019 Georgia Milestones Assessment System (Georgia Milestones) included a series of criterion-referenced assessments in English language arts (ELA) and mathematics for grades 3 through 8 and high school and in science and social studies for grade 5, grade 8, and high school. Each core assessment included multiple-choice items that reflect key domains within the Georgia state-mandated content standards. The ELA and mathematics core assessments also included constructed-response items, technology-enhanced items, extended constructed-response items, and extended writing-response items (i.e., writing prompts).

The test design for the Georgia Milestones End-of-Grade (EOG) and End-of-Course (EOC) assessments involved using items from multiple sources, including newly developed Georgia Milestones items, items from the TerraNova norm-referenced test (NRT), and items from the existing Georgia Milestones item bank, which contains items previously developed by Georgia

educators for Georgia assessments. The design of the Georgia Milestones assessments included the following elements.

- Number of core forms
  - There were two or three EOG core forms per grade per content area. (Two core forms [i.e., spring A and spring B] were used for spring 2019; in the case of ELA grades 3, 5, and 8 and mathematics grades 5 and 8, one other core form [i.e., summer A] was developed for the summer retest administration.)
    - Core forms were constructed using items that had been previously field tested in the EOG program and in another testing program (i.e., TerraNova).
  - There were six EOC core forms per course. (Two core forms were used for winter 2018 [i.e., winter A and winter B], spring 2019 [i.e., spring A and spring B], and summer 2019 [i.e., summer A and summer B].)
    - Core forms were constructed using items that had been previously field tested in the EOC program and in another testing program (e.g., TerraNova).
    - Summer 2019 core forms were a re-administration of the winter 2018 core forms.
- Item use within the core forms
  - Georgia Milestones items
    - These items counted only toward Georgia Milestones scores.
  - TerraNova items
    - Single-purpose TerraNova items
      - These items counted only toward TerraNova scores.
      - Of the 20 TerraNova items administered for NRT reporting, approximately 10 counted only toward a student's TerraNova score and were not reflected in the Georgia Milestones score.
      - These same 10 (approximately) single-purpose TerraNova items were common to both core forms for a given grade/content area or course.
    - Dual-purpose TerraNova items
      - These items counted toward both Georgia Milestones and TerraNova scores.
      - Of the 20 TerraNova items administered for NRT reporting, approximately 10 counted toward a student's Georgia Milestones score.
      - These same 10 (approximately) dual-purpose TerraNova items were common to both core forms for a given grade/content area or course.
- Anchor plan
  - Anchor items are used to link the different test forms across administrations to a common scale of measurement. For this testing program, two sets of anchor items (i.e., anchor 1 ~ anchor 2) were used to link multiple forms within and across

administrations. (Note that these anchors were all selected from previous operational administrations.)

- Anchor set 1—approximately 10 dual-purpose TerraNova items that were common to all forms for a given administration by grade/content area or course
- Anchor set 2—10 to 15 additional anchor items
  - These were common items between operational forms for a given grade/content area or course.
  - These anchor items needed to have been previously administered in spring 2018 (or in spring 2017, as necessary).
  - If the number of dual-purpose TerraNova items ended up being less than 10, every attempt was made to increase the number of items in anchor set 2 until there were about 25 total linking items.

Note that the TerraNova items that counted toward the Georgia Milestones scores were considered part of the anchor set used to link the Georgia Milestones tests to the operational scale.

- Available pool of items
  - TerraNova items (fixed for the 2019 administrations)—20 items per grade/content area or course
  - Georgia Milestones items from previous spring operational administrations (i.e., spring core A and spring core B)
  - Georgia Milestones items from previous winter operational administrations (i.e., winter core A and winter core B) (EOC only)
  - Items that were field tested in spring 2018
  - Field tested but unused Georgia Milestones items in the Georgia Milestones item bank

2.1.2 Content Standards

The Georgia Milestones program is designed to measure how well students have acquired the skills and knowledge described in the Georgia state-mandated content standards. It provides a consistent and clear signal about student preparedness for the next level, whether it is the next grade or course, college, or a career. The assessments are used to inform state and federal accountability, including educator effectiveness, at the school, system, and state levels. The assessments are designed to provide a clear signal about student achievement within the assessment system and with external measures. The assessments are intended to be fair for all students, including those with disabilities or limited English proficiency, and are used to gauge the overall quality of education throughout Georgia.

At the inception of the Georgia Milestones program, the state-mandated content standards for ELA and mathematics were the College and Career Georgia Performance Standards (CCGPS), and the standards for science and social studies were the Georgia Performance Standards (GPS).

Beginning with the 2015–2016 school year, both the Georgia state-mandated content standards and the Georgia Milestones program transitioned to the Georgia Standards of Excellence (GSE) for ELA and mathematics. The state curriculum followed the CCGPS for ELA and mathematics and the GPS for science and social studies. The transition to the GSE for science and social studies began with the embedded field test in the 2016–2017 school year and culminated with the introduction of operational assessments for science and social studies aligned to the GSE starting with the 2017–2018 school year.

The Georgia Department of Education maintains a website devoted to Georgia's content standards. The website contains a wide range of information, including the standards, content area frameworks, resources and videos, and professional learning tools. More information can be found at https://www.georgiastandards.org/.

2.1.3 Test Blueprint Targets

A critical part of the evidence that supports the use of Georgia Milestones for its intended purposes is based on test content and the extent to which the content domain is represented in the test. According to the *Standards*, content-based evidence "can include logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretation of test scores" (AERA, APA, & NCME, 2014, p. 14). Hence, documentation of the content domain, how the content is sampled and represented, and alignment of items to the content must be well articulated.

Table 2.1 shows the published blueprint for the 2018–2019 administration, indicating the prescribed percentage of points for each content domain for all grade/content areas or courses. In addition to the test blueprints, the test specifications provide details on how the points for each domain are distributed across standards and elements. The operational form for each grade/content area and course matched the blueprint at the domain level.

Table 2.1: Test Blueprint for Approximate Percentage of Points by Domain

Content Area	Grade						Course			
Domain	**3**	**4**	**5**	**6**	**7**	**8**	**EOC 1**	**EOC 2**	**EOC 3**	**EOC 4**
English Language Arts										
Reading and Vocabulary	53	53	53	53	53	53	53	53		
Writing and Language	47	47	47	47	47	47	47	47		
Mathematics										
Operations and Algebraic Thinking	25	20	10							
Number and Operations	35									
Number and Operations in Base 10		20	25							
Number and Operations—Fractions		30	30							
Measurement and Data	30	20	20							
Geometry	10	10	15	18	23	28				
The Number System				30	21					
Ratio and Proportional Relationships				12	19					
Statistics and Probability				17	20	12				
Numbers, Expressions, and Equations						20				
Expressions and Equations				23	17					
Algebra and Functions						40				
Algebra (includes Number and Quantity)							30			
Functions							30		35	
Algebra Connections to Geometry							25			
Algebra Connections to Statistics and Probability							15		15	
Equations									30	
Expressions									20	
Congruence and Similarity								30		35
Circles								15		15
Equations and Measurement								15		35
Expressions, Equations, and Functions (includes Number)								28		
Statistics and Probability								12		15
Science										
Earth Science			23							
Physical Science			35							
Life Science			42							
Matter						28				
Energy						18				
Motion						15				
Waves						24				
Force						15				

Table 2.1: Test Blueprint for Approximate Percentage of Points by Domain (continued)

Content Area	Grade						Course			
Domain	**3**	**4**	**5**	**6**	**7**	**8**	**EOC 1**	**EOC 2**	**EOC 3**	**EOC 4**
Science (cont.)										
Cells							20			
Cellular Genetics and Heredity							23			
Classification and Phylogeny							13			
Ecology							27			
Evolution							17			
Chemistry: Atomic and Nuclear Theory and the Periodic Table								28		
Chemistry: Chemical Reactions and Properties of Matter								22		
Physics: Energy, Force, and Motion								28		
Physics: Waves, Electricity, and Magnetism								22		
Social Studies										
History			58			50				
Geography			12			15				
Government/Civics			15			20				
Economics			15			15				
Colonization through the Constitution							16			
New Republic through Reconstruction							20			
Industrialization, Reform, and Imperialism							16			
Establishment as a World Power							24			
Post-World War II to the Present							24			
Fundamental Economic Concepts								23		
Microeconomic Concepts								23		
Macroeconomic Concepts								21		
International Economics								14		
Personal Finance Economics								19		

Note: Per content area, EOC 1 and EOC 2 respectively refer to 9th Grade Literature & Composition and American Literature & Composition, Coordinate Algebra and Analytic Geometry, Biology and Physical Science, and United States History and Economics. EOC 3 and EOC 4 respectively refer to Algebra I and Geometry.

DRC psychometricians provided test development criteria for the selection of the 2018–2019 operational test forms. The criteria addressed the statistical targets for the anchor item set and the total test for each grade/content area and course. The item selection criteria for the statistical targets are presented here.

- Content guidelines
  - Georgia Milestones test blueprint
    - The whole test (i.e., all items that will be used for a Georgia Milestones score) should match the Georgia Milestones test blueprint. Thus, this criterion excludes the approximately 10 TerraNova items that count only toward the NRT score.
    - The full anchor set should meet the Georgia Milestones blueprint.
      - Anchor 1 + anchor 2 must match the test blueprint.
  - TerraNova test blueprint
    - The set of 20 TerraNova items should not vary from previous administrations.
- Psychometric guidelines
  - TerraNova tests
    - TerraNova Form G was used as the reference test characteristic curve (TCC).
    - As long as TerraNova item sets do not change, there is no need to evaluate the TerraNova selection versus the TCC.
  - Georgia Milestones assessments
    - Target test difficulty means and standard deviations were provided.
      - Targets for EOG and EOC were based on all forms administered until spring 2018.
    - The whole test (i.e., all items that will be used for a Georgia Milestones score) should match Georgia Milestones target difficulties. Thus, this criterion excludes the approximately 10 TerraNova items that count only toward the NRT score.
    - The reading portion of the ELA assessment must also meet the target difficulties. High-stakes decisions are made on the basis of the reading score, so the difficulty of the reading set needs to be consistent across administrations.
    - Ideally, domain difficulty targets should be consistent as well.
      - The full anchor set should meet the Georgia Milestones blueprints.
        - ✓ Anchor 1 + anchor 2 must match targets for test difficulties.
    - For both EOG and EOC selections, avoid items based on the following statistics if possible.
      - too difficult: *p*-value <0.20

        - too easy: *p*-value >0.90
        - not discriminating: ptb <0.15
        - positive discrimination by a distractor
        - high discrimination by a distractor relative to its key
        - high omit (not-reached were included in omit): omit >5%
        - differential item functioning (DIF) C flag: gender and ethnicity
    - TCC, standard error of measurement (SEM), and test information function (TIF) curves should match so that identical scoring tables are created.
- Blueprint
    - The test forms selected should match the established blueprint.
    - Psychometrics should confirm the match upon receipt of preliminary item selections.
    - The linking items should match the blueprint in terms of content representation.
- TerraNova item selection
    - Twenty TerraNova items should be selected for each grade/content area and course.
    - Selected TerraNova items should have a TCC (i.e., true score divided by total raw score) within 0.05 of the corresponding TerraNova Form G TCC.
    - Selected TerraNova items should have content coverage similar to the corresponding TerraNova Form G content coverage.

Table A.1.1 in Appendix A presents the target item difficulty means for 2018–2019 operational items.

2.1.4 Universal Design

Assessments that are universally designed allow participation of the widest possible range of students, resulting in more valid inferences about students' performance. Universally designed assessments may reduce the need for accommodations by decreasing or eliminating access barriers associated with the tests themselves. Table 2.2 presents the elements of universal design (Thompson & Thurlow, 2002). The elements of universal design are relevant to both item development and form construction. This section describes how the elements of universal design were addressed in the construction of the 2018–2019 test forms. This section also provides validity evidence that the assessments have been designed to measure the knowledge and skills across the full achievement continuum described in the content standards and that the assessments are fair for all students at all levels of proficiency.

Table 2.2: Elements of Universal Design

Element	Explanation
Inclusive Assessment Population	Tests designed for state, district, or school accountability must include every student except those in the alternate assessment, and this is reflected in assessment design and field testing procedures.
Precisely Defined Constructs	The specific constructs tested must be clearly defined so that all construct-irrelevant cognitive, sensory, emotional, and physical barriers can be removed.
Accessible, Nonbiased Items	Accessibility is built into items from the beginning, and bias review procedures ensure that quality is retained in all items.
Amenable to Accommodations	The test design facilitates the use of needed accommodations.
Simple, Clear, and Intuitive Instructions and Procedures	All instructions and procedures are simple, clear, and presented in understandable language.
Maximum Readability and Comprehensibility	A variety of readability and plain language guidelines are followed (e.g., sentence length and number of difficult words are kept to a minimum) to produce readable and comprehensible text.
Maximum Legibility	Characteristics that ensure easy decipherability are applied to text, tables, figures, illustrations, and response formats.

Universal design requires that assessments measure the performance of students with a wide range of abilities and skill repertoires, ensuring that students with diverse learning needs receive opportunities to demonstrate competence on the same content. Because field test items are embedded in operational forms and multiple forms are spiraled within classrooms, field test items are administered to students with a wide range of disabilities, to students with limited English proficiency, and to students across racial, ethnic, and socioeconomic lines. A student with a disability or who is an English Learner (EL) may be provided a test administration accommodation based on his or her Individualized Education Program (IEP), Individual Accommodation Plan (IAP), or English Learner/Testing Participation Committee (EL/TPC) Plan. Accommodations in setting, presentation, response, and scheduling (both standard and conditional) are provided in the *Accommodations Manual* (Georgia Department of Education, 2017), which is available on the GaDOE website at http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/Information-For-Educators.aspx. EL students who are enrolled for the first time in a school in the United States may receive a one-time deferment from assessments in ELA and social studies. Braille and large-print forms are provided to students with vision disabilities.

In addition to the NCEO guidelines, in our item and test development process, we also consider the Frameworks for Universal Design for Computer-Based Testing (UD-CBT) and Universal Design for Learning. These guidelines specify how digital technologies can create tests that more accurately assess students who possess a diverse range of physical, sensory, and cognitive abilities and challenges. UD-CBT has been found to level the playing field for students with disabilities and for English language learners. In adherence with the federal Individuals with Disabilities Education Act (IDEA) of 2004, DRC has incorporated universal design principles into the design and development of DRC INSIGHT, ensuring that the system used to deliver Georgia Milestones is accessible to the widest possible range of students. The system was designed to ensure that appropriate accommodations are available to students with disabilities under IDEA, students covered by Section 504, and ELs. INSIGHT makes available universal tools and appropriate accommodations and ensures that Georgia Milestones tests are accessible to students with disabilities and to ELs. The online system is designed to provide tools, available for use by all students, that mirror those used in instructional environments. The

accommodations are appropriate and effective for meeting the individual student's need(s) to participate in the Georgia Milestones program, do not alter the construct being assessed, and allow meaningful interpretations of results and comparison of scores for students who utilize them.

Universal design is present in the accommodations available to students taking a Georgia Milestones test, including:

- Color overlays/high contrast: Background screens or text may be changed to provide the student with high contrast.
- Highlighter: Text in items or passages may be highlighted using an available digital highlighter.
- Magnification: Two additional levels of magnification are available.
- Line guide: Text can be isolated using the available line guide.
- Mask: Content that is not of immediate need or that may be distracting to the student may be blocked off with an available mask.

2.1.5 Item Selection Process

(1) Anchor Sets

Form selection specifications provided the criteria for the selection of sets of linking items used to anchor the operational forms to a common scale of measurement. Note that new scales of measurement were established after the spring 2015 administration. As the test development specialists were designing the 2018–2019 test forms, their goal was to create forms in which the average difficulty of the operational and anchor items selected matched specific psychometric targets and maintained the same blueprint representation as the 2015 administration. Table A.1.1 in Appendix A presents the target item difficulty means for 2018–2019 operational items. Table 2.3 shows the content representation of the anchor items for each grade/content area and course. Note that the number of anchor items included was expressly designed to provide representation in terms of content and item difficulty.

(2) Operational Test Selection

The test development specialists completed the selection of the operational items to fulfill the test blueprint and to meet the psychometric targets as closely as possible. The selections were submitted to the DRC psychometricians for review, additional analyses, and approval. The psychometricians requested changes to the selections, as needed, until the selections were satisfactory and met the psychometric targets as closely as possible. The selection's TCC, TIF, SEM, and raw-score-to-scale-score conversion tables should be as similar as possible to those on the target form.

Table 2.3: Number of Criterion-Reference Anchor Points by Assessment for Spring 2019 EOG and EOC

Content Area	Grade or Course	Points per Assessment					Anchor as Percentage of Core
		Non-anchor Core	Anchor Sets			Total Core	
			One	Two	Total		
English Language Arts	3	34	10	11	21	55	38%
	4	34	10	11	21	55	38%
	5	34	10	11	21	55	38%
	6	34	10	11	21	55	38%
	7	35	10	10	20	55	36%
	8	34	10	11	21	55	38%
	9LCO	34	10	11	21	55	38%
	AMLC	34	10	11	21	55	38%
Mathematics	3	33	9	16	25	58	43%
	4	33	9	16	25	58	43%
	5	33	9	16	25	58	43%
	6	33	9	16	25	58	43%
	7	33	9	16	25	58	43%
	8	33	10	15	25	58	43%
	CALG	33	8	17	25	58	43%
	AGEO	33	9	16	25	58	43%
	ALG1	33	9	16	25	58	43%
	GEOM	33	10	15	25	58	43%
Science	5	35	9	16	25	60	42%
	8	35	6	19	25	60	42%
	BIOL	35	7	18	25	60	42%
	PHSC	35	5	20	25	60	42%
Social Studies	8	35	10	15	25	60	42%
	5	35	9	16	25	60	42%
	HIST	35	9	16	25	60	42%
	ECON	35	8	17	25	60	42%

Note: Anchor set 1 is a dual-purpose set of TerraNova items common to all core forms at a given grade/content area or course. Anchor set 2 is a common set of items between the two core forms (i.e., A and B) within a given administration at a given grade/content area or course.

2.1.6 Braille Edition Process

Form A of each grade/content area and course was brailled to enable visually impaired students to participate in the Georgia Milestones assessments. Two meetings for the spring 2019 test forms were conducted with a committee of teachers of the visually impaired; two meetings were held for EOG (an item review in November 2018 and a proof review in January 2019) and no meetings were held for EOC as the EOC Braille forms were repeats of Braille forms developed and administered in prior years. The committee reviewed the items to determine which items could not be brailled, to make recommendations for how to braille the items, and to develop the teachers' notes that accompany the braille form. The committee subsequently reviewed the braille transcriptions and made recommendations, as needed, for how to modify the transcription.

While the goal is to maximize the number of items on the braille form, it is not always possible to convert all items into braille. In other cases, some items may represent concepts that are not appropriate for students who take the braille form. For some grade/content areas and courses, it was necessary to suppress items from the assessment due to difficulties associated with the braille transcription. Table 2.4 lists the items that were suppressed from the braille forms. Note that the use of item response theory (IRT) models to construct the Georgia Milestones assessments means that it is possible to suppress items from the assessment and still provide scores of comparable quality to the full Georgia Milestones form. More details about the use of IRT in building the Georgia Milestones assessments can be found in Chapter 5.

Table 2.4: Items Omitted from Braille Forms within EOG and EOC 2018–2019 Test Administrations

Program	Grade or Course	Content Area	Item Number	OP or FT	TerraNova Status	
					NRT	CRT
EOG	4	ELA	752745	OP	Yes	No
	8	ELA	752426	OP	Yes	Yes
	8	Science	746760	OP	No	No
EOC	CALG	Mathematics	744475	OP	Yes	No
	ALG1	Mathematics	744475	OP	Yes	No
	GEOM	Mathematics	746104	OP	No	No
	BIOL	Science	744434	OP	Yes	No
	PHSC	Science	744593	OP	Yes	No

Note: Items are classified as operational (OP) or field test (FT). OP items also could be TerraNova (TN) items, and TN items could count for both CRT and NRT purposes or for NRT-only purposes.

2.2 Field Test Development

2.2.1 Assessment of Item Bank for Item Needs

Prior to the initiation of the new item development effort in early 2018 for the items to be field tested in 2019, the contents of the item bank were evaluated, giving consideration to the distribution of items in the item bank across domains, standards, and elements. Item development was primarily focused on developing items for each domain, standard, and element in quantities proportionate to the distribution of score points per the test blueprint and on filling gaps in the item bank. The scope of the field test item development can be found in Table 2.5.

Development for the 2018–2019 Georgia Milestones assessment was based on the test blueprint and item specification documents. The test blueprint identifies the number of items for each content domain (i.e., reporting category), the number and location of the field test items, and the psychometric priorities for the selection of operational items. The item specifications are detailed prescriptions for how items are to be written and may include sample stems and sample items that provide item writers with clear models of phrasing, formatting, and graphical presentations for acceptable test items. These documents describe all content in detail, including areas of classification, types of items, and estimated levels of cognitive complexity (i.e., depths of knowledge [DOKs]). The item specifications provide detailed information regarding the following:

- Eligible item type(s)
- Content domain, standard, and element to be measured
- Clarification statement of the task students will perform when answering each item type
- Stimulus attributes (e.g., stems, graphics, narratives)
- Response attributes (e.g., general, correct response, acceptable distractors, unacceptable distractors)

Table 2.5: Item Development Scope for the Spring 2019 Field Test

Content Area	Grade or Course	Available Field Test Positions (Max Number of FT Positions in Spring 2019)	Percentage of Development (Including Overage)	Number of Items Developed	Number of Items Taken to Content Review in July 2018
English Language Arts	3	65	108%	70	70
	4	39	110%	43	43
	5	65	109%	71	70
	6	39	108%	42	43
	7	39	108%	42	43
	8	65	108%	70	70
	9LCO	65	111%	72	71
	AMLC	65	108%	70	70
Mathematics	3	51	116%	59	59
	4	51	120%	61	61
	5	102	101%	103	103
	6	51	120%	61	61
	7	51	122%	62	62
	8	102	100%	102	102
	CALG	44	111%	49	49
	AGEO	44	116%	51	51
	ALG1	107	106%	113	112
	GEOM	107	104%	111	111
Science	5	50	144%	72	61
	8	50	132%	66	57
	BIOL	100	121%	121	115
	PHSC	100	123%	123	115
Social Studies	5	50	130%	65	65
	8	50	130%	65	64
	HIST	100	102%	102	121
	ECON	100	121%	121	95
Total		**1752**	**113%**	**1,987**	**1,944**

2.2.2 Universal Design Principles

The elements of universal design were addressed throughout the item development and field testing process. The elements of universal design are relevant to both item development and form construction. This section addresses how the elements of universal design were incorporated throughout the item development process and provides validity evidence that supports the use of the assessments to serve their intended purposes.

Regarding precisely defined constructs, the annotated content descriptions identify the Georgia state-mandated content standards and elements that can be assessed. The item specifications also clarify how the items may or may not assess the standard/elements. To ensure that accessibility is incorporated into Georgia Milestones, items are written and edited to have clear, unambiguous language and to remove construct-irrelevant content, context, art, or language. For mathematics problem-solving items, the reading level is targeted to be one or two grade-levels below the grade-level of the item. During item review meetings, participants are trained on how to review items for sensitivity issues to avoid potential sources of bias. In addition to guidelines for reviewing items for content, participants are provided with a list of sensitivity review guidelines. Following field testing, statistical methods of detecting DIF are applied (see Section 8.2.3.1) and participants review the item statistics, including DIF flags, for flagged field test items. At the 2019 data review meetings, participants reviewed data from the spring 2018 embedded field test items and made recommendations to accept or reject items based on the statistics.

The elements of universal design regarding simple, clear, and intuitive instructions and procedures; maximum readability and comprehensibility; and maximum legibility are addressed in the item development and form construction processes. During the item development, editing, and review processes, the grade-level difficulty of vocabulary is checked using *EDL Core Vocabularies* (Taylor, Frackenpohl, & White, 1989), *Children's Writer's Word Book* (Mogilner & Mogilner, 2006), *The Living Word Vocabulary* (Dale, 1976), and *Basic Reading Vocabularies* (Harris & Jacobson, 1982). Reading passages are checked using the Flesch-Kincaid readability formula (Flesch, 1948) as well as other appropriate readability formulas. During item development, consideration is given to the ability of items to be transcribed into braille, particularly the ability of art and graphics to be converted into tactiles. During item content and bias review meetings, training for participants addresses linguistic accessibility, the principles of plain language, and attentiveness to such issues when reviewing items. To ensure maximum legibility and contrast, art specifications define how art is to be rendered; whether or how to use shading and fill patterns in graphics; and the font face, style, and size to be used in tables, graphs, or other types of art.

2.2.3 Passage and Item Writing Phase

2.2.3.1 Summary Demographics of Writers

DRC test development specialists and supervisors selected item and passage writers for the Georgia Milestones program on the basis of their academic backgrounds and experience as classroom teachers in content areas and grade-levels for which they were given item writing assignments. All writers have bachelor's degrees in their content field, in education, or in a related field, and many writers hold advanced degrees. The writers have varied levels of experience in creating and submitting peer-reviewed items for publication in large-scale

assessments. Many of the item and passage writers have previous experience writing items for Georgia Milestones. All passage and item writers were required to sign confidentiality agreements prepared by DRC's Contract Management department.

2.2.3.2 Training of Writers

DRC test development specialists and supervisors provided the item writers with training on the Georgia state-mandated content standards and item specifications. Emphasis was placed on ensuring that writers understood the Georgia state-mandated content standards so that items clearly aligned to content standards, thus contributing to test validity.

All reading passages for new field test items were written or selected specifically for Georgia Milestones by experienced passage writers or passage finders. Reading passage writers were provided with the passage specifications, which define the types and characteristics of passages for each grade-level, allowable lengths of passages, appropriate or inappropriate topics, and bias and sensitivity considerations. Writers were also provided with the content frameworks and a set of guidelines for writing passages, which elaborated on information in the passage specifications and guidance on how to craft passages to ensure that the elements could be assessed. In addition, writers were provided with *The Georgia Milestones Style Guide* (DRC, 2017). DRC made passage writing assignments to ensure appropriate distribution across genres as well as a diversity of topics. The DRC ELA content lead provided writers with assignments, which included grade-levels, passage types, suggested topics, and any requirements for which elements needed to be assessed for a passage. Writers who were writing for Georgia for the first time attended a telephone orientation meeting with the DRC content lead. DRC instructed passage writers on passage specifications to ensure that passages were grade appropriate in content, difficulty, and length. Training for passage writers also consisted of feedback on their draft passages with specific suggestions for revision. DRC submitted sample passages for each grade to the GaDOE for review to ensure that passages conformed to specifications and met expectations for quality.

Item writers were provided with the content frameworks, item specifications, DOK descriptions, item review guidelines checklist, and *The Georgia Milestones Style Guide* (2017). Item writing assignments were generated within DRC's item banking system, the Item Development and Educational Assessment System (IDEAS). Writers were sent one item template for each item to be written, which indicated the content element and DOK level to which the item was to be aligned. Any item writers who were writing for Georgia for the first time were provided training and orientation either in person or over the telephone. When writers submitted new items, the DRC test development specialist reviewed the items to ensure that the criteria for item acceptance were met and that the items aligned to the assigned Georgia state-mandated content standard, element, and DOK level. If an item did not meet these initial acceptance criteria, the item was either rejected or returned to the writer with instructions to revise the item. The writer was provided feedback regarding the reason an item was rejected or how the item needed to be revised to meet the acceptance criteria.

2.2.4 Initial Review for Content Alignment and Sensitivity

New test items were developed using templates designed to capture all item attributes and supporting information, such as content standards, elements, assessment limits, DOK, and content reference documentation. Test items and passages were edited and revised by DRC test development specialists, content supervisors, style editors, and art specialists before being presented to the GaDOE for review.

Item writers adhered to the content frameworks and item specifications as they drafted and revised items. DRC test development specialists also used the item specifications during editorial reviews and revisions of the items. Throughout the item development and review processes, the alignment between each item and the Georgia state-mandated content standards was checked. All test items were carefully reviewed for content and style by DRC test development specialists. All test items developed in 2018 for field testing in spring 2019 were reviewed internally by DRC senior test development specialists, who were familiar with the content frameworks and item specifications, and by GaDOE assessment and curriculum staff. During all item reviews, careful attention was paid to verifying that each item measured the intended Georgia state-mandated content standard. If there was any misalignment, the item was edited to achieve greater alignment or the item was realigned to a different content standard.

After GaDOE-mandated edits were applied to the test items and passages and any misalignment issues were resolved, the items were then presented to the content/bias review committees for review, revision, and approval. For items that were developed to be embedded as field test items in the spring 2019 assessments, the content review meetings were held in July 2018. During training, GaDOE staff presented background information on the Georgia Milestones program and DRC staff provided training materials that detailed the item review principles that were followed for the evaluation of items. The training emphasized the review of items for

- content and grade appropriateness;
- DOK;
- standards alignment to GSE (including new GSE standards for science and social studies);
- bias, fairness, and sensitivity;
- plain language; and
- linguistic accessibility.

DRC developed a PowerPoint presentation for training, and participants were also provided with a list of guidelines for reviewing items for content and sensitivity issues. GaDOE reviewed and approved all training materials. Copies of training materials presented at review meetings can be found in Figures A.1.1 through A.1.3 in Appendix A. Tables 2.6 through 2.9 show the demographic distribution of the content and bias review participants.

Within each content area and grade-level group, DRC facilitators guided the process and recorded the results. A committee member also acted as a scribe to create a second record of the committee results. The two sets of records were compared on a regular basis to ensure continuity of the record keeping. Participants reviewed each item and recommended that items be accepted,

accepted with revision, or rejected. Participants ensured that each item clearly measured a single standard and identified the DOK level for each item. Edits, if needed, were made to the item during the meeting, and final revisions were documented in the review books. Table 2.10 shows the results of the item review meeting.

This set of carefully implemented and documented activities includes the participation of qualified experts and certified, experienced classroom instructional practitioners in content and bias review committees that are tasked with approving, revising, and selecting all test items to be used within the assessments. The activities provide key validation evidence supporting the use of the assessments for their intended purposes.

Table 2.6: July 2018 Content and Bias Review Participant Count by Grade/Course and Content Area

Grade or Course	English Language Arts	Mathematics	Science	Social Studies
3	7	8		
4	7	8		
5	7	8	6	5
6	7	8		
7	7	8		
8	7	8	6	6
9LCO	6			
AMLC	6			
CALG		8		
AGEO		8		
ALG1		8		
GEOM		8		
BIOL			6	
PHSC			6	
HIST				5
ECON				6

Table 2.7: July 2018 Content and Bias Review Participant Count by Content Area and Region

Content Area	Atlanta Metro	North	Central	South	Higher Ed	Statewide	Total
English Language Arts	5	9	7	10	2	0	33
Mathematics	10	8	10	10	2	0	40
Science	5	6	6	7	0	0	24
Social Studies	6	5	6	5	0	0	22
Total	**26**	**28**	**29**	**32**	**4**	**0**	119

Table 2.8: July 2018 Content and Bias Review Participant Count by Content Area and Ethnicity

Content Area	American Indian or Alaska Native	Asian/ Pacific Islander	Black or African American	Hispanic or Latino	Multi-race or Other	White	Decline to State	Total
English Language Arts	0	0	8	1	0	24	0	33
Mathematics	0	1	17	1	0	21	0	40
Science	0	0	2	0	1	20	1	24
Social Studies	0	0	2	0	0	20	0	22
Total	**0**	**1**	**27**	**2**	**1**	**85**	**1**	**119**

Table 2.9: July 2018 Content and Bias Review Participant Count by Content Area and Representation of Special Populations (English Learner, Special Education, and Gifted) and by Gender

Content Area	English Learner	Special Education	Gifted	Female	Male
English Language Arts	10	6	20	30	3
Mathematics	7	10	25	36	4
Science	6	6	16	17	7
Social Studies	3	6	13	12	10
Total	**26**	**28**	**74**	**95**	**24**

Table 2.10: July 2018 Content and Bias Review Results

Content Area	Grade or Course	Number of Items to Review	Number of Items Accepted as Is	Number of Items Revised	Number of Items Rejected	Other	% of Items Accepted	% of Items Rejected
English Language Arts	3	70	52	18	0	0	100%	0%
	4	43	33	10	0	0	100%	0%
	5	70	56	14	0	0	100%	0%
	6	43	29	13	0	0	98%	0%
	7	43	28	14	0	0	98%	0%
	8	70	50	20	0	0	100%	0%
	9LCO	71	41	30	0	0	100%	0%
	AMLC	70	42	28	0	0	100%	0%
Mathematics	3	59	27	32	0	0	100%	0%
	4	61	34	27	0	0	100%	0%
	5	103	54	47	2	0	98%	2%
	6	61	39	21	1	0	98%	2%
	7	62	53	8	1	0	98%	2%
	8	102	77	24	1	0	99%	1%
	CALG	49	31	18	0	0	100%	0%
	AGEO	51	29	21	1	0	98%	2%
	ALG1	112	70	38	4	0	96%	4%
	GEOM	111	81	29	1	0	99%	1%
Science	5	61	46	11	4	0	93%	7%
	8	57	47	10	0	0	100%	0%
	BIOL	115	62	49	4	0	97%	3%
	PHSC	115	63	29	5	0	80%	4%
Social Studies	5	65	40	23	2	0	97%	3%
	8	64	42	21	1	0	98%	2%
	HIST	95	52	41	2	0	98%	2%
	ECON	121	83	37	1	0	99%	1%
Total		**1944**	**1261**	**633**	**30**	**0**	**97%**	**2%**

2.2.5 Other Item Editing and Graphic Creation

Upon completion of items and associated artwork, DRC test development specialists reviewed the items in relation to established criteria. Edits were made to each item as needed to ensure these criteria were met. Criteria for review included the following:

- Item validly measures intended content standard.
- Distractors are plausible, parallel, and mutually exclusive.
- Item has only one correct response.
- Item adheres to item specifications and content limits.
- Item stem, answer choices, and any graphic materials are clear and concise.
- Item corresponds to appropriate level of difficulty.
- Language and content are age appropriate.
- Item is free of bias.
- Content is significant and relevant.

Following in-house content and bias and sensitivity reviews, style editors reviewed all items for grammar, punctuation, and adherence to the GaDOE-approved style sheets and specifications. Items were checked to ensure language was clear and consistent within and across items, formatting was in accordance with agreed-upon type fonts and sizes, words were hyphenated correctly, and emphasis techniques were applied consistently. These types of checks also ensured that test items followed the principles of universal design, such as clear and unambiguous items and art, limited use of shading in art, appropriate size of text in graphics, and avoidance of text on top of shading in graphs.

2.2.6 Field Test Design/Plan

The spring 2019 Georgia Milestones administration consisted of 26 operational assessments. All the items developed in 2018 for the 2019 EOG assessments were field tested. Twelve forms' worth of field test items (per the field test design) were administered for each grade, though not every field test slot contained a unique field test item. Recall that two core forms that fully covered the test blueprint for spring were initially selected for the EOG assessments.

For the EOC assessments, there were also 12 forms' worth of field test items (per the field test design) administered for each course, though not every field test slot contained a unique field test item. Recall that two core forms that fully covered the test blueprint for spring were initially selected for the EOC assessments.

Tables 1.1 and 1.3 in Chapter 1 provide an overview of the 2019 Georgia Milestones field test design.

2.2.6.1 Reading and Evidence-Based Writing Field Test Design/Plan

The fall 2017 Georgia Milestones standalone field test administration for Reading and Evidence-Based Writing (REBW) was conceived to bolster the Georgia Milestones item bank with a new set of operational writing prompt passages and items. Each set consists of two short

stimulus passages, three multiple-choice items, a two-point constructed-response item, and a seven-point essay written in response to a prompt that is based on the stimulus passages.

The viability of the current item bank for the writing section varies from grade to grade, but, overall, the current item bank does not support the sustained rate of use without additional field testing. The standalone field test event bolstered the current Georgia item bank and provided the GaDOE with additional flexibility to administer newly developed test items without requiring increased student testing time during an operational administration.

Each standalone field test form was built to be administered in two sections. The first section consisted of 10 on-grade operational core items with associated passages and functioned as a common link between all standalone field test forms at a given grade. These items were generally selected from the operational items on the spring 2016 operational administration. The second section of the standalone field test form was built to mimic the layout, format, and content of an operational writing prompt test section. It consisted of 3 multiple-choice items, 1 two-point constructed-response item, and 1 seven-point essay. All 5 of these items were linked to 2 short stimulus passages (i.e., a passage set). For each passage set, a total of 6 multiple-choice items were developed, but only 3 appeared on each field test form. Therefore, to ensure that all 6 multiple-choice items were field tested, each passage set appeared in 2 field test forms.

A total of 15 items were field tested per form, including 10 linking multiple-choice items, 3 field test multiple-choice items, 1 field test constructed-response item, and 1 field test writing prompt. A total of 8 items were field tested per passage set, including 6 field test multiple-choice items, 1 field test constructed-response item, and 1 field test writing prompt.

The field test was designed to be administered online only, with no paper/pencil test forms generated. For the EOG assessment, there were 12 field test forms per grade—covering 6 passage sets and 6 writing prompts. For the EOC assessment, there were 24 field test forms per course—covering 12 passage sets and 12 writing prompts.

Since the field test was administered in the fall rather than in the spring (when the traditional operational field test was administered), the EOG test was administered off-grade in grades 4 through 9. The EOC test was administered to students who had successfully completed the given course in the prior school year.

2.2.7 Data Review

This section discusses 2018 field test items. These field test items were developed in 2017 and field tested in 2018; their data were brought to the data review educator committees in July 2018 and were reviewed for their first operational use on tests to be administered during the 2018–2019 school year.

2.2.7.1 Materials and Training Provided to Reviewers

The review of statistics and content alignment of the items that were field tested in spring 2018 to the Georgia state-mandated content standards occurred in July 2018, following the content and bias review of newly developed items to be field tested in spring 2019.

Training for the data review was conducted by a psychometrician from DRC, providing participants with the background necessary to understand the science behind the statistics that

were to be reviewed. The primary focus of this training was to ensure an understanding of the flagging criteria (see the next section for more information).

At the beginning of the data review meeting, participants attended a general training session. A PowerPoint presentation provided information on the topics listed below. As needed, the following topics were clarified by the grade/content area or course facilitator:

- The importance of the data review in the overall item and test development process
- Participants' role in providing recommendations on item use based on the statistics
- Item difficulty, as indicated by *p*-value, and the importance of having items on a test that represent a wide range of difficulty
- Discrimination, as measured by the point biserial correlation, and the importance of having highly discriminating items on a test
- DIF and how it is detected
- Sample items with statistics and how to interpret the data
- Questions to think about when reviewing items flagged for difficulty, discrimination, or DIF
- Procedures for reviewing items and making recommendations

Following training, participants were provided with printed copies of the items as they appeared in the test materials. A copy of the training materials presented at the data review meeting can be found in Figure A.1.3 in Appendix A. Items were presented one per page on data cards, with the standard, element, and answer key identified. Each data card contained the performance statistics of the item. The statistics included the *p*-value, point biserial, and DIF. Participants reviewed all flagged field test items and closely examined each item that had one or more flags. The DRC facilitator guided each committee to compare the results to the flagging criteria. As participants conducted the review, they examined each flagged item to determine why the item was flagged. They recommended to either accept the item as is or reject the item. If items were flagged for mild or high levels of DIF, participants discussed possible reasons the item performed differently for the focal group than it did for the reference group. In addition, review participants had access to the content frameworks. As needed, the participants considered the content alignment and grade-level appropriateness of the items if the item statistics indicated a possible misalignment. A committee member also acted as a scribe to create a second record of the committee results. The two sets of records were compared on a regular basis to ensure continuity of the record keeping. Figure A.1.4 in Appendix A presents a sample of the data card used for the data review. For security reasons, item identifying information has been removed.

2.2.7.2 Flagged Items

Items received flags based on the criteria listed below in Table 2.11. The Mantel-Haenszel (1959) method was used to detect DIF. Further details about DIF can be found in Section 8.2.3.1.

Table 2.11: Item Data Flagging Criteria

Item Type	Criterion	Flag	Borderline Flag
Selected-Response	Item-total correlation for the correct response	Less than 0.1	Between 0.10 and 0.15
	Item-total correlation for any incorrect response	Greater than 0.0	
	p-value	Less than 0.20 or greater than 0.95	
	Any bias code of	Either C- or C+	
Constructed-Response	Any bias code of	Either C- or C+	
	Item-total correlation		Less than 0.1
Technology-Enhanced	Any bias code of	Either C- or C+	
	Item-total correlation		Less than 0.1

Note: The intent of these criteria is to flag everything that should be reviewed. For this purpose, the preference is to over-identify rather than under-identify. Any of these flags should cause the item to be reviewed by content experts, but there are many reasons the experts might want to keep an item in spite of the statistics.

2.2.7.3 Summary Demographics of Data Reviewers

The data review participants attended the content and bias review earlier in the week to review the field test items for spring 2019 and were, therefore, familiar with the process of reviewing items for content and alignment to the Georgia state-mandated content standards. Tables 2.6 through 2.9 show the demographic distribution of the review participants for the July data review.

2.2.7.4 Data Reviewer Results

The summation of the results from the July item data review is shown in Table 2.12. In addition to the review of items flagged, there was a census review of the entire ELA field test. The results of this review are also provided in Table 2.12.

Table 2.12: July 2018 Item Data Review Committee Results

Content Area	Grade or Course	Number of Items to Review	Number of Items Accepted as Is	Number of Items Rejected	% of Items Accepted	% of Items Rejected
English Language Arts	3	24	22	2	92%	8%
	3 (REBW)	48	48	0	100%	0%
	4	10	10	0	100%	0%
	4 (REBW)	48	47	1	98%	2%
	5	19	18	1	95%	5%
	5 (REBW)	47	46	1	98%	2%
	6	12	11	1	92%	8%
	6 (REBW)	46	44	2	96%	4%
	7	10	9	1	90%	10%
	7 (REBW)	48	46	2	96%	4%
	8	16	15	1	94%	6%
	8 (REBW)	45	44	1	98%	2%
	9LCO	21	19	2	90%	10%
	9LCO (REBW)	87	84	3	97%	3%
	AMLC	28	26	2	93%	7%
	AMLC (REBW)	94	91	3	97%	3%
Mathematics	3	8	4	4	50%	50%
	4	4	1	3	25%	75%
	5	9	9	0	100%	0%
	6	7	6	1	86%	14%
	7	8	7	1	88%	13%
	8	7	6	1	86%	14%
	CALG	20	16	4	80%	20%
	AGEO	21	20	1	95%	5%
	ALG1	8	6	2	75%	25%
	GEOM	18	14	4	78%	22%
Science	5	7	4	3	57%	43%
	8	6	4	2	67%	33%
	BIOL	16	6	10	38%	63%
	PHSC	8	1	7	13%	88%
Social Studies	5	6	2	4	33%	67%
	8	7	6	1	86%	14%
	HIST	9	8	1	89%	11%
	ECON	16	11	5	69%	31%
Total		**788**	**711**	**77**	**90%**	**10%**

2.3 Materials Developed to Inform the Public about the Testing Program

Many publications have been developed to disseminate information about Georgia Milestones to a variety of audiences. The GaDOE has authored and published several documents for educators, parents or guardians, and the general public. DRC, as the contractor for Georgia Milestones, has developed other documents, such as study guides, a parent brochure, and a score

interpretation guide. These documents communicate essential information about the background and purpose of Georgia Milestones and the GSE content measured on the tests. The common purpose of the documents described here is to provide a firm foundation for the design, construction, administration, and scoring of the Georgia Milestones assessments. The purpose is also to make the relationships transparent among the GSE content frameworks, the content of the tests, the meaning of the test scores, and the interpretations that can be supported by the test scores.

2.3.1 Assessment Guides

The Georgia Milestones Assessment guides (GaDOE, 2018) are supplemental documents to the Georgia state-mandated content standards that are developed and published by the GaDOE to acquaint Georgia educators with the content coverage of the Georgia Milestones assessments. Only the knowledge, concepts, and skills reflected in the Georgia state-mandated content standards are assessed on Georgia Milestones. While the specific content standards designed for classroom/individual assessment are not included in the Georgia Milestones Assessment guides (2018), the knowledge, concepts, and skills outlined are often required for the mastery of the standards that are assessed. Therefore, the Georgia Milestones Assessment guides (2018) are in no way intended to substitute for the Georgia state-mandated content standards; they are provided to help educators better understand how the curriculum will be assessed. Furthermore, the Georgia Milestones Assessment guides (2018) do not suggest when concepts and skills should be introduced in the instructional sequence; rather, their purpose is to communicate what concepts and skills could be assessed on Georgia Milestones.

The documents are organized by grade/content area or course and content domain. The Georgia state-mandated content standards and related concepts, skills, and abilities that are assessed in each domain are provided. The EOG Georgia Milestones Assessment guides (2018) are available on the GaDOE website at https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/Georgia-Milestones-End-of-Grade-Assessment-Guides.aspx. An excerpt of the table of contents from that document is provided below in Figure 2.1. The EOC Georgia Milestones Assessment guides (2018) are available on the GaDOE website at http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/Georgia-Milestones-End-of-Course-Assessment-Guides.aspx. An excerpt of the table of contents from that document is provided below in Figure 2.2.

Figure 2.1: Excerpt from an EOG Georgia Milestones Assessment Guide

Table of Contents

THE GEORGIA MILESTONES ASSESSMENT SYSTEM 4
- GEORGIA MILESTONES END-OF-GRADE (EOG) ASSESSMENTS 5
- ASSESSMENT GUIDE 6

TESTING SCHEDULE 7
DEPTH OF KNOWLEDGE DESCRIPTORS. 8
SCORES 11
ENGLISH LANGUAGE ARTS (ELA) 12
- DESCRIPTION OF TEST FORMAT AND ORGANIZATION 12
- CONTENT MEASURED 13
- GRADE 5 ENGLISH LANGUAGE ARTS (ELA): DOMAIN STRUCTURES AND CONTENT WEIGHTS . 14
- ITEM TYPES 15
- ENGLISH LANGUAGE ARTS (ELA) DEPTH OF KNOWLEDGE EXAMPLE ITEMS 16
- ENGLISH LANGUAGE ARTS (ELA) ADDITIONAL SAMPLE ITEMS 25
- ENGLISH LANGUAGE ARTS (ELA) ADDITIONAL SAMPLE ITEM KEYS 43
- ENGLISH LANGUAGE ARTS (ELA) EXAMPLE SCORING RUBRICS AND EXEMPLAR RESPONSES 46
- ENGLISH LANGUAGE ARTS (ELA) WRITING RUBRICS 50

MATHEMATICS. 57
- DESCRIPTION OF TEST FORMAT AND ORGANIZATION 57
- CONTENT MEASURED 58
- GRADE 5 MATHEMATICS: DOMAIN STRUCTURES AND CONTENT WEIGHTS 59
- ITEM TYPES 60
- MATHEMATICS DEPTH OF KNOWLEDGE EXAMPLE ITEMS. 61
- MATHEMATICS ADDITIONAL SAMPLE ITEMS 64
- MATHEMATICS ADDITIONAL SAMPLE ITEM KEYS 74
- MATHEMATICS EXAMPLE SCORING RUBRICS AND EXEMPLAR RESPONSES 77

SCIENCE 81
- DESCRIPTION OF TEST FORMAT AND ORGANIZATION 81
- CONTENT MEASURED 82
- GRADE 5 SCIENCE: DOMAIN STRUCTURES AND CONTENT WEIGHTS 83
- ITEM TYPES 83
- SCIENCE DEPTH OF KNOWLEDGE EXAMPLE ITEMS 84
- SCIENCE ADDITIONAL SAMPLE ITEMS 91
- SCIENCE ADDITIONAL SAMPLE ITEM KEYS 105

Copyright © 2018 by Georgia Department of Education. All rights reserved.

Figure 2.2: Excerpt from an EOC Georgia Milestones Assessment Guide

Table of Contents

THE GEORGIA MILESTONES ASSESSMENT SYSTEM 3
GEORGIA MILESTONES END-OF-COURSE (EOC) ASSESSMENTS 4
ASSESSMENT GUIDE 5
TESTING SCHEDULE 6
TEST STRUCTURE 7
DESCRIPTION OF TEST FORMAT AND ORGANIZATION 7
CONTENT MEASURED 8
ITEM TYPES 10
DEPTH OF KNOWLEDGE DESCRIPTORS 11
SCORES 14
EXAMPLE ITEMS 15
ADDITIONAL SAMPLE ITEMS 24
ADDITIONAL SAMPLE ITEM KEYS 40
EXAMPLE SCORING RUBRICS AND EXEMPLAR RESPONSES 43
WRITING RUBRICS 48
APPENDIX: LANGUAGE PROGRESSIVE SKILLS, BY GRADE 55

Copyright © 2018 by Georgia Department of Education. All rights reserved.

2.3.2 Content Weights

The tables showing the percentage of score points for each grade/content area and course by domain (i.e., reporting category) are published on the GaDOE website and are available to educators and the general public. The content weights are similar to the test blueprint tables but show percentages of score points rather than numbers of items. The content weights for the Georgia Milestones assessments are presented in Table A.2.1 in Appendix A.

The EOG content weights are available on the GaDOE website at https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/Georgia-Milestones-EOG-Resources.aspx. The EOC content weights are available on the GaDOE website at https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/Georgia-Milestones-EOC-Resources.aspx. Excerpts of the EOG and EOC content weights documents are provided below in Figures 2.3 and 2.4.

Figure 2.3: Excerpt of the EOG Content Weights for the 2018–2019 School Year

Georgia Milestones
Content Weights

The chart below shows the approximate weights for domains on each Georgia Milestones Assessment in grade 5. Each Georgia Milestones Assessment is aligned to the state-adopted content standards.

GRADE 5		
Content Area	**Domain**	**Approximate Percent of Test**
English Language Arts	Reading and Vocabulary	53%
	Writing and Language	47%
Mathematics	Operations and Algebraic Thinking	10%
	Number and Operations in Base 10	25%
	Number and Operations - Fractions	30%
	Measurement and Data	20%
	Geometry	15%
Science	Earth Science	23%
	Physical Science	35%
	Life Science	42%
Social Studies	History	58%
	Geography	12%
	Government/Civics	15%
	Economics	15%

Georgia Department of Education
July 2018 • Page 3 of 6
All Rights Reserved

Figure 2.4: Excerpt of the EOC Content Weights for the 2018–2019 School Year

Georgia Milestones Assessment System

Georgia Milestones Content Weights

The chart below shows the approximate weights for domains on each Georgia Milestones End of Course Assessment in the content area of mathematics. Each Georgia Milestones Assessment is aligned to the state-adopted content standards.

Mathematics		
End of Course	**Domain**	**Approximate Percent of Test**
Coordinate Algebra	Algebra *(includes Number and Quantity)*	30%
	Functions	30%
	Algebra Connections to Geometry	25%
	Algebra Connections to Statistics and Probability	15%
Analytic Geometry	Congruence and Similarity	30%
	Circles	15%
	Equations and Measurement	15%
	Expressions, Equations, and Functions	28%
	Statistics and Probability	12%

Georgia Department of Education
July 2018 • Page 2 of 5
All Rights Reserved

2.3.3 Study/Resource Guides

The *Study/Resource Guide for Students and Parents* (GaDOE, 2018) is designed for use by parents or guardians to help prepare students for the Georgia Milestones assessments. A guide was developed for each grade-level or course, and each guide focuses on the knowledge and skills that are tested on Georgia Milestones in ELA, mathematics, science, and social studies. The guides present an overview of the Georgia Milestones assessments and test-taking strategies to review with students. The chapters are organized by content. The skills assessed within each domain are reviewed, and the instructional activities designed to build and reinforce the knowledge and skills measured by the Georgia state-mandated content standards are provided. Each chapter includes a practice test with annotated solutions to help assess student progress. Samples of the table of contents from EOG and EOC study/resource guides are shown in Figures 2.5 and 2.6. Sample pages with suggested instructional activities and test questions from the EOG and EOC study/resource guides are shown in Figures 2.7 and 2.8.

In the process of developing the study/resource guides, DRC test development specialists, GaDOE staff, and Georgia educators reviewed the suggested instructional activities and developed the sample test items. Current versions of the study/resource guides for EOG are available on the GaDOE website at http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/EOG-Study-Resource-Guides.aspx. The EOC study/resource guides are available at http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/EOC-Study-Resource-Guides.aspx.

Figure 2.5: Sample Table of Contents from an EOG Study/Resource Guide

Table of Contents

THE GEORGIA MILESTONES ASSESSMENT SYSTEM 3
HOW TO USE THIS GUIDE 4
PREPARING FOR TAKING TESTS 5
OVERVIEW OF THE END-OF-GRADE ASSESSMENT. 6
TYPES OF ITEMS 6
DEPTH OF KNOWLEDGE 8
ENGLISH LANGUAGE ARTS (ELA) 11
DESCRIPTION OF TEST FORMAT AND ORGANIZATION 11
CONTENT 11
ITEM TYPES 11
ENGLISH LANGUAGE ARTS (ELA) DEPTH OF KNOWLEDGE EXAMPLE ITEMS 12
ENGLISH LANGUAGE ARTS (ELA) CONTENT DESCRIPTION AND ADDITIONAL SAMPLE ITEMS . 23
ENGLISH LANGUAGE ARTS (ELA) ADDITIONAL SAMPLE ITEM KEYS 62
ENGLISH LANGUAGE ARTS (ELA) SAMPLE SCORING RUBRICS AND EXEMPLAR RESPONSES . 67
ENGLISH LANGUAGE ARTS (ELA) WRITING RUBRICS 75
MATHEMATICS. 83
DESCRIPTION OF TEST FORMAT AND ORGANIZATION 83
CONTENT 83
ITEM TYPES 83
MATHEMATICS DEPTH OF KNOWLEDGE EXAMPLE ITEMS. 84
MATHEMATICS CONTENT DESCRIPTION AND ADDITIONAL SAMPLE ITEMS 92
MATHEMATICS ADDITIONAL SAMPLE ITEM KEYS 128
MATHEMATICS SAMPLE SCORING RUBRICS AND EXEMPLAR RESPONSES 133
SCIENCE 145
DESCRIPTION OF TEST FORMAT AND ORGANIZATION 145
CONTENT 145
ITEM TYPES 145
SCIENCE DEPTH OF KNOWLEDGE EXAMPLE ITEMS 146
SCIENCE CONTENT DESCRIPTION AND ADDITIONAL SAMPLE ITEMS 150
SCIENCE ADDITIONAL SAMPLE ITEM KEYS 180
SOCIAL STUDIES. 185
DESCRIPTION OF TEST FORMAT AND ORGANIZATION 185
CONTENT 185
ITEM TYPES 185
SOCIAL STUDIES DEPTH OF KNOWLEDGE EXAMPLE ITEMS 186
SOCIAL STUDIES CONTENT DESCRIPTION AND ADDITIONAL SAMPLE ITEMS 189
SOCIAL STUDIES ADDITIONAL SAMPLE ITEM KEYS 213
APPENDIX: LANGUAGE PROGRESSIVE SKILLS, BY GRADE 220

Copyright © 2018 by Georgia Department of Education. All rights reserved.

Figure 2.6: Sample Table of Contents from an EOC Study/Resource Guide

Table of Contents

THE GEORGIA MILESTONES ASSESSMENT SYSTEM 3
GEORGIA MILESTONES END-OF-COURSE (EOC) ASSESSMENTS 4
HOW TO USE THIS GUIDE 5
OVERVIEW OF THE COORDINATE ALGEBRA EOC ASSESSMENT 6
ITEM TYPES 6
DEPTH OF KNOWLEDGE DESCRIPTORS 7
DEPTH OF KNOWLEDGE EXAMPLE ITEMS 10
DESCRIPTION OF TEST FORMAT AND ORGANIZATION 16
PREPARING FOR THE COORDINATE ALGEBRA EOC ASSESSMENT 17
STUDY SKILLS 17
ORGANIZATION—OR TAKING CONTROL OF YOUR WORLD 17
ACTIVE PARTICIPATION 17
TEST-TAKING STRATEGIES 18
PREPARING FOR THE COORDINATE ALGEBRA EOC ASSESSMENT 18
CONTENT OF THE COORDINATE ALGEBRA EOC ASSESSMENT 19
SNAPSHOT OF THE COURSE 20
UNIT 1: RELATIONSHIPS BETWEEN QUANTITIES 21
UNIT 2: REASONING WITH EQUATIONS AND INEQUALITIES 34
UNIT 3: LINEAR AND EXPONENTIAL FUNCTIONS 66
UNIT 4: DESCRIBING DATA 127
UNIT 5: TRANSFORMATIONS IN THE COORDINATE PLANE 161
UNIT 6: CONNECTING ALGEBRA AND GEOMETRY THROUGH COORDINATES 172
COORDINATE ALGEBRA ADDITIONAL PRACTICE ITEMS 187
ADDITIONAL PRACTICE ITEMS ANSWER KEY 206
ADDITIONAL PRACTICE ITEMS SCORING RUBRICS AND EXEMPLAR RESPONSES 210

Copyright © 2018 by Georgia Department of Education. All rights reserved.

Figure 2.7: Sample Page with Suggested Instructional Activities from an EOG Study/Resource Guide

English Language Arts (ELA)

ACTIVITY

The following activity develops skills in Unit 1: Reading Literary Text.

Standards: ELAGSE5RL1, ELAGSE5RL2, ELAGSE5RL3, ELAGSE5W3

Story Time!

Try this activity after reading a story, book, or play with family or friends:

- Choose a character you just read about.
- Make a list of clues the author gave about your character.
- Write your ideas down on paper or on a board.
  You can also use a chart like this one:

Character: _____

Type of Clue	Sentence or Clue from the Story	My Opinion about the Character
Says		
Does		
Worries about		
Looks like		
What others say		

- Share your ideas or charts with your family or friends.

Game: Who Am I?

- Another fun thing to do is to not let anyone know which character you chose.
- After you have completed your notes or chart, play a guessing game.
- Pretend to act like your character, and have your family or friends guess who you are.

Put On a Play

- Write a story idea or plot with family or friends.
- Act out your story as if you were the character you chose. How would your character speak and act in a new situation?

Page 82 of 222 *Georgia Milestones Grade 5 EOG Study/Resource Guide for Students and Parents*

Copyright © 2018 by Georgia Department of Education. All rights reserved.

Figure 2.8: Sample Page with Sample Test Questions from an EOC Study/Resource Guide

Item 1

Selected-Response

Sandra makes necklaces and sells them at a school craft fair. She uses the equation $P = 7.5n - (2.25n + 15)$ to determine her total profit at the fair when n necklaces are sold. Based on this equation, how much does she charge for each necklace?

A. $2.25
B. $7.50
C. $15.00
D. $17.25

Item 2

Selected-Response

The perimeter of a rectangle is $P = 2w + 2l$, where w is the width of the rectangle and l is the length of the rectangle. Rearrange this formula to find the width of the rectangle.

A. $w = P - 2l$

B. $w = \frac{P}{4 - l}$

C. $w = 2P - l$

D. $w = \frac{P}{2} - l$

Copyright © 2018 by Georgia Department of Education. All rights reserved.

2.3.4 Parent Brochures

Each year, a parent brochure is published, in both English and Spanish, that provides an overview of the Georgia Milestones testing program. The GaDOE assessment staff had opportunities to review, provide feedback on, and give final approval of the brochure. The brochure, which was originally published in 2015 and is updated annually, explains the purpose of Georgia Milestones, what scores are reported, and how scores are used. In addition, the brochure offers suggestions for how parents or guardians can help prepare students for the Georgia Milestones assessments and where additional information can be found on the GaDOE website. The first page of the brochure for the 2018–2019 school year is presented in Figure 2.9. The entire brochure is presented in Figure A.2.1 of Appendix A and is available on the GaDOE website at http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/Georgia-Milestones-Assessment-System.aspx.

Figure 2.9: Sample Page of the Parent Brochure

Georgia Milestones Assessment System

End-of-Grade Assessments
Grades 3–8

End-of-Course Assessments
High School

What is the Georgia Milestones Assessment System?

The Georgia Milestones Assessment System (Georgia Milestones) is a comprehensive assessment system spanning grades 3 through high school. Georgia Milestones measures how well students have learned the knowledge and skills outlined in the state-adopted content standards in English Language Arts, Mathematics, Science, and Social Studies. Students in grades 3 through 8 will take End-of-Grade (EOG) assessments in English Language Arts and Mathematics, while grades 5 and 8 students will also take the EOG Science and Social Studies assessments. High school students will take End-of-Course (EOC) assessments for each of the ten courses in which they are enrolled, as designated by the State Board of Education.

What is the purpose of Georgia Milestones?

Georgia Milestones is designed to provide information about how well students are mastering the state-adopted content standards in the core content areas of English Language Arts, Mathematics, Science, and Social Studies. Importantly, Georgia Milestones is designed to provide students with critical information about their own achievement and their readiness for their next level of learning—be it the next grade, the next course, or the next endeavor (college or career).

© Photodisc/PunchStock

What types of questions will a student see on the Georgia Milestones assessments?

Georgia Milestones includes the item types described below:

- open-ended (constructed-response) items in English Language Arts and Mathematics (all grades and courses);
- a writing component (in response to passages read by students) at every grade level and course within the English Language Arts assessments;
- nationally norm-referenced items in all content areas and courses to complement the Georgia criterion-referenced information and to provide a national comparison; and
- multiple-choice items in all content areas and courses.

The mode of administration for the Georgia Milestones program is online. Paper/pencil test materials, such as Braille forms, are available for the small number of students who cannot interact with a computer or device due to their disability.

Where can I find more information about Georgia Milestones?

Talk with your student's teacher or school principal. They can provide specific information about the dates your student will be taking the assessments this year. Resources to help your student prepare for Georgia Milestones are available on the Georgia Department of Education's website at http://testing.gadoe.org. To see what online testing is like, you and your student may visit the Experience Online Testing Georgia website at http://gaexperienceonline.com. The items on the demonstration tests are general and represent multiple grade levels. They do not assess student achievement.

What can I do to help my student?

Students who are prepared, calm, and rested perform better on tests. Here are some of the many ways to help your student approach Georgia Milestones in a relaxed, positive way:

- Encourage your student to employ good study and test-taking skills. These skills include following directions carefully, avoiding careless errors, and reviewing work.
- Explain the purpose of the tests. The assessments give students an opportunity to show what they have learned in school. They also give teachers information that helps them plan instruction.
- Point out that some items may be more difficult than others.
- Be certain your student gets plenty of sleep and has a healthy breakfast and lunch. Taking tests is hard work for many students and can require a lot of energy.
- Be certain your student is at school on time. Rushing and worrying about being late could affect performance on the tests.
- Remember to ask your student about the testing at the end of each day.

Questions and Answers for Parents

DEPARTMENT OF EDUCATION

2018–2019

2.3.5 *Interpretive Guide for Score Reports*

The *Interpretive Guide for Score Reports* (GaDOE, 2018) is written for Georgia teachers and administrators who receive score reports from the 2018–2019 administration of Georgia Milestones. The GaDOE assessment staff had opportunities to review, provide feedback on, and give final approval of the guide. This guide has four sections. The first section provides background information on the Georgia Milestones Assessment System. The second section presents an overview of key terms and test-related concepts. The third section offers general guidelines for interpreting Georgia Milestones scores. The fourth section provides a snapshot and an overview of each score report.

The *Interpretive Guide for Score Reports* (2018) is developed collaboratively by DRC and the GaDOE and updated annually. The tables of contents from the EOG and EOC guides are presented in Figures 2.10 and 2.11, respectively. Current versions of the EOG and EOC score interpretation guides for the Georgia Milestones assessments are presented in Figures A.2.2 and A.2.3 of Appendix A and are available on the GaDOE website at http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/Georgia-Milestones-EOG-Resources.aspx for EOG and http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/Georgia-Milestones-EOC-Resources.aspx for EOC.

Figure 2.10: Table of Contents from the EOG *Interpretive Guide for Score Reports*

Table of Contents

PURPOSE OF THIS GUIDE **1**

BACKGROUND OF GEORGIA MILESTONES **1**
- EOC Administrations 1
- EOG Administrations 1
- Grade Levels and Content Areas Assessed 2
- Alignment to Standards 2
- End-of-Grade Test Contents 2
- Format of Georgia Milestones EOG Assessments 2

KEY TERMS **3**
- Accommodations 3
- Achievement Level 4
- Achievement Level Descriptor (ALD) 4
- Administration Type 4
- Criterion-Referenced Test 5
- Domain 5
- GTID 5
- Lexile® 5
- Lexile® Range 5
- Lexile "Stretch" Bands 5
- Mean Scale Score 5
- Norm-Referenced Scores 6
- Scale Score 6
- Standard Deviation 7
- Standard Error of Measurement (SEM) 7
- Test Form 7

GENERAL GUIDELINES FOR SCORE INTERPRETATION **7**
- Understanding the Use of Scale Scores 7
- Scale Scores and Achievement Levels 8
- Promotion and Retention 9
- Standard Error of Measurement 10
- Students Not Receiving Scale Scores 10
- Students With Conditional Scale Scores 11
- Interpreting Lexile Measures 12
- Interpreting Domain Level Information 14
- Interpreting Group Data in Summary Reports 17

Copyright © 2015–2019 by Georgia Department of Education. All rights reserved.

i

Figure 2.10: Table of Contents from the EOG Interpretive Guide for Score Reports (continued)

Table of Contents

GEORGIA MILESTONES SAMPLE REPORTS WITH ANNOTATIONS....18
Preliminary Reports....19
Individual Student Reports....20
Sample Individual Student Report (Grade 8 example)....22
Sample Individual Student Report (Grade 8 EOG Retest example)....28
Student Labels....32
Sample Student Label (Grade 8 example)....33
Class Roster Reports....34
Sample Class Roster Report (Grade 8 example)....38
Sample Class Roster Report (Grade 8 EOG Retest example)....42
Remediation and Retest Roster Reports....45
Sample Remediation and Retest Roster Report (Grade 8 example)....47
Local Coding Error (LCE) Roster Reports....48
Sample Local Coding Error (LCE) Roster Report....49
Content Area Summary Reports....50
Sample School Content Area Summary Report (Grade 8 English Language Arts example)....53
Sample System Content Area Summary Report (Grade 8 Social Studies example)....55
Summary Reports of All Student Populations....57
Sample State Summary Report (Grade 8 English Language Arts example)....58

Copyright © 2015–2019 by Georgia Department of Education. All rights reserved.

ii

Figure 2.11: Table of Contents from the EOC *Interpretive Guide for Score Reports*

Table of Contents

PURPOSE OF THIS GUIDE **1**

BACKGROUND OF GEORGIA MILESTONES **1**
- EOC Administrations 1
- Courses Assessed 2
- Using the End-of-Course Assessment as a Final Exam 2
- Alignment to Standards 2
- End-of-Course Assessment Contents 2
- Format of Georgia Milestones End-of-Course Assessments 3

KEY TERMS **4**
- Accommodations 4
- Achievement Level 4
- Achievement Level Descriptor (ALD) 4
- Administration Type 5
- Criterion-Referenced Test 5
- Domain 5
- Grade Conversion Score (GCS) 5
- GTID 5
- Lexile® 5
- Lexile® Range 6
- Lexile "Stretch" Bands 6
- Mean Scale Score 6
- Norm-Referenced Scores 6
- Scale Score 7
- Standard Deviation (SD) 7
- Standard Error of Measurement (SEM) 7
- Test Form 7

Copyright © 2015–2019 by Georgia Department of Education. All rights reserved.

i

Figure 2.11: Table of Contents from the EOC Interpretive Guide for Score Reports (continued)

Table of Contents

GENERAL GUIDELINES FOR SCORE INTERPRETATION 8
Understanding the Use of Scale Scores 8
Scale Scores and Achievement Levels 8
Grade Conversion Score (GCS) 9
Standard Error of Measurement (SEM) 10
Students Not Receiving Scale Scores 11
Students With Conditional Scale Scores 12
Interpreting Lexile Measures 12
Interpreting Domain Level Information 14
Interpreting Group Data in Summary Reports 16

GEORGIA MILESTONES SAMPLE REPORTS WITH ANNOTATIONS 17
Preliminary Reports 18
Individual Student Reports 19
Sample Individual Student Report (Ninth Grade Literature and Composition example) 21
Sample Individual Student Report (United States History example) 23
Student Labels 25
Sample Student Labels (Ninth Grade Literature and Composition and United States History examples) 26
Class Roster Reports 27
Sample Class Roster Report (Ninth Grade Literature and Composition example) 29
Sample Class Roster Report (United States History example) 31
Local Coding Error (LCE) Roster Reports 33
Sample Local Coding Error (LCE) Roster Report 34
Content Area Summary Reports 35
Sample School Content Area Summary Report (Ninth Grade Literature and Composition example) 38
Sample System Content Area Summary Report (United States History example) 40
Summary Reports of All Student Populations 42
Sample State Summary Report (United States History example) 43

Copyright © 2015–2019 by Georgia Department of Education. All rights reserved.

ii

2.3.6 *Student Assessment Handbook*

Each year, the GaDOE prepares and publishes the *Student Assessment Handbook* (GaDOE, 2018), which describes the entire Georgia assessment program, not just the Georgia Milestones program. The handbook provides information about the legislative mandates for the assessment program, test security, test administration procedures, assessing students with disabilities and providing testing accommodations, and test preparation. The table of contents is presented in Figure 2.12. The handbook is available on the GaDOE website at http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/Information-For-Educators.aspx.

Figure 2.12: Table of Contents from the *Student Assessment Handbook*

TABLE OF CONTENTS

Table of Contents iii
Introduction to Georgia's Student Assessment Program 1
Glossary 2
Acronyms 7
State Board Policies and Rules 8
Test Security 9
Security and Accountability 9
Security Breaches 9
Testing Irregularities 10
Cell Phone/Electronic Device Use on Georgia Standardized Assessments 13
Calculator Allowances and Restrictions 14
Professional Ethics 15
Georgia Student Assessment Program Responsibilities 17
GaPSC Code of Ethics for Educators (505-6-.01) 17
Test Security Information for School Test Coordinators/Teachers/Examiners 20
Roles and Responsibilities 23
Superintendent 23
System Test Coordinator 23
System Technology Coordinator 25
System Special Education Coordinator 25
System Title III/ESOL Coordinator 26
Principal 27
School Test Coordinator 28
Examiner 29
Proctor 30
Assessment Descriptions 31
Georgia Student Assessment Program Quick Reference Chart: 2018-2019 31
State-Administered Tests 32
Assessing Comprehension and Communication in English State to State for English Language Learners (ACCESS for ELLs 2.0/Alternate ACCESS) 32
Georgia Alternate Assessment 2.0 (GAA 2.0) 34
Georgia Kindergarten Inventory of Developing Skills (GKIDS) 40
GKIDS Readiness Check 42
Georgia Milestones Assessment System (Georgia Milestones) 46
End-of-Grade Measures (EOG) – Grades 3 through 8 47

Georgia Department of Education
Page iii of vi • Rev. November 2018
All Rights Reserved

Figure 2.12: Table of Contents from the *Student Assessment Handbook* (continued)

Georgia Milestones and the Promotion, Placement, and Retention Law 50
End-of-Course (EOC) Measures – Middle/High School 50
National Assessment of Educational Progress (NAEP) 60
Other Assessment Options and Assessment Literacy Resources 63
College Admissions Placement Tests 63
Lexile Scores 65
Formative Instructional Practices Online Professional Learning (FIP) 67
Georgia Online Formative Assessment Resource (GOFAR) 68
General Assessment Administration Guidance 70
Scheduling 70
Test Environment 70
Pre-Administration Preparation 71
During Test Administration 71
Homebound/Hospitalized Students 72
Makeup Sessions 73
Dealing with Emergency/Unexpected Situations 73
Post-Test Administration 75
Transferring Student Test Scores 76
Transferring Records for Students with Disabilities, 504 Students, and EL Students 76
Dissemination of Test Information 76
State Dissemination of Scores 76
Accessing Assessment Data Files 77
Interpreting the Results 78
Public Reporting of Statewide Test Results 78
Local System Dissemination of Scores 79
Rescore Requests 79
Retention of the Files of Test Results and Student Remediation Files 79
Assessing Special Populations 81
Students with Disabilities under the Individuals with Disabilities Education Act (Students with IEPs) 81
Individualized Education Program (IEP) Teams 82
Participation in Assessments 82
Section 504 Students (Students with Disabilities who are not served in special education) 82
Local School System Responsibilities 83
Accommodation Guidelines for Students with Disabilities 83

All Rights Reserved

Figure 2.12: Table of Contents from the *Student Assessment Handbook* (continued)

Test Administration Accommodations Considerations for Students with Disabilities and English Learners 83
Types of Accommodations 83
Standard and Conditional Administrations 84
Modifications vs. Accommodations 84
Additional Considerations for Accommodations 85
Scribing Procedures – Standard Accommodation 86
Scribe Qualifications and Preparation 86
General Scribing Procedures for Online and Paper Administrations (Applicable to Georgia Milestones and ACCESS for ELLs 2.0) 87
English Language Arts Guidelines for Constructed Response Items 87
Mathematics Guidelines for Constructed Response Items 88
Allowable Accommodations for Students with Disabilities 89
Eligibility Guidelines: Reading of English Language Arts (ELA) Passages 91
Guidance for Use of Conditional Accommodation 13: Sign ELA passages. 91
Eligibility Guidelines: Calculator Usage 91
State Required Codes (SRC) for Students with Disabilities and EL students 92
State Directed Codes (SDU-A) 92
Georgia Network for Educational and Therapeutic Support (GNETS) State Directed Codes 92
Participation of English Learners (EL) in State Assessments 93
Data Collections Reporting for Exited ELs 94
Use of Scores of EL Students in Accountability Determinations 95
Accommodation Guidelines for EL Students 95
EL-Test Participation Committee Plan 96
Allowable Accommodations for English Learners 98
Eligibility Guidelines for EL students: Reading of English Language Arts (ELA) passages 99
Guidance for Use of Conditional Accommodations 8: Reading of English Language Arts (ELA) passages 99
Guidance for Special Needs Scholarship Students to Participate in State Assessments (SB10) 100
Test Preparation 103
Preparation of Students 103
Practicing Test-Taking Skills vs. Teaching the Test 103
Inappropriate Use of Test Materials 103
Recommended Test Preparation 104
Practice test-taking strategies. 104
Communication with Students and Parent(s)/Guardian(s) 105
Strategies/Tips for Preparing Students for Test Taking 105

All Rights Reserved

Figure 2.12: Table of Contents from the *Student Assessment Handbook* (continued)

Tips for Students 105
Helpful Web Site for Test Taking Tips and Strategies 106
Students in Home School 107
Required Forms 108
Superintendent's Certification 2018-2019 109
Superintendent's Certification 2018-2019 110
Principal's Certification 111
Testing Irregularities Report Form 112
Statement Report 113
Validation Form for Transcription of Answer Documents 114
Eligibility Criteria for the Georgia Alternate Assessment 2.0 (GAA) 115
Sample Forms/Optional 118
Examiner's Certification of Adherence to Prescribed Test Administration Procedures 119
Participation Documentation for Eligible EL Students 120
Web Resources 122
Important Memoranda and Announcements for System Test Coordinators 122
National Assessment of Educational Progress Resources 123
O.C.G.A. § 20-2-281 124
§ 20-2-281. Student assessments 124
History 128
Notes 129
Commentary 129
Research References & Practice Aids 130

All Rights Reserved

2.3.7 *Accommodations Manual*

The GaDOE prepared and published the *Accommodations Manual: A Guide to Selecting, Administering, and Evaluating the Use of Test Administration Accommodations for Students with Disabilities* (2018), which presents a five-step process to guide the selection and use of accommodations during test administration. The intended audiences include IEP teams, IAP/Section 504 plan committees, educators, and administrators. The manual serves as a guide during the selection, administration, and evaluation of test administration accommodations for students with disabilities. The table of contents is presented in Figure 2.13. The manual is available on the GaDOE website at http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/Information-For-Educators.aspx.

Figure 2.13: Table of Contents from the *Accommodations Manual*

TABLE OF CONTENTS

INTRODUCTION **5**

GENERAL PRINCIPLES GUIDING GEORGIA'S STUDENT ASSESSMENT PROGRAM **6**

Participation of Students with Disabilities in Assessments 6

Local School System Responsibilities 6

Students Eligible for Accommodations 6

Requirements for Use of Accommodations 7

THE FIVE STEP PROCESS **7**

Step 1: Expect Students with Disabilities to Achieve Grade-Level Academic Content Standards . 7

Federal and State Laws Requiring Participation by Students With Disabilities 7

Including All Students with Disabilities in State-mandated Assessments 8

Equal Access to Grade-level Content 8

Step 2: Learn About Accommodations for Instruction and Assessment 9

What are Accommodations? 9

Description of Accommodations Categories 9

Modifications versus Accommodations 10

Step 3: Select Accommodations for Instruction and Assessment for Individual Students 10

Involving Students in Selecting, Using, and Evaluating Test Accommodations 11

Determining the Consequences of Assessment Accommodations Use 11

Standard and Conditional Accommodations 12

Questions To Guide Accommodation Selection 12

Step 4: Administer Accommodations During Instruction and Assessment 13

Accommodations during Instruction 13

Accommodations during Assessment 13

Ethical testing practices 15

Standardization 15

Test Security 15

Step 5: Evaluate and Improve Accommodations Use 15

Questions To Guide Evaluation of Accommodation Use at the School or System Level . 16

Questions To Guide Evaluation at the Student Level 16

All Rights Reserved

Figure 2.13: Table of Contents from the *Accommodations Manual* (continued)

Allowable Accommodations for Georgia's Student Assessment Program 17
SPECIAL CONSIDERATIONS FOR CONDITIONAL ACCOMMODATIONS 20
ACCOMMODATIONS NOT SPECIFIED IN THIS DOCUMENT 21
FACT SHEET 1: CATEGORY OF ACCOMMODATIONS 22
FACT SHEET 2: EXAMPLES OF ACCOMMODATIONS BASED ON STUDENT CHARACTERISTICS 23
FACT SHEET 3: DO'S AND DON'TS WHEN SELECTING ACCOMMODATIONS 27
FACT SHEET 4: GUIDELINES FOR ADMINISTERING SPECIFIC ACCOMMODATIONS 28
Presentation Accommodations 28
Response Accommodations 28
Setting Accommodations 29
Scheduling Accommodations 29
TEACHER TOOL 1: ACCESS NEEDS THAT MAY REQUIRE ACCOMMODATIONS 30
TEACHER TOOL 2: ACCOMMODATIONS FROM THE STUDENT'S PERSPECTIVE 32
TEACHER TOOL 3: ASSESSMENT ACCOMMODATIONS AGREEMENT 34
TEACHER TOOL 4: LOGISTICS PLANNING CHECKLIST 35
APPENDIX 36
Federal and State Laws Requiring Participation by Students with Disabilities 36
Every Student Succeeds Act (ESSA) 36
Individuals with Disabilities Education Improvement Act of 2004 37

All Rights Reserved

CHAPTER 3:
STANDARDS, STANDARD SETTING, AND STANDARDS VALIDATION

Chapter 3 of this technical report provides background information on the standard setting activities and functions to address Standard 5.21 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014): "When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly" (p. 107). In terms of the validity of the Georgia Milestones assessments, it is essential to understand that descriptors and cut scores are established in a collaborative, participatory process largely driven by the input of Georgia teachers and educators. In addition, as cited in the Standards, validity extends to the interpretation of test scores. The descriptors clearly establish, in plain language, the proper frame of reference for understanding how to interpret test scores, particularly cut scores.

The achievement standards for Georgia Milestones, including the achievement level descriptors (ALDs) and cut scores, were established in the summer following the 2014–2015 administration of the tests. The achievement standards for science and social studies were reviewed by Georgia educators after the winter 2017 and spring 2018 administrations of the tests. This section summarizes the findings from the 2015 standard setting and from the 2017–2018 standards validation workshop.

3.1 August 2015 Georgia Milestones Standard Setting Overview

In August 2015, the GaDOE partnered with DRC to conduct a standard setting workshop for the Georgia Milestones Assessment System. Standard setting was conducted in English language arts (ELA), reading, mathematics, science, and social studies for grades 3 through 8 and for eight courses in high school. Committees of Georgia educators and stakeholders worked individually and in concert to recommend achievement standards associated with four achievement levels: *Beginning Learner*, *Developing Learner*, *Proficient Learner*, and *Distinguished Learner*.

A three-phase process was conducted to establish achievement standards for the Georgia Milestones assessments, which yielded achievement standards for ELA, mathematics, science, and social studies:

1) ALDs were developed by Georgia educators and then edited by DRC and the GaDOE. ALDs summarize the knowledge, skills, and abilities expected of students in each achievement level. Initial draft ALDs were developed in a workshop that was held on March 23 and 24, 2015, and the drafts were amplified in a second workshop that was held from July 21 through 24, 2015.
2) Standard setting was conducted from August 24 through 29, 2015. The Bookmark Standard Setting Procedure (Lewis, Mitzel, & Green, 1996; Lewis, Mitzel, Mercado, & Schulz, 2012) was implemented to recommend cut scores for the Georgia Milestones assessments. A total of 271 participants from across the state of Georgia participated in the standard setting workshop.
3) A review committee convened to consider the recommended achievement standards on September 1, 2015. A total of 22 Georgia educators, decision makers, and stakeholders considered the recommendations and made adjustments to promote articulation and consistency.

For a detailed description of the standard setting workshop, see the complete standard setting report, *Georgia Milestones Standard Setting Technical Report* (DRC, 2015).

3.2 Achievement Level Descriptors

The first step in the 2015 standard setting was to develop and approve ALDs, which summarize the knowledge, skills, and abilities expected of students in each achievement level. Specifically, there are four types of ALDs (Egan, Schneider, & Ferrara, 2012), each with a different focus.

- **Policy ALDs** define the GaDOE's vision for each achievement level. Policy ALDs are not specific to any given test; rather, they represent a policy vision for each achievement level.
- **Range ALDs** specify the knowledge, skills, and abilities expected of students in each achievement level on a given test. For example, a range ALD may list the expectations of students at the *Developing Learner* level in grade 5 science. These expectations include those for students who have just entered the *Developing Learner* level, those who are well within the *Developing Learner* level, and those who are nearly (but not quite) at the *Proficient Learner* level.
- **Threshold ALDs** specify the knowledge, skills, and abilities expected of students who are at the point of entry in each achievement level on a given test. For example, a threshold ALD may list the expectations of students who have just enough skill to be considered at the *Proficient Learner* level in grade 8 social studies. While the range ALD specifies the expectations for all students at the *Proficient Learner* level on this test, the threshold ALD seeks to specify the expectations for a student who has just entered the *Proficient Learner* level.
- **Reporting ALDs**, like range ALDs, specify the knowledge, skills, and abilities expected of students in each achievement level on a given test; however, they are designed to communicate this information to stakeholders and educators in the field through score reporting.

Georgia educators convened for two complementary workshops to develop range and threshold ALDs based on the policy ALDs.

3.3 Standard Setting Workshop

The 2015 standard setting workshop took place over a six-day period. The first three days were devoted to mathematics and social studies. The second three days were devoted to ELA and science. The workshop procedures used in each three-day period were the same.

All the materials used at the standard setting workshop were based on test items and results from the spring 2015 administration of the Georgia Milestones assessments. The state-mandated content standards formed the basis for all decisions at the standard setting. These content standards, as adopted by the State Board of Education, detail the knowledge, skills, and abilities that students should be taught in each grade/content area and course. Copies of the content standards were distributed to workshop participants.

The workshop committee was composed of a total of 271 educators, all of whom were recruited, selected, and invited to the workshop by DRC with the guidance and supervision of the GaDOE. The committee comprised a purposeful mix of educators with a variety of backgrounds. Special care was taken to promote geographic diversity among participants, with representation from across the state.

3.4 Cut Scores for Reading

A key requirement of Georgia Milestones is to provide an indication of whether a student is reading below grade-level or at/above grade-level. To identify cut scores to determine reading status, the GaDOE linked the Georgia Milestones reading tests with the MetaMetrics Lexile Framework for Reading. By doing so, the GaDOE acknowledged the state's widespread use of Lexiles to inform inferences in the ELA classroom. DRC and MetaMetrics worked under the direction of the GaDOE to link Georgia Milestones and Lexile scales.

The Lexile stretch bands are associated with the types of challenging, grade-appropriate materials that are associated with the Georgia state-mandated content standards. The Georgia Milestones reading score that is associated with the lower bound of this stretch band in each grade was identified as the cut score for reading. This linking study is documented in a separate technical report produced by MetaMetrics titled *Linking the Georgia Milestones End-of-Grade and End-of-Course Reading and Vocabulary with the Lexile Framework for Reading* (DRC, 2015).

3.5 Policy Review Committee

On September 1, 2015, a separate group of Georgia educators, administrators, and stakeholders convened to consider the recommendations of the standard setting committee. This policy review committee comprised 22 participants, including two participants who also took part in the standard setting committee. The committee consisted of a purposeful mix of educators with a variety of backgrounds well suited to address the standard setting recommendations in terms of the assessment system and the policy perspective. Special care was taken to promote diversity among participants, with representation from across the state. The policy review workshop was facilitated by DRC and took place at the GaDOE offices in Atlanta, Georgia.

The policy review committee had three main goals:

- to study the recommended cut scores from the standard setting to consider their reasonableness, their articulation across grades within each content area, and their articulation across content areas
- to examine the cut scores for reading and consider their reasonableness
- to examine the proposed links between the End-of-Course (EOC) reporting scale and the grade-equivalent metric

This final goal was particular to the policy review committee. State regulations specify that students' performance on the EOC assessment composes a fixed percentage of their classroom grade in the course. To assist schools and systems in equitably incorporating EOC test performance into students' grades, the GaDOE developed a system of correspondence between the EOC score (on the reporting scale) and grade-equivalent score (on the 0–100 scale). Grade-

equivalent scores are derived from the scale score in the following manner for the Georgia Milestones scales: The value of 0 is assigned to the lowest obtainable scale score (LOSS), which differs by content area. The value of 68 is assigned to the cut score for the *Developing Learner* level, which is 475. The value of 80 is assigned to the cut score for the *Proficient Learner* level, which is 525. The value of 92 is assigned to the cut score for the *Distinguished Learner* level, which differs by content area. The value of 100 is assigned to the highest obtainable scale score (HOSS), which differs by content area. Finally, linear transformation is applicable for all possible scale scores between any of the two key points referenced above.

Following the policy review workshop, the GaDOE presented the recommended cut scores to the State Board of Education as part of a specially called board meeting. The cut scores and associated impact data for all tests were approved on September 3, 2015.

3.6 2017–2018 Georgia Milestones Standards Validation Overview

In 2015, the Georgia Standards of Excellence (GSE) replaced the existing Georgia Performance Standards (GPS) for science and social studies. This change was reflected in the Georgia Milestones tests for science and social studies in the 2017–2018 school year. (This transition mirrors that of ELA and mathematics, both of which transitioned from the GPS to the GSE in the 2015–2016 school year.) Accordingly, the test blueprints for the science and social studies assessments changed as a result of alignment to the GSE. In addition, the tests incorporated new item types (i.e., technology-enhanced items) to best measure a given standard.

The test constructs for the eight tests did not change markedly. Although the GSE reflect updated language when discussing the tested content (and, for science, an updated organization of that tested content), the underlying knowledge, skills, and abilities measured by the updated tests are consistent with the existing tests. To be cautious, however, the GaDOE sponsored a standards validation study for the Georgia Milestones science and social studies tests. The purpose of this standards validation was to review the existing achievement standards for science and social studies, to update the ALDs with language from the GSE, and to determine whether the existing cut scores were still valid for continued use on the tests. In this way, the standards validation was an extension of the original 2015 standard setting, as it sought to review the cut scores established during that workshop. The standards validation had three phases, structured to be similar in methodology to the 2015 standard setting.

1) ALDs for science and social studies were developed by Georgia educators and then edited by DRC and the GaDOE. Draft ALDs were developed in a workshop that was held on July 22, 2017. A total of 56 Georgia educators took part in this ALD development process.
2) Content-based standards validation workshops for science and social studies were conducted for the EOC tests on November 7, 2017, and for the End-of-Grade (EOG) tests on March 27, 2018. A modification of the Bookmark Standard Setting Procedure (Lewis, Mitzel, & Green, 1996; Lewis, Mitzel, Mercado, & Schulz, 2012) was implemented to review the cut scores for the Georgia Milestones assessments. A total of 77 participants from across the state of Georgia participated in the content-based standards validation workshops.
3) The GaDOE then reviewed the recommendations from the standards validation workshops and found that Georgia educators' recommendations were highly

consistent with the existing cut scores. Accordingly, the existing cut scores for science and social studies were retained for continued use.

3.7 Updated Achievement Level Descriptors for Science and Social Studies

Using a process similar to the one that was implemented in 2015 to create the initial ALDs for science and social studies, Georgia educators convened to develop ALDs for the updated science and social studies tests. A total of 28 educators focused on the science tests, and another 28 focused on the social studies tests. In each content area committee, participants divided into groups of roughly equal size to focus on the test for grade 5, the test for grade 8, and the two EOC tests.

Educators drafted the Georgia policy ALDs and the updated ALDs using the GSE as a base. The GaDOE then worked with DRC to refine those ALDs for consistency across tests.

Georgia educators convened for two complementary workshops to develop range and threshold ALDs based on the policy ALDs.

3.8 Content-Based Standards Validation Workshops for Science and Social Studies

Two content-based standards validation workshops were held for science and social studies. On November 7, 2017, a content-based standards validation workshop was held for the EOC tests. Specifically, this workshop focused on the tests for biology, physical science, United States history, and economics. A similar workshop was held on March 27, 2018, to focus on the grade 5 and 8 EOG tests.

The materials used at the standards validation workshops were based on test items and results from the most recent administration of the Georgia Milestones assessments. For the EOC standards validation workshop, materials were based on the winter 2017 administration of the assessments. For the EOG workshop, materials were based on the spring 2018 administration.

Thirty-seven Georgia educators took part in the content-based standards validation workshop for the EOC tests, and forty participated in the workshop for the EOG tests. Within each committee, participants divided into groups of roughly equal size to review the cut scores for the tests. The committees were recruited, selected, and invited to the workshop by DRC with the guidance and supervision of the GaDOE. The committee comprised a purposeful mix of educators with a variety of demographic and geographic backgrounds.

At each content-based standards validation workshop, participants were presented with copies of the GSE and the updated ALDs. Participants were also given ordered item booklets (OIBs) that included representations of the existing cut scores. For example, participants in the grade 5 social studies group were shown an OIB with a *Proficient Learner* bookmark on page 50; participants were informed that the existing *Proficient Learner* cut score for the test was associated with a bookmark on that page in the OIB.

During two rounds of discussion and judgments, participants considered whether the existing cut scores were still associated with the knowledge, skills, and abilities expected of students in each achievement level. To accomplish this, participants reviewed the content measured by the items before and after the bookmarks associated with the existing cut scores (termed *benchmarks* during the workshop) and considered how well the content measured by the items before these

bookmarks reflected the knowledge, skills, and abilities specified in the ALDs (specifically the threshold ALDs).

Participants were told that they should retain the existing bookmarks if there was good correspondence between the content knowledge expected of students in each achievement level (as shown in the ALDs) and the content measured by the items before the benchmarks in the OIB. If there was not good correspondence, participants were instructed to recommend an alternative cut score that provided better correspondence; participants did so by moving the bookmarks in the OIB. Participants were asked to record content-based rationales for all their judgments at the workshops.

3.9 Review of the Recommendations from the Content-Based Standards Validation Workshops

The recommendations from the content-based standards validation workshops were reviewed by the GaDOE and DRC after each workshop. Participants' recommendations were found to be consistent with the existing cut scores. Moreover, other validity evidence collected on the existing cut scores suggested that the cut scores were still valid for continued use on the tests.

This section summarizes the findings from that process. Further information can be found in the *Georgia Milestones Standard Setting Technical Report for End-of-Grade and End-of-Course Science and Social Studies* (DRC, 2018).

The GaDOE considered the bookmarks placed by members of the standards validation committees and their relationship with the existing cut scores. In collaboration with DRC, the GaDOE made four primary findings:

- **The content-based expectations for students in each achievement level have not changed significantly since 2015.** The underlying expectations for student achievement in each achievement level have not changed, although the content standards and ALDs have been updated, especially in terms of wording and organization.

  The GaDOE noted that a minority of test items on each of the EOC 2017 tests were brand-new; for both tests, the majority of test items were taken from the existing test pool. The test scales also remained the same in 2017; equating methods were used to place the 2017 test on the existing test scale, which was specified as the baseline in 2015. This process was possible because of the consistency in content-based expectations for students on these tests.

  Lastly, DRC asked participants to record content-based rationales for their bookmark judgments during the standards validation workshops. If participants indicated that there were major differences between the expectations of students in each of the achievement levels that were envisioned in 2015 (as represented by the existing cut scores) and those articulated in 2017 (as summarized by the ALDs), one would expect participants' rationales for their bookmark placements to refer to significant content shifts. Instead, participants' rationales referred only to minor, item-level elements of content. Participants did not appear to note significant shifts in the content-based expectations for students in each achievement level.

- **Participants' bookmark and cut score recommendations were very similar to those from the 2015 standard setting committee.** The final-round recommendations of participants at the standards validation workshops were remarkably similar to those from the final round of the 2015 standard setting.

  When the cut score recommendations from the 2017 standards validation workshops were compared to the 2015 final-round cut score recommendations, there was a great deal of overlap between the two distributions of recommendations. This overlap suggests that participants in 2017–2018 were validating the existing cut scores and that the minor differences between their recommendations and the existing cut scores were due to the random error associated with having a different group of participants make recommendations.

- **The cut scores associated with the median bookmark placements were all very close to one another on the test scale, to the point of being statistically indistinguishable.** The cut scores associated with the median bookmark placements were all within a range of plus-or-minus one-third (±0.33) conditional standard error of measurement (CSEM) values.

  The CSEM quantifies the amount of statistical error associated with the test instrument: If a student were to be tested multiple times, one would expect his or her test scores to fall within a range of ±1.0 CSEM approximately two-thirds of the time. Within this range, it is difficult to argue that scale scores are significantly different. Within a range of ±0.33 CSEM, scale scores are nearly indistinguishable.

- **The impact data observed in winter 2017 and spring 2018 were similar to those from previous administrations of the tests when the existing cut scores were applied.** In 2017–2018, DRC calculated the percentages of students who would be classified in each achievement level based on their 2017–2018 test performance. These impact data were then compared with other administrations of Georgia Milestones. If the expectations for students were markedly different in 2017–2018, or if the underlying test construct was significantly different, then one would expect the impact data to show dramatic differences.

  The impact data associated with the 2017–2018 administrations were not markedly different from those observed in 2015 or 2016. This similarity again suggests that the expectations for students in each achievement level have not changed and that the existing cut scores are valid for continued use.

The GaDOE weighed the small differences between the judgments made at the standards validation and the original standard setting with the relative costs to the testing system of changing cut scores (e.g., impairing existing longitudinal studies).

Based on the totality of the evidence, the GaDOE determined that there were no significant differences between the content assessed in the new science and social studies tests and the content assessed in the previous tests. The GaDOE also determined that the differences between the judgments made at the 2017–2018 standards validation workshops and the original 2015 cut scores could be attributed to random statistical error.

As a matter of policy and based on the available evidence, the GaDOE decided that the existing cut scores in these eight tests of science and social studies were still valid for continued use.

3.10 Final Reporting Scale for All Tests

Following a long-standing tradition by the GaDOE on Georgia's large-scale assessments, significant cut scores are represented by fixed points on the test scale for each assessment. Specifically, the test scale for each assessment was created so that the *Developing Learner* cut score is always equal to 475 and the *Proficient Learner* cut score is always equal to 525.

To create the final reporting scale for each assessment, linear transformation constants (additive and multiplicative) were calculated to transform each temporary test scale into the final reporting scale. As such, the cut scores had the same interpretations in terms of relative value and content-based expectations. However, the *Developing Learner* and *Proficient Learner* values were tied to the values listed above. The *Distinguished Learner* cut scores for all tests (and the reading cut scores) were not tied to any particular value on the final reporting scales. The cut scores were placed on the final reporting scales using the same linear transformations as described above. The final, approved cut scores are shown in Tables 3.1 and 3.2.

Table 3.1: Approved Cut Scores

Content Area	Grade or Course	Approved Cut Scores		
		Developing Learner	***Proficient Learner***	***Distinguished Learner***
English Language Arts	3	475	525	581
	4	475	525	574
	5	475	525	587
	6	475	525	599
	7	475	525	592
	8	475	525	581
	9LCO	475	525	587
	AMLC	475	525	590
Mathematics	3	475	525	580
	4	475	525	585
	5	475	525	580
	6	475	525	580
	7	475	525	580
	8	475	525	579
	CALG	475	525	594
	AGEO	475	525	596
	ALGI	475	525	594
	GEOM	475	525	596
Science	5	475	525	595
	8	475	525	593
	BIOL	475	525	609
	PHSC	475	525	604
Social Studies	5	475	525	555
	8	475	525	572
	HIST	475	525	590
	ECON	475	525	610

Table 3.2: Approved Reading Cut Scores

Content Area	Grade or Course	Approved Cut Scores
		On-Grade Reading
Reading	3	470
	4	494
	5	483
	6	492
	7	472
	8	482
	9LCO	478
	AMLC	482

Note: Reading cut scores are used only to provide a grade-level status indication based on a link between the Georgia Milestones tests of reading and the MetaMetrics Lexile Framework for Reading. See Section 3.4 for more information.

CHAPTER 4:
TEST ADMINISTRATION

Chapter 4 of this technical report describes the processes and activities implemented and the information disseminated to help ensure standardized test administration procedures and, thus, uniform test administration conditions for all students.

Chapter 4 examines how test administration procedures implemented for the Georgia Milestones Assessment System (Georgia Milestones) strengthen and support the intended score interpretations and reduce construct-irrelevant variance that could threaten the validity of score interpretations. Chapter 4 demonstrates adherence to AERA, APA, and NCME (2014) Standards 4.15, 4.16, 6.1, 6.2, 6.3, 6.4, 6.6, 6.7, and 6.10 in the Georgia Milestones program.

4.1 Training of Districts

According to the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014), the "usefulness and interpretability of test scores require that a test be administered and scored according to the test developer's instructions" (p. 111). To ensure that the Georgia Milestones tests were administered and scored in accordance with the GaDOE mandates, the GaDOE took a primary role in communicating and training district personnel. While the development of Georgia Milestones was a collaborative effort between the GaDOE and DRC, the GaDOE wanted to be sure that the districts understood the purpose of the Georgia Milestones program and how its administration must be standardized to fit testing industry standards and to meet the mandates of state and federal legislation and State Board of Education policies. To accomplish these goals, the GaDOE conducted several training sessions specifically geared toward system test coordinators.

System test coordinators are the district liaisons for the GaDOE Assessment and Accountability Division. They are responsible for disseminating information and test materials to each school within their districts and for enforcing the legal and ethical rules for administering tests. The main contacts for district staff are the Georgia Milestones Administration team within the GaDOE's Assessment Administration Division and staff from the Assessment Research and Development Division. These departmental staff members communicate frequently with districts about questions particular to the Georgia Milestones program, answer general assessment questions, and provide assistance with Georgia Milestones data and interpretation of test results.

The GaDOE workshops for system test coordinators regarding the Georgia Milestones program included the following:

- System Test Coordinator Conference
- Pre-administration Workshops
- Lunch and Learn Webinars

The remainder of this section will describe each of these activities, which occurred between August 7, 2018, and March 8, 2019.

4.1.1 System Test Coordinator Conference

The purpose of the System Test Coordinator Conference was to provide information necessary for the successful administration of the 2018–2019 Georgia assessment programs. All system test coordinators were expected to attend in-depth sessions that were delivered as webinars. Multiple dates were provided to accommodate districts. These half-day meetings were held at the beginning of the school year. For the 2018–2019 school year, the meetings were held on the dates shown in Table 4.1. Figure C.1.1 in Appendix C includes a copy of the presentation used for the conference.

Table 4.1: Schedule for System Test Coordinator Conference

Date	Session/Time
8/7/18	Fall Assessment Conference: Part 1 9:00 a.m.–noon EDT
8/9/18	Fall Assessment Conference: Part 1 (live repeat of August 7) 1:00 p.m.–4:00 p.m. EDT
9/4/18	Fall Assessment Conference: Part 2 9:00 a.m.–noon EDT
9/6/18	Fall Assessment Conference: Part 2 (live repeat of September 4) 1:00 p.m.–4:00 p.m. EDT
9/13/18	Fall Assessment Conference: Part 3 1:00 p.m.–4:00 p.m. EDT

The conference showcased each testing program. The Georgia Milestones component of the conference covered the following details:

- General announcements with links to available resources, state board rules, transmission of secure information, and data reporting
- Test Security and Assessment administration
    - Professional ethics for fair and ethical testing for all students and stakeholders
    - Code of ethics for Georgia educators
    - Roles and responsibilities—superintendent, system test coordinator, system tech coordinator, school test coordinator, examiner, and proctor
    - Training plan that system test coordinators must develop within the district
    - Online testing considerations
    - Materials management and security
    - Scheduling considerations
    - Security for online testing environments
    - Cell phones and electronic devices
    - Steps for avoiding testing irregularities and reporting an irregularity
    - Assessment online forms

- Accommodations and special populations
- Post-assessment communication, dissemination of test scores, and transferring student test scores
- Overview of 2018–2019 state assessments
- Central Office Services (COS)
- Overview of Georgia Milestones and participants for EOG and EOC assessments
- Georgia Milestones, administration windows, and winter EOC preliminary student reports and reporting timeline
- Georgia Milestones EOG scheduling flexibility
- Assessment and Accountability Division contact information

Each year the Assessment and Accountability Division produces a handbook that includes a wealth of information related to general topics such as test security, test administration scheduling, and assessing special populations (see Section 2.3.6). Figure 2.12 shows a copy of the table of contents for the *Student Assessment Handbook* (Georgia Department of Education, 2018).

The System Test Coordinator Conference provides all system test coordinators with a broad overview of each program. Those new to their role are provided with an introduction to each program, and veteran coordinators are provided with program updates that will be in place for the new school year. The fall conference is followed by more detailed workshops specific to individual programs throughout the school year.

4.1.2 Georgia Milestones Pre-Administration Workshops

In preparation for the winter and spring Georgia Milestones assessments, the GaDOE hosted online workshops to present information related to the administration of the Georgia Milestones tests. The target audience was primarily system test coordinators, and the mode of delivery was through a webinar. Because this web-delivery model does not restrict the number of participants, many districts opt to have their school test coordinators participate as well.

The GaDOE arranged several dates and times to accommodate the schedules of district personnel. The pre-administration workshops for Georgia Milestones were offered on the dates shown in Table 4.2. Recordings were made available for both system and school coordinators who were unable to attend the live sessions and for the redelivery of information at the district or school level.

Table 4.2: Schedule for Georgia Milestones Pre-administration Workshops

Date	Description	Program	Time
10/16/18	EOC winter 2018 and Spring 2019 mid-month Pre-administration Workshop	EOC	10:00 a.m.–noon
10/18/18	EOC winter 2018 and Spring 2019 mid-month Pre-administration Workshop	EOC	1:00 p.m.–3:00 p.m.
2/26/19	EOG (spring main and summer retest 2019)/EOC (spring main, summer main and fall mid-month 2019) Pre-administration Workshop	EOG/EOC	10:00 a.m.–noon
2/28/19	EOG (spring main and summer retest 2019)/EOC (spring main, summer main and fall mid-month 2019) Pre-administration Workshop	EOG/EOC	1:00 p.m.–3:00 p.m.

Figures C.1.5 and C.1.6 in Appendix C include copies of the presentations used for the workshops. The presentations for the workshops go into great detail about procedures for online testing, ensuring test security, and administering the test in accordance with the state's established rules. The following topics were major foci of the training:

- Program overview and policies—administration dates, EOG/EOC participants, score utilization
- General Test Security training—security hierarchy, code of ethics for Georgia educators, roles and responsibilities, Student Assessment Handbook references, test security, testing irregularities
- Readiness and preparation—key dates, Georgia Milestones administration windows
- Understanding Georgia Milestones as an online assessment—advantages of online testing, online readiness, Central Office Service (COS), eDIRECT accounts, test tickets and test rosters, universal tools, overview of new item types
- Testing accommodations—accommodations in eDIRECT, transcribing paper/pencil administrations, accommodations in the online environment, online audio, selecting background color options, online accommodation tools for large print, read-aloud in eDIRECT
- Test practice for online administrations—secure practice test, practice test with response transmission, and experience online testing Georgia
- Accessibility and Accommodations—accommodations information in the Student Assessment Handbook, standard and conditional accommodations, special accommodations requests
- Training and resources—train-the-trainer model
    - System Test Coordinator training—links to DOE training series and presentations, checklist for system test coordinators, key dates, scheduling reminders, EOC preliminary reports for spring graduates, EOC resources, EOG resources

  - School Test Coordinator training—checklist for school test coordinator, QuickStart videos and guides, school test coordinator training PowerPoint, school test coordinator online training course
  - Test Examiner training—test administration manuals (TAMs), online versus paper/pencil, scratch paper, calculator policy, testing environment setup
  - Preparing students—links to available resources (updated assessment guides, updated study guides, item and scoring samplers, quick start guides and videos), test practice for online administrations, test practice resources (online tools training, practice videos for technology-enhanced items, secure practice test directions, score descriptions for new technology-enhanced items), INSIGHT content guide
- Monitoring online test sessions—student status, student status dashboard, monitoring accommodations, managing test sessions, managing test administration
- DRC customer service team and the GaDOE contact information

4.1.3 Georgia Milestones Lunch and Learn Webinars

To assist districts with the implementation of Georgia Milestones, the GaDOE offered a series of optional Lunch and Learn Webinars for system test coordinators. The purpose of these webinars was to facilitate ongoing communication with system test coordinators and equip them with the information they needed to support their local schools. The webinars were intended to be informal, provide pertinent assessment updates, and provide a forum for questions and feedback from system test coordinators. Each webinar was scheduled for a one-hour session on a Friday afternoon from noon–1:00 p.m. For the 2018–2019 school year, a total of five Lunch and Learn Webinars were offered from October 12, 2018, through March 8, 2019.

The following topics are a sample of those discussed during these webinars:

- Policy updates
- 2019–2020 Georgia Milestones assessment window
- Georgia Milestones English language arts (ELA) test design
- Test development and educator involvement activities
- Best practices
  - Read the test administration manual prior to the day of testing
  - School testing environment and setup
  - Testing environment—materials
  - Test tickets and student roster
  - Utilize test practice resources
  - Familiarize students with universal tools
  - Familiarize oral reading students with text-to-speech (TTS) audio and other online accommodations

- New technology-enhanced item types and a demo
- INSIGHT content guide
- Online Readiness updates and tools

4.2 Ancillary Materials

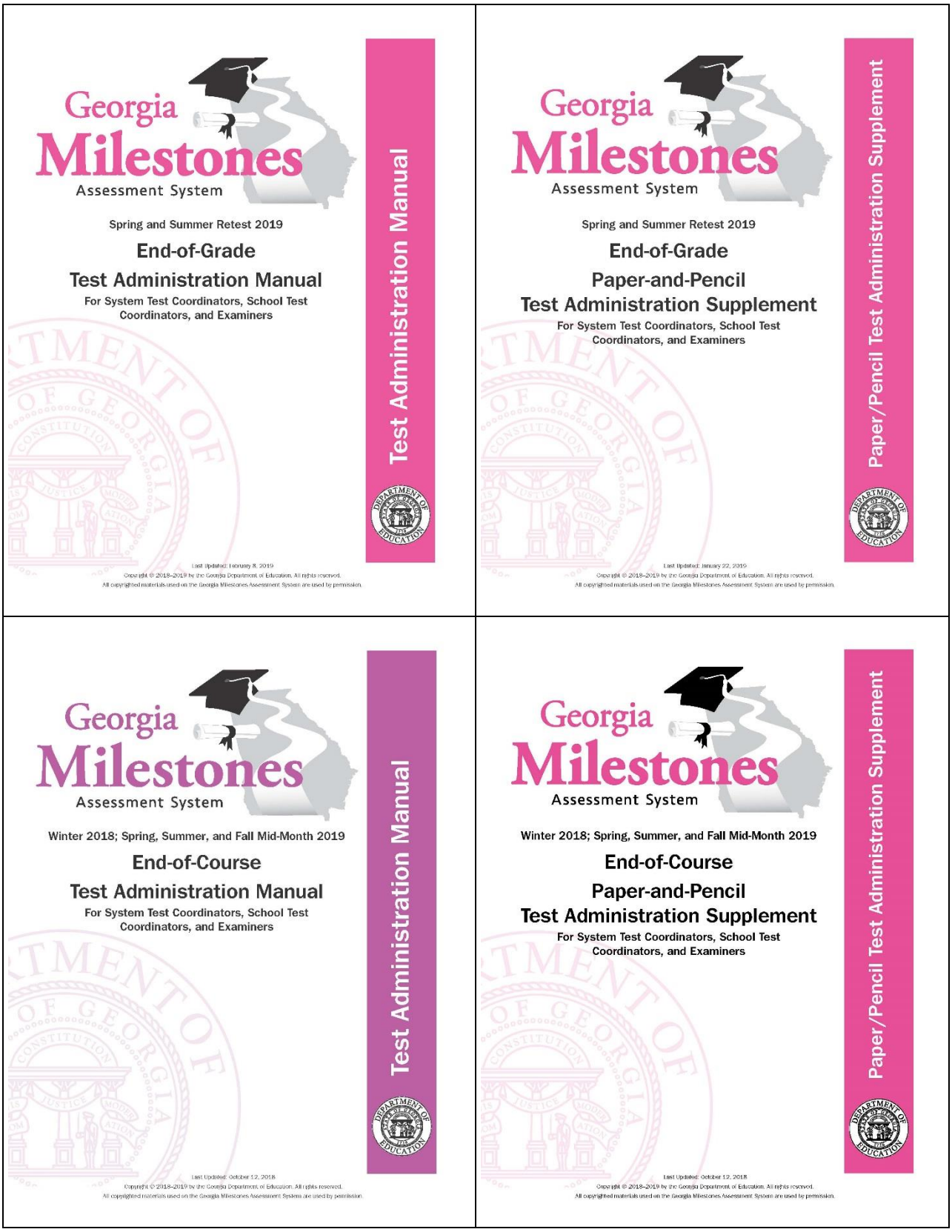
Test administration ancillary materials for the Georgia Milestones program contribute to the body of evidence of the validity of score interpretation. This section details how the test materials specifically address the AERA, APA, and NCME (2014) standards related to test administration procedures.

4.2.1 Test Administration Manuals

For the 2018–2019 test administrations, a series of TAMs (Georgia Department of Education, 2018–2019) covering grades 3 through 8 and high school, as shown in Figure 4.1, was produced by DRC. The GaDOE assessment staff had opportunities to review, provide feedback on, and give final approval of these manuals. The manuals provided instructions for the following pretest and posttest procedures:

- Roles and responsibilities
- Test security
- Students to be tested
- Scheduling the tests
- Testing accommodations
- Online test administration guidelines
- User experience
- Network and Workstation security
- The eDIRECT Online Data portal
- Encoding student information
- Directions for System Test Coordinators
- Directions for School Test Coordinators
- Directions for Examiners
- Directions for returning test materials to school test coordinators after each day of testing, with a checklist for Examiners

Figure 4.1: Cover Pages of the *Test Administration Manuals*



This section presents the AERA, APA, and NCME (2014) standards relevant to test administration and discusses how information in the Georgia Milestones *TAM*s (2018–2019) addresses these standards.

Standard 4.15 The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. (90)

The *TAM*s (2018–2019) present information for pretesting, testing, and posttesting activities with sufficient detail and clarity to support reliable test administrations by qualified test administrators. To ensure uniform administration conditions throughout the state, instructions identify the materials that the test examiner and students need.

The paper/pencil *TAM* (2018–2019) supplements the EOG TAM for student populations who will utilize paper test materials during the Georgia Milestones test administration. This applies to those students whose disability does not allow them access to the online platform with appropriately assigned accommodations. Student responses marked in a student test booklet or on a student answer document are to be transcribed to an online test form.

The online *TAM* (2018–2019) explains how to manage organization and user account information in the online testing system, add and edit information in student management, manage test sessions, assign accommodations for qualified students, and review and print test tickets for students testing online. The manual presents information on how to prepare the classroom and workstations to properly administer the online version of the tests, monitor student test progress during testing, and ensure all students are in complete status.

Standard 4.16 The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. (90)

To ensure clarity of instructions to students, the *TAM*s (2018–2019) include scripts that the test examiner is instructed to read verbatim to students. Test examiners may use professional judgment to respond to student questions, but they may not reword test items, suggest answers, or evaluate student work during the test session. Samples of scripts from EOG and EOC *TAM*s (2018–2019) are presented in Figures 4.2 and 4.3.

Sample test items are provided for each test to familiarize students with how to fill in answers.

Figure 4.2: Sample Script of the EOG *Test Administration Manual*

Part 4: Directions for Examiners

DIRECTIONS FOR ADMINISTERING THE ENGLISH LANGUAGE ARTS EOG SPRING ASSESSMENT

1. *Before you begin English Language Arts Section 1, verify student identity, record attendance, and check that students have the following test materials:*

   - *One* Test Ticket *for each student for each section of the test*
   - *Two No. 2 pencils with erasers*
   - *Two sheets of scratch paper for Section 1, Reading and Evidence-Based Writing. One sheet of scratch paper each for Section 2 and Section 3. For two-day and three-day administrations, students may receive a new sheet of scratch paper. It is permissible for the* Test Ticket *to be used as scratch paper. Consult with your School Test Coordinator for local guidance. The* Test Ticket *is a secure document and must be returned at the end of a testing session.*

   After all students have materials,

   SAY **The testing period has begun. Please do not talk until after you have been dismissed. Cell phones, smartphones, smartwatches, tablets, or other electronic devices are NOT allowed in the testing room. The use, or intended use, of an electronic device to look up correct answers, take pictures of test materials, or share information with others will result in an invalidation of your test score and possible disciplinary action. Invalidation means that you would not receive a test score for this content area test. Are there any devices in the room that I need to collect?**

   If so, collect and secure the device(s) per your school's/system's established procedures.

2. SAY **Look at your *Test Ticket*. Please verify your last name and first name are correct. If your name is correct, please sign your *Test Ticket* now. If you have someone else's *Test Ticket*, please raise your hand now.**

   If a student's hand is raised, provide student(s) with the correct Test Ticket.

   SAY **Please write your name in the upper right-hand corner of the scratch paper. If you need additional scratch paper during the test, raise your hand and I will bring you a sheet. You must write your name in the upper right-hand corner of the scratch paper. I will collect all scratch paper at the end of testing.**

   The following tools are available on your test:

   Refer to the table below and name the tools available.

Online Tool	English Language Arts
Cross-off	Yes
Highlighter	Yes
Line Guide	Yes
Sticky Note	Yes
Magnifier	Yes
Calculator	No
References	No
Writer's Checklist	Yes
Undo	Yes

Examiner – Spring – ELA

End-of-Grade Test Administration Manual Page 35

Copyright © 2018–2019 by the Georgia Department of Education. All rights reserved.

Figure 4.3: Sample Script of the EOC *Test Administration Manual*

DIRECTIONS FOR ADMINISTERING THE MATHEMATICS EOC ASSESSMENTS

These directions should be used to administer the Coordinate Algebra, Analytic Geometry, Algebra I, and Geometry EOC assessments.

1. *Before you begin, verify student identity, record attendance, and check that students have the following test materials:*

   - *One* Test Ticket *for each student for each section of the test*
   - *Two No. 2 pencils with erasers*
   - *Two sheets of scratch/graph paper. It is permissible for the* Test Ticket *to be used as scratch paper. Consult with your School Test Coordinator for local guidance. The* Test Ticket *is a secure document and must be returned at the end of a testing session.*
   - *Graphing calculator or scientific calculator (**IMPORTANT:** Graphing calculators or scientific calculators are **not** allowed for Section 1, Part A, for Analytic Geometry, Coordinate Algebra, Algebra I, and Geometry. So withhold issuing any student a calculator until Section 1, Part A, is completed.)*

   You must ensure that all calculators have been cleared of any stored text before beginning the assessment. A failure to confirm that calculator memory has been cleared will constitute a testing irregularity.

   After all students have materials,

   SAY **Before we begin, I must remind you that you may not have in your possession during testing a cell phone or any other electronic device that would allow you to access, retain, or transmit information. The use, or intended use, of such a device to access information and/or to photograph, post, retain, share, or transmit information/images from any portion of a secure test will result in invalidation of your test score. Are there any devices in the room that I need to collect?**

   If so, collect and secure the device(s) per your school's/system's established procedures.

   Confirm that any stored memory on students' hand-held calculators has been cleared.

   SAY **You may not use a cell phone as a calculator. To do so will result in invalidation of your test score.**

2. **Look at your *Test Ticket*. Please verify your last name and first name are correct. If your name is correct, please sign your *Test Ticket* now. If you have someone else's *Test Ticket*, please raise your hand now.**

   If a student's hand is raised, provide student(s) with the correct Test Ticket.

   SAY **Please write your name in the upper right-hand corner of the scratch/graph paper. If you need additional scratch/graph paper during the test, raise your hand and I will bring you a sheet. You must write your name in the upper right-hand corner of the scratch/graph paper. I will collect all scratch/graph paper at the end of testing.**

Examiner – Mathematics

Copyright © 2018–2019 by the Georgia Department of Education. All rights reserved.

Standard 6.1 Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user. (114)

To ensure the usefulness and interpretability of test scores and to minimize sources of construct-irrelevant variance, Georgia Milestones tests are administered according to the prescribed test schedule and test administration order. For the EOG assessments, the *TAM*s (2018–2019) include instructions for scheduling the four content area tests over nine consecutive school days or twelve consecutive school days within the state testing window of April 8 through May 17, 2019. For the EOG retest assessment, the *TAM*s (2018–2019) include instructions for scheduling ELA at grades 3, 5, and 8 and mathematics at grades 5 and 8 within two five-day windows between May 13 and July 19, 2019. For the EOC assessments, the *TAM*s (2018–2019) include information about the winter, spring, and summer main administration testing windows as well as the mid-month administrations. The manuals also include instructions for administering the tests in the order prescribed and following the schedule for timing each test session. The test administration schedule is presented in Section 1 of the manual.

Standard 6.3 Changes or disruptions to standardized test administration procedures or scoring should be documented and reported to the test user. (115)

If testing is interrupted at any time during a test session, the *TAM* (2018–2019) instructs the test examiner to move the student to another available computer that has been properly setup for testing. If test examiners have questions about an individual situation, they are instructed to contact the school or system test coordinator or the technology coordinator.

The EOC and EOG *TAM*s (2018) provide additional information about situations that would constitute a breach of test security and a reminder that any irregularities need to be reported to the system test coordinator. School or district personnel may cause testing irregularities by doing the following:

- Coaching examinees during testing or altering or interfering with examinees' responses in any way
- Giving examinees access to test questions or prompts prior to testing
- Copying, reproducing, or using all or any portion of secure test booklets/online testing forms in any manner inconsistent with test security regulations
- Making answers available to examinees
- Reading or reviewing test questions before, during (unless specified in the Individualized Educational Program [IEP], Individual Accommodation Plan [IAP], or English Learner/Testing Participation Committee [EL/TPC]), or after testing
- Questioning students about test content after the test administration
- Failing to follow security regulations for distribution and return of secure test materials as directed or failing to account for all secure test materials before, during, and after testing (lost test booklets constitute a breach of test security and result in a referral to the Georgia Professional Standards Commission)
- Using or handling secure test booklets, answer documents, or online testing logins/passwords/test forms for any purpose other than examination

- Failing to follow administration directions for the test
- Failing to properly secure and safeguard logins/passwords necessary for online test administration
- Erasing or marking answers or altering responses on an answer document or within an online test form
- Participating in, directing, aiding, counseling, assisting, encouraging, or failing to report any of these prohibited acts

If testing irregularities occur, the system test coordinator is responsible for contacting the GaDOE. If directed to do so by a GaDOE representative, the system test coordinator must then prepare a report of any testing irregularities and submit the report to the GaDOE. The form for reporting irregularities is available on the GaDOE secure portal at https://portal.doe.k12.ga.

Standard 6.4 The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance. (116)

The EOG *TAM* specifically states, "Workspace for each student should be large enough to accommodate test materials. Workspace should be cleared of all other materials. Posters, charts, and other classroom materials related to the content being tested should be removed or covered during testing" (2018–2019, p. 29).

The EOC *TAM* states, "Be sure that all students have a comfortable and adequate workspace" (2018–2019, p. 35). In addition, the manual also states, "Workspace for each student should be large enough to accommodate test materials. Workspace should be cleared of all other materials. Posters, charts, and other classroom materials related to the content being tested should be removed or covered during testing" (2018–2019, p. 30).

Standard 6.6 Reasonable efforts should be made to ensure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent or deceptive means. (116)

The online *TAM* (2018–19) gives instructions to the examiners to check the classroom or computer lab to make sure there is a visual barrier between workstations or the placement of workstations is such that students cannot easily view other students' answers.

The paper/pencil *TAM* (2018–2019) presents instructions for posttest activities to ensure that test materials are handled properly and to preserve the integrity of student information and test scores. Detailed instructions guide test examiners in completing required information on the student answer document. For students who are administered a large-print version of the Georgia Milestones assessment, test examiners are instructed to transcribe student responses from the large-print test booklet to an online administration exactly as they appear in the large-print test booklet. Pre-administration training clearly specifies that the transcription process should be done with at least one additional certified staff member present as a witness. For students who are administered the braille version of the assessment, the test examiner or proctor is instructed to fill in student responses on the answer document or an online test exactly as the student dictates.

Standard 6.7 Test users have the responsibility of protecting the security of test materials at all times. (117)

Throughout the manuals, test coordinators and test examiners are reminded of test security requirements and procedures to maintain test security. Specific actions that are direct violations of test security are so noted. Detailed information about test security procedures and reminders is presented in Section 1 of the EOG *TAM* (2018–2019).

4.2.2 Return Material Forms and Guidelines

The *TAM*s (2018–2019) instruct test coordinators in procedures for organizing and packing materials and returning them to DRC for scanning and scoring. The purpose of the instructions is to ensure that used and unused test materials are properly accounted for and student answer documents are properly organized for return shipment. Proper organization of materials contributes to accurate score reports and helps in the timely delivery of such reports.

The *TAM*s (2018–2019) provide a checklist to guide the test coordinator in preparing materials for return shipping, including graphics that demonstrate the proper organization of materials.

4.2.3 Network and Workstation Security

The Technology Coordinator and School Test Coordinator are responsible for ensuring that all workstations and the school network meet the configuration requirements detailed to the systems and are working efficiently and securely. The technology and test coordinators should have enough knowledge about the technical details of the Windows, Mac, iOS, and Chrome operating systems, and have the necessary security rights and privileges to perform the necessary tasks.

4.2.4 Score Interpretation Guides

Understanding what the test scores mean and how to interpret score reports is essential to making valid interpretations of test scores. The EOG and the EOC *Interpretive Guide for Score Reports* (2019) are written for Georgia teachers and administrators who receive Georgia Milestones score reports from the 2018–2019 administrations. More details about the guides can be found in Section 2.3.5.

4.3 Test Security Measures

Maintaining the security of all test materials is crucial to preventing the possibility of random or systematic errors (such as unauthorized exposure to test items) that would affect the valid interpretation of test scores. Several test security measures have been implemented for Georgia Milestones. Test security procedures are discussed extensively throughout the *TAM*s (2018–2019). In addition, all secure test materials are marked with the secure test materials notice, as shown in Figure 4.4. For online testing, the technology coordinator and the school test coordinator are responsible for ensuring that all workstations and the school network meet the configuration requirements detailed in the systems requirement guide.

Figure 4.4: Notice for Secure Test Materials

SECURE TEST MATERIALS—DO NOT COPY

All test booklets and supporting materials associated with the Georgia Milestones Assessment System are confidential and secure. No part of any test booklet or electronic online test form may be reproduced or transmitted in any form or by any means, including but not limited to electronic, mechanical, manual, or verbal (e.g., photocopying, photography, scanning, recording, paraphrasing—rewording or creating mirror items for instruction—and/or copying). Georgia Milestones materials must remain secure at all times and (excluding the *School and System Test Coordinator's Manual*, the *Paper-and-Pencil Examiner's Manual*, and the *Online Examiner's Manual*) cannot be viewed by any individual or entity prior to or after testing. To do so is a direct violation of testing policies and procedures established by Georgia law (§20-2-281) and State Board of Education (Rule 160-3-1-.07(4) in addition to copyright laws and the Georgia Code of Ethics for Educators. Georgia Milestones materials may not be provided to any persons except those conducting the test administration and those being tested. All test booklets (used and unused) and all supporting materials must be accounted for and returned at the completion of the test administration and in the manner prescribed in the *School and System Test Coordinator's Manual*.

For online testing, the school test coordinators provide examiners with student rosters that identify test sessions, usernames and passwords, and test tickets for all students scheduled to participate. The school test coordinators are instructed to deliver these secure materials immediately before testing and collect them immediately after testing. The school test coordinators are instructed to retain test tickets in a secure location until score results for each test taker have been received. Upon receipt of test scores, the school test coordinators are instructed to securely destroy the test tickets. Before starting the test, the Examiner should know how to contact the Technology Coordinator or School Test Coordinator for technical support without leaving the room unattended. Students may be told to bring a book to read or some other work to do in the event they finish the test early. This material must be unrelated to the content being tested, and students should not be permitted to use the computer again once they have completed the test.

For paper/pencil testing, all secure materials that are sent to the system for testing are also tracked by the security barcode on the front covers. The GaDOE also requires that the superintendent complete an assurance form. DRC requires the return of all materials (i.e., used, unused, scorable, and nonscorable) after the completion of testing and scans every secure barcode to ensure that all materials have been returned. If any materials are missing and the reason is not properly documented on the security checklist, DRC contacts the systems to either locate the missing materials or to determine why the materials were not returned (e.g., a booklet was destroyed after something was spilled on it). This process continues until DRC accounts for all secure documents.

Georgia routinely implements a series of data forensic analyses as an integral part of test security. This work is maintained by the Governor's Office of Student Achievement (GOSA) and includes extensive analysis of answer changes as well as unexpected score fluctuations from year to year.

4.4 Test Administration

The 2019 Georgia Milestones EOG tests were administered to students within the state testing window of April 8 through May 17, 2019. Systems chose one testing period within the state testing window and a content area to be completed within the same week following the district's protocols. The school district can set the order of content areas for EOG administration based on their schedules and computer lab availability. Beginning in the 2016–2017 school year, additional flexibility was provided so that systems may elect to schedule test administrations using one of the two models demonstrated in Figures 4.5 and 4.6. Using mathematics as an example, the following list details the two possible models:

- One-day administration: As in past years, all students in a grade took and completed Sections 1 and 2 of mathematics in one session that was divided by a break.
- Two-day administration: All students in a grade took and completed Section 1 of mathematics on day 1 and Section 2 of mathematics on day 2.

This flexibility was available for all four content areas. However, a system may, for example, have elected to administer one section of ELA and mathematics per day while continuing to administer both sections of science and social studies in one day.

There were three main administrations of the 2018–2019 Georgia Milestones EOC assessments: winter, spring, and summer. In addition, online mid-month administrations were available in August, September, October, November, January, February, and March. Mid-month administrations were available via electronic testing. Note that systems could request paper/pencil testing materials for mid-month administrations for students with disabilities. The schedule of test administration windows is shown in Table 4.3. Each school system determined the sequence and scheduling option for all EOC assessments. The actual time of day at which the tests were administered may have varied from school to school. The decision to use this flexibility (one-day or two-day administrations) was a district decision to be implemented in all schools in the district.

Table 4.3: End-of-Course Winter 2018–Summer 2019 Test Administration Windows

Administration	Test Administration Testing Window
Fall mid-month	August 6–24, 2018 September 10–21, 2018 October 9–26, 2018 November 5–16, 2018
Winter main	November 26, 2018–January 4, 2019
Spring mid-month	January 14–25, 2019 February 11–22, 2019 March 11–22, 2019
Spring main	April 22–May 31, 2019
Summer main	June 17–July 19, 2019

4.4.1 Time

Each section of each grade/content area or course test is timed to provide sufficient time for students to attempt all items. The *TAM*s (2018–2019) instruct test examiners to allow all students a minimum of 45 to 70 minutes, depending on the grade/content area or course, to complete each test section. If students are still productively engaged in completing the section, testing is to be continued for the full amount of allotted time, which is 70 to 90 minutes, depending on the grade/content area or course. The administration schedules of the Georgia Milestones tests are presented in Figures 4.5 and 4.6.

Figure 4.5: Georgia Milestones EOG Suggested Administration Schedule

One-Day Administration

Content Area	Task	Time Allotted
English Language Arts: Section 1, Reading and Evidence-Based Writing • Grades 3–8 Literature and Composition	Distribute test materials, read test instructions, and Practice Questions	10 minutes
	Testing Time	70–90 minutes
	Collect test materials	15 minutes
English Language Arts: Sections 2 and 3 • Grades 3–8 Literature and Composition	Distribute test materials and read test instructions	10 minutes
	Testing Time Section 2	60–85 minutes
	Break	Local Determination
	Testing Time Section 3	60–85 minutes
	Collect test materials	15 minutes
Mathematics • Grades 3–8 Mathematics	Distribute test materials, read test instructions, and Practice Questions	10 minutes
	Testing Time Section 1	60–85 minutes
	Break	Local Determination
	Testing Time Section 2	60–85 minutes
	Collect test materials	15 minutes
Science • Grades 5 and 8 Science	Distribute test materials, read test instructions, and Practice Questions	10 minutes
	Testing Time Section 1	45–70 minutes
	Break	Local Determination
	Testing Time Section 2	45–70 minutes
	Collect test materials	15 minutes
Social Studies • Grades 5 and 8 Social Studies	Distribute test materials, read test instructions, and Practice Questions	10 minutes
	Testing Time Section 1	45–70 minutes
	Break	Local Determination
	Testing Time Section 2	45–70 minutes
	Collect test materials	15 minutes

Figure 4.5: Georgia Milestones EOG Suggested Administration Schedule (continued)

Multiple-Day Administration

Content Area	Day	Task	Time Allotted
English Language Arts: Section 1, Reading and Evidence-Based Writing • Grades 3–8 Literature and Composition	1	Distribute test materials, read test instructions, and Practice Questions	10 minutes
		Testing Time	70–90 minutes
		Collect test materials	15 minutes
English Language Arts: Sections 2 and 3* • Grades 3–8 Literature and Composition	2	Distribute test materials and read test instructions	10 minutes
		Testing Time Section 2	60–85 minutes
		Collect test materials	15 minutes
	2 or 3	Distribute test materials and read test instructions	10 minutes
		Testing Time Section 3	60–85 minutes
		Collect test materials	15 minutes
Mathematics • Grades 3–8 Mathematics	1	Distribute test materials, read test instructions, and Practice Questions	10 minutes
		Testing Time Section 1	60–85 minutes
		Collect test materials	15 minutes
	2	Distribute test materials and read test instructions	10 minutes
		Testing Time Section 2	60–85 minutes
		Collect test materials	15 minutes
Science • Grades 5 and 8 Science	1	Distribute test materials, read test instructions, and Practice Questions	10 minutes
		Testing Time Section 1	45–70 minutes
		Collect test materials	15 minutes
	2	Distribute test materials and read test instructions	10 minutes
		Testing Time Section 2	45–70 minutes
		Collect test materials	15 minutes
Social Studies • Grades 5 and 8 Social Studies	1	Distribute test materials, read test instructions, and Practice Questions	10 minutes
		Testing Time Section 1	45–70 minutes
		Collect test materials	15 minutes
	2	Distribute test materials and read test instructions	10 minutes
		Testing Time Section 2	45–70 minutes
		Collect test materials	15 minutes

* Some school districts may choose to separate ELA Sections 2 and 3, which would result in the three ELA assessments being administered over three days. Although permissible, school districts must still ensure that all sections are completed within the same week.

Figure 4.6: Georgia Milestones EOC Suggested Administration Schedule

One-Day Administration

Content Area	Task	Time Allotted
English Language Arts: Section 1, Reading and Evidence-Based Writing • Ninth Grade Literature and Composition • American Literature and Composition	Distribute test materials, read test instructions, and Practice Questions	10 minutes
	Testing Time	70–90 minutes
	Collect test materials	15 minutes
English Language Arts: Sections 2 and 3 • Ninth Grade Literature and Composition • American Literature and Composition	Distribute test materials and read test instructions	10 minutes
	Testing Time Section 2	60–85 minutes
	Break	Local Determination
	Testing Time Section 3	60–85 minutes
	Collect test materials	15 minutes
Mathematics • Coordinate Algebra • Analytic Geometry • Algebra I • Geometry	Distribute test materials, read test instructions, and Practice Questions	10 minutes
	Testing Time Section 1	60–85 minutes
	Break	Local Determination
	Testing Time Section 2	60–85 minutes
	Collect test materials	15 minutes
Science • Biology • Physical Science	Distribute test materials, read test instructions, and Practice Questions	10 minutes
	Testing Time Section 1	45–70 minutes
	Break	Local Determination
	Testing Time Section 2	45–70 minutes
	Collect test materials	15 minutes
Social Studies • United States History • Economics/Business/Free Enterprise	Distribute test materials, read test instructions, and Practice Questions	10 minutes
	Testing Time Section 1	45–70 minutes
	Break	Local Determination
	Testing Time Section 2	45–70 minutes
	Collect test materials	15 minutes

Figure 4.6: Georgia Milestones EOC Suggested Administration Schedule (continued)

Multiple-Day Administration

Content Area	Day	Task	Time Allotted
English Language Arts: Section 1, Reading and Evidence-Based Writing • Ninth Grade Literature and Composition • American Literature and Composition	1	Distribute test materials, read test instructions, and Practice Questions	10 minutes
		Testing Time	70–90 minutes
		Collect test materials	15 minutes
English Language Arts: Sections 2 and 3* • Ninth Grade Literature and Composition • American Literature and Composition	2	Distribute test materials and read test instructions	10 minutes
		Testing Time Section 2	60–85 minutes
		Collect test materials	15 minutes
	2 or 3	Distribute test materials and read test instructions	10 minutes
		Testing Time Section 3	60–85 minutes
		Collect test materials	15 minutes
Mathematics • Coordinate Algebra • Analytic Geometry • Algebra I • Geometry	1	Distribute test materials, read test instructions, and Practice Questions	10 minutes
		Testing Time Section 1	60–85 minutes
		Collect test materials	15 minutes
	2	Distribute test materials and read test instructions	10 minutes
		Testing Time Section 2	60–85 minutes
		Collect test materials	15 minutes
Science • Biology • Physical Science	1	Distribute test materials, read test instructions, and Practice Questions	10 minutes
		Testing Time Section 1	45–70 minutes
		Collect test materials	15 minutes
	2	Distribute test materials and read test instructions	10 minutes
		Testing Time Section 2	45–70 minutes
		Collect test materials	15 minutes
Social Studies • United States History • Economics/Business/Free Enterprise	1	Distribute test materials, read test instructions, and Practice Questions	10 minutes
		Testing Time Section 1	45–70 minutes
		Collect test materials	15 minutes
	2	Distribute test materials and read test instructions	10 minutes
		Testing Time Section 2	45–70 minutes
		Collect test materials	15 minutes

* Some school districts may choose to separate ELA Sections 2 and 3, which would result in the three ELA assessments being administered over three days. Although permissible, school districts must still ensure that all sections are completed within the same week.

4.4.2 Accommodations

Testing accommodations are permissible for a student with a current IEP, IAP/Section 504 plan, or EL/TPC plan on file. Regular program students who do not qualify under the Individuals with Disabilities Education Act (IDEA) or Section 504, including those served by Student Support Teams (SST), will not be granted test administration accommodation(s).

A student with a disability or who is an English learner (EL) may be provided a test administration accommodation based on his or her IEP, IAP, or EL/TPC plan. An EL student enrolled for the first time in a school in the United States may receive a one-time deferment from assessments in content areas other than mathematics and science. Braille and large-print forms are provided to any student with vision disabilities.

Accommodations must be specified in the student's IEP, IAP, or EL/TPC plan and must be consistent with accommodations used during daily classroom instruction and testing. The use of any accommodation must be indicated on the student answer document.

Any departure from the list of standard accommodations may alter the nature of the test content and the knowledge and skills being assessed, resulting in a conditional administration. If an accommodation that is not listed in the table of allowable accommodations (or is listed as a conditional accommodation) is given based on a student's IEP, IAP, or EL/TPC plan, it will result in a conditional administration and should be coded as such under "conditional administration." The table of accommodations presented in the *Accommodations Manual* (2017) is shown in Figure 4.7.

In addition to the clear guidance on accommodations in these Georgia Milestones test-specific manuals, the GaDOE attends to this topic in general during the System Test Coordinator Conference (Section 4.1.1). The GaDOE also details the available and acceptable use of accommodations in the *Student Assessment Handbook* (2018) (Section 2.3.6) and in the *Accommodations Manual* (2017) (Section 2.3.7).

Figure 4.7: Georgia Milestones Accommodations

ALLOWABLE ACCOMMODATIONS FOR STUDENTS WITH DISABILITIES

Setting Accommodation	GAA 2.0	ACCESS FOR ELLS 2.0[16]	GKIDS	Georgia Milestones	NAEP
1. Special education classroom	A	A	S	S	S
2. Special or adapted lighting	A	A	S	S	S
3. Small group	A	A	S	S	S
4. Preferential seating	A	A	S	S	S
5. Sound field adaptations	S	A	S	S	S
6. Adaptive furniture (e.g. slant board)	A	A	S	S	S
7. Individual or study carrel	A	A	S	S	S
8. Individual administration	A	A	S	S	S
9. Test administered by certified educator familiar to student	A	A	S	S	S
Presentation Accommodations	**GAA 2.0**	**ACCESS FOR ELLS 2.0[16]**	**GKIDS**	**Georgia Milestones**	**NAEP**
10. Large Font/Large Print	S	S	S	S	S
11. Video Sign Language/Sign the directions	S	S	S	S	S
12. Video Sign Language/Sign test questions	S		S	S	S
13. Sign English Language Arts (ELA) passages	S		S	C[1]	
14. Oral reading of test questions in English	A	S[3]		S[11]	S
15. Text to Speech/Oral reading of English Language Arts (ELA) passages in English	A	S[3]		C[1]	
16. Explain or paraphrase the directions for clarity (in English only)	S	A	S	S	
17. Braille	S	S	S	S	S
18. Color overlays, templates, or place markers	S	S[2]	S	S	S
19. Use of highlighter by student	S	S[2]			S
20. Magnification/Low vision aids (e.g. CCTV, other magnifying equipment)	S	S[2]	S	S	S
21. Repetition of directions (in English only)	A	S	S	S	
22. Audio amplification devices or noise buffer/listening devices	S	S[2]	S	S	S
23. Use directions that have been marked by teacher		S			
Response Accommodations	**GAA 2.0**	**ACCESS FOR ELLS 2.0[16]**	**GKIDS**	**Georgia Milestones**	**NAEP**
24. Technology applications, such as Brailler or other communications device with grammar and spell checks disabled; Internet disabled for device	S	S	S	S	S
25. Student marks answers in test booklet	A	S[12]			
26. Student points to answers[15]	A	S[12]	S	S	S
27. Verbal response in English only	A	S[4]	S	S	S
28. Scribe	A[6]	S[6,7]	S[6]	S[6]	S[8]
29. Braille writer/Braille Note-Taker	S	S[14]	S	S	S
30. Abacus	A			S[9]	

Figure 4.7: Georgia Milestones Accommodations (continued)

31. Basic function calculator or adapted basic calculator (e.g. Braille or talking calculator)	S			C[1]	S[5]
32. Adapted writing tools (e.g. pencil grips, large diameter pencil)	A	S[2]	S	S	S
33. Adapted/lined paper	A			S	
Scheduling Accommodations	**GAA 2.0**	**ACCESS FOR ELLS 2.0[16]**	**GKIDS**	**Georgia Milestones**	**NAEP**
34. Frequent monitored breaks	A	A	S	S	S
35. Optimal time of day for testing	A	S	S	S	S
36. Extended time	A	S	S	S	S
37. Flexibility in the order of administration for content areas	A	S[13]		S	
38. Extending sessions over multiple days	A	S[10]			

A = Administration procedures allowable for all eligible students.
S = Standard accommodation required for eligible students.
C = Conditional accommodation for required for eligible students.

Footnotes

1. Restricted to eligible students only; see guidance for eligibility. For oral reading, screen reader is the preferred method of administration (exception Signing of ELA Passages). Where a human reader delivers the accommodation, examiners must adhere to directions provided in the Read-Aloud Guidelines.
2. If prescribed as an accommodation, allowable for all ACCESS FOR ELLS 2.0 domains. Allowable as an accessibility tool for all students.
3. Allowed for ACCESS FOR ELLS 2.0 Writing, Listening, and Speaking domains only.
4. Allowed for ACCESS for ELLs 2.0 Listening, Reading, Speaking domains; not allowed for Writing domain.
5. Only NAEP calculator active blocks will be given to students who need this accommodation.
6. Use of a scribe is allowable if guidelines are followed exactly *from the Student Assessment Handbook*.
7. Use of a scribe is not allowed for ACCESS FOR ELLS 2.0 Speaking domain.
8. Accommodation not allowed on NAEP Writing assessments.
9. Allowed for students with visual impairments only.
10. Paper-Pencil Mode: Allowed for ACCESS FOR ELLS 2.0 Listening and Reading domains; not allowed for Writing and Speaking domains. Not allowed for any domain via online testing mode.
11. Screen reader is the preferred method of administration. Where a human reader delivers the accommodation, examiners must adhere to directions provided in the posted Read-Aloud Guidelines.
12. Paper-Pencil Mode: Allowed for ACCESS FOR ELLS 2.0 Listening, Reading, and Writing domains; not allowed for Speaking domain. Online Mode is not recommended for students requiring this accommodation.
13. ACCESS for ELLs 2.0 Reading domain must be administered prior to Speaking and Writing domains.
14. Allowed for ACCESS FOR ELLS 2.0 Writing, Listening, and Reading domains; not allowed for Speaking domains.
15. Student selects answer by eye gaze allowed for GAA 2.0 participants.
16. Allowable accommodations for Alternate ACCESS can be found in the *WIDA Accessibility and Accommodations Supplement*.

In summary, the overall purpose of each of the various workshops and ancillary materials is to keep districts informed about policies and procedures related to testing in general and to the Georgia Milestones program in particular. The information imparted is clearly related to standardizing the administration of Georgia Milestones, maintaining the security of the assessments, allowing access to the assessments for special populations by clearly delineating appropriate accommodations, and providing guidance on appropriate interpretations of the test results. These communication and training efforts by the GaDOE and the ancillary information developed by DRC address multiple best practices of the testing industry, particularly related to the following *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014):

- Standard 6.1—Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user. (114)
- Standard 6.2—When formal procedures have been established for requesting and receiving accommodations, test takers should be informed of these procedures in advance of testing. (115)
- Standard 6.3—Changes or disruptions to standardized test administration procedures or scoring should be documented and reported to the test user. (115)
- Standard 6.4—The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance. (116)
- Standard 6.6—Reasonable efforts should be made to ensure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent or deceptive means. (116)
- Standard 6.7—Test users have the responsibility of protecting the security of test materials at all times. (117)
- Standard 6.10—When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used. (119)

CHAPTER 5:
PERFORMANCE SCORING

Chapter 5 of this technical report describes the processes and activities implemented to ensure consistent and accurate standardized test handscoring procedures for all students.

Chapter 5 examines how the performance assessment procedures implemented for the handscored portions of the Georgia Milestones assessments work together to maximize scoring accuracy and consistency both within and across administrations in a pre-equated testing model. Chapter 5 demonstrates the Georgia Milestones program's adherence to AERA, APA, and NCME (2014) Standards 6.8, 6.9, 7.10, 12.14, and 12.15.

Scoring of the Georgia Milestones Assessment System (Georgia Milestones) included the scoring of multiple-choice items against the answer key and the raw scores obtained from constructed-response items that required handscoring. A student's raw score is the total number of points achieved by the student on the multiple-choice and the multipoint items for a given assessment. From the raw scores, the scale scores were calculated using procedures described in Section 6.4.

5.1 Scoring of Multiple-Choice Items

Student responses to multiple-choice questions were scored against the final, approved answer keys. Sections of the test were evaluated as a whole, and an attempt status was determined for each subject area. DRC's item banking and test publishing system, IDEAS, includes all relevant data elements that are used to facilitate scoring of each multiple-choice item on the Georgia Milestones assessments.

5.2 Scoring of Constructed-Response Items

The development of Georgia Milestones constructed-response items—items for which the student constructs, rather than selects, his or her response—involved DRC's Performance Assessment Services (PAS), or "handscoring," division. The test development processes or steps that involved PAS included the following:

- pre-rangefinding
- rangefinding
- handscoring training materials development
- field test scoring
- operational scoring

5.3 Pre-rangefinding

Pre-rangefinding was the first opportunity for the Georgia Department of Education (GaDOE) staff to converse with DRC's handscoring staff about how to score individual constructed-response items. As such, pre-rangefinding meetings represented both a chance for the GaDOE to communicate specific, per-item score point characteristics to DRC's handscoring staff and a chance for DRC's handscoring staff to ask questions about how best to score

particular response types. The GaDOE staff also provided preliminary confirmation of scores on the papers that ultimately composed DRC's handscoring training materials. These included scores on "grounding papers," or papers used as exemplars of score points. What follows is a description of the steps in the constructed-response item development process, which involved DRC's PAS division.

After student answer documents were received and processed for the embedded field test portion of the spring 2019 Georgia Milestones End-of-Grade (EOG) and End-of-Course (EOC) assessments, DRC's PAS Scoring Directors (SDs) selected student responses for pre-rangefinding. Table 5.1 shows the breakdown of field tested items by assessment, content area, grade or course, and item type.

Table 5.1: Number of Items by Item Type in the 2019 Georgia Milestones Embedded Field Test

Assessment	Content Area	Grade or Course	Number of 1–4-Point (ELA)/ 0–4-Point (Mathematics) Extended Constructed-Response Items	Number of 0–2-Point Constructed-Response Items
EOG	English Language Arts	3	5	5
		4	3	3
		5	5	5
		6	3	3
		7	3	3
		8	5	5
	Mathematics	3	4	4
		4	4	5
		5	4	3
		6	5	4
		7	5	4
		8	3	3
EOC	English Language Arts	9LCO	5	5
		AMLC	5	5
	Mathematics	CALG	4	4
		AGEO	4	4
		ALG1	10	7
		GEOM	10	7

Note: No writing prompts were field tested in 2019.

For each field tested item, DRC SDs reviewed 300–500 responses and, from that sample, selected the 45–55 responses that composed each pre-rangefinding set. For each item, the set of responses was assembled to represent the full range of the score points delineated by the applicable Georgia Milestones scoring rubric. As part of this process, SDs selected grounding papers. These functioned as a starting point for discussions surrounding the scoring of each item. One to three grounding papers per score point were selected for mathematics, while two grounding papers per score point were selected for English language arts (ELA). Grounding papers provided a means of focusing the discussion on the general attributes of each score point and achieving a basic level of agreement among pre-rangefinding participants.

Responses that contained elements of adjacent score points were also selected so that a consensus score based on the pre-rangefinding group's interpretation of the scoring rubric could be determined. Additionally, SDs made sure to include responses about which they had questions in the pre-rangefinding sets to get consensus scores and response-specific scoring feedback from the group.

When all papers had been selected, sets were assembled for pre-rangefinding calls with the GaDOE. PDFs of each set were created and posted to the GaDOE secure FTP site. As SDs completed and posted pre-rangefinding sets, pre-rangefinding calls were scheduled with the GaDOE. Thirteen pre-rangefinding calls were held—eight for ELA and five for mathematics. All pre-rangefinding calls for the embedded field test were held between April 30 and May 23, 2019. For all field tested items, DRC was able to provide GaDOE staff at least five business days to review each posted set in advance of that item's scheduled pre-rangefinding call. Pre-rangefinding calls were conducted via Skype to allow participants to view and discuss individual student responses simultaneously as a group.

SDs facilitated the pre-rangefinding calls with participation from DRC's Georgia handscoring project management team, DRC's Georgia test development team, and the GaDOE, including assessment and curriculum experts, who provided scoring insight and helped with the interpretation and application of rubrics. Pre-rangefinding discussions often revolved around how best to employ rubrics, previous scoring decisions, and rationales to meet the department's vision for consistency and rigor.

After the pre-rangefinding meetings were completed, DRC scoring staff incorporated consensus scores and scoring rationales that resulted from the calls into their sets. When necessary—for example, due to a scarcity of examples of a particular score point—new responses were found, approved by the GaDOE, and added to the sets at that point. As sets were finalized, they were considered to be rangefinding sets as opposed to pre-rangefinding sets. This change in set characterization concluded the pre-rangefinding process.

5.4 Rangefinding

The next step in the constructed-response item development process, which involved DRC's PAS division, was rangefinding. The rangefinding meetings provided a chance for DRC handscoring staff to meet, in person, with Georgia educators to further fine-tune the scoring of individual constructed-response items. The rangefinding committees either confirmed the consensus scores generated during pre-rangefinding or provided new consensus scores along with accompanying scoring rationales. Both the consensus scores and the associated scoring rationales generated during rangefinding were incorporated into handscoring training materials later used in field test scoring.

The rangefinding meetings for the embedded field test were held at the Westin-Buckhead hotel in Atlanta from June 3 to 5, 2019, for EOG ELA grades 4, 6, and 7; June 3 to 6, 2019, for EOG mathematics; June 3 to 7, 2019, for EOG ELA grades 3, 5, and 8 as well as all EOC ELA and mathematics. All EOG mathematics utilized grade bands, in which two adjacent grades were assigned to a single conference room. EOC mathematics utilized the following course bands: Coordinate Algebra/Algebra I and Analytic Geometry/Geometry. All ELA conference rooms rangefound 1–4-point narrative prompts and 0–2-point constructed-response items. All mathematics rooms rangefound 0–4-point extended constructed-response and 0–2-point

constructed-response items. At rangefinding, each item had an associated set of student responses that the rangefinding committee scored. Table 5.2 shows the breakdown by assessment, content area, and grade or course of the number of responses that were rangefound in each room and the time frame during which the associated rangefinding meeting was held.

Table 5.2: Approximate Number of Responses Rangefound and Days Required for the 2019 Georgia Milestone Embedded Field Test

Assessment	Content Area	Grade or Course	Approximate Number of Responses Rangefound (by room)	Number of Days: 3 = M, T, W 4 = M, T, W, Th 5 = M, T, W, Th, F
EOG	English Language Arts	3	475	5
		4	285	3
		5	475	5
		6	285	3
		7	285	3
		8	475	5
	Mathematics	3	855	4
		4		
		5	790	4
		6		
		7	745	4
		8		
EOC	English Language Arts	9LCO	475	5
		AMLC	475	5
	Mathematics	CALG	1,235	5
		ALG1		
		AGEO	1,235	5
		GEOM		

Note: The numbers of responses are based on approximations of 45 responses per rangefinding set for 0–2-point constructed-response items and 55 responses per set for 0–4-point extended constructed-response items.

5.4.1 Rangefinding Committees

Rangefinding committees were composed of Georgia educators and facilitated by DRC SDs. Six-member committees were used in all conference rooms for the majority of meeting days. For the EOG mathematics 5 and 6 grade band, and for the EOC mathematics course bands, eight-person rangefinding committees were used. Rangefinding committee members were selected with the following six factors taken into consideration:

- Appropriateness of grade/content area or course of current or previous work focus
- Previous rangefinding and other meeting (e.g., data review, new item review) experience
- Region of Georgia
- Ethnicity
- Gender

- Experience with special student populations (e.g., gifted, special education, English learner [EL])

Above all, each committee member needed to have past or present work experience that related in some way to the grade/content area or course for which they were selected. An attempt was made to include at least one committee member with previous rangefinding experience on each committee. A mix of ethnicities, genders, and regions within Georgia was sought within each committee.

Tables 5.3 and 5.4 provide a breakdown of how the factors outlined above were represented within each committee.

Table 5.3: Demographics of the Rangefinding Committee for the 2019 Georgia Milestones EOG Embedded Field Test

Rangefinding Committee Demographics		ELA Grades						Mathematics Grades		
		3	4	5	6	7	8	3/4	5/6	7/8
Previous Item Development Meeting Experience (by committee member)	Rangefinding	2	1	2	2	2	2	2	3	3
	Other Meetings, but No Rangefinding	-	1	-	1	-	-	-	2	-
	No Previous Experience	4	4	4	3	4	4	4	3	3
Experience with ELs, Students with Disabilities, and/or Gifted Students	ELs	2	1	1	3	1	1	1	3	3
	Special Education	3	1	2	1	2	2	1	1	2
	Gifted	3	5	1	3	4	3	3	4	3
Number of Participants by Region	Central	2	2	3	1	2	-	2	2	1
	Metro	2	1	1	1	1	2	1	2	2
	North	1	1	1	2	2	2	2	2	1
	South	1	2	1	2	1	2	1	2	2
	Higher Ed	-	-	-	-	-	-	-	-	-
Number of Participants by Educators' Race/Ethnicity	American Indian or Alaska Native	-	-	-	-	-	-	-	-	1
	Asian/Pacific Islander	-	-	-	-	-	-	1	1	-
	Black or African American	1	1	1	1	1	1	2	2	3
	White	4	5	4	4	4	5	3	4	2
	Hispanic or Latino	-	-	-	-	-	-	-	1	-
	Multi-race	1	-	-	1	-	-	-	-	-
	Declined to State	-	-	1	-	1	-	-	-	-
Gender	Male	-	-	1	-	-	1	1	1	1
	Female	6	6	5	6	6	5	5	7	5

Table 5.4: Demographics of the Rangefinding Committee for the 2019 Georgia Milestones EOC Embedded Field Test

Rangefinding Committee Demographics		Course			
		9LCO	AMLC	CAL/ALGI	GEOM/AGEO
Previous Item Development Meeting Experience (by committee member)	Rangefinding	3	3	3	2
	Other Meetings, but No Rangefinding	-	1	-	1
	No Previous Experience	3	2	3	5
Experience with ELs, Students with Disabilities, and/or Gifted Students	ELs	1	1	1	2
	Special Education	2	1	2	3
	Gifted	2	1	4	2
Number of Participants by Region	Central	1	1	2	2
	Metro	-	-	2	2
	North	2	1	1	1
	South	1	2	1	1
	Higher Ed	2	2	2	2
Number of Participants by Educators' Race/Ethnicity	American Indian or Alaska Native	1	-	-	-
	Asian/Pacific Islander	-	-	-	-
	Black or African American	-	1	3	4
	White	4	5	5	3
	Hispanic or Latino	-	-	-	1
	Multi-race	1	-	-	-
	Declined to State	-	-	-	-
Gender	Male	3	1	1	3
	Female	3	5	7	5

5.4.2 Rangefinding Process

As committee members were being confirmed, DRC scoring staff finalized their rangefinding sets. These were the same sets that were employed during the pre-rangefinding process, but now they included the input and feedback from the pre-rangefinding calls. All sets were finalized for rangefinding before rangefinding began.

Rangefinding started with an opening session that included all grades/content areas and courses. This large-group session began with a review of relevant Georgia Milestones background information presented by the GaDOE and a discussion of the item development

process. The large group then broke into grade/course-specific groups by content area. In each breakout room, each member signed a confidentiality agreement form, which was collected by the DRC facilitator. Next, the facilitator led a discussion of the Georgia Milestones Scoring Philosophy's pertinent points, focusing on the section of that document relevant to the content area and item type being rangefound. Once that discussion was completed, a paper set of student responses (i.e., a rangefinding set) was presented to each committee member. Before discussing any student responses, the facilitator had the committee members read the passage for ELA items or discuss the pertinent concepts addressed in each mathematics item. Finally, a discussion of the item and rubric ensued. Only after all questions about the passages, items, and/or rubrics had been answered did the facilitator move the committee's focus on to the associated student responses.

Each committee began by looking at student work in batches of between six and ten responses. In the ELA rooms, the grounding papers from pre-rangefinding were positioned within the first batch of student responses within the set for each item. These papers were used to ground rangefinding participants' thinking regarding the general attributes of each score point. In the mathematics rooms, the grounding papers were spread throughout each set due to mathematics rangefinding sets being organized in descending order of score points. In both ELA and mathematics rooms, whenever a committee's consensus score diverged from the pre-rangefinding score on a grounding paper, the facilitator let the appropriate GaDOE and/or DRC manager know and the committee's rationale was sought. After there was evidence of room-wide scoring consistency, committee members went on to independently score up to approximately 20 student responses at a time before reconvening as a group for discussion. For every student response, committee members' scores were noted and discussed until a consensus score was reached.

Discussions of student responses were conducted in a manner that emphasized scoring guideline language and the relevant points from the Georgia Milestones scoring philosophy. This ensured that committee members remained focused on the specific Georgia scoring requirements for each score level. PAS staff took notes addressing how and why the committees arrived at score point decisions; this information, referred to as the scoring rationale, was subsequently used by the SDs in rater training and scoring.

The committee in each rangefinding room worked item by item until all items for that grade/content area or course were completed. Consensus scores and committee-generated rationales were noted on tally sheets kept by the SD and by the official notetaker—a designated committee member—in each room. The SD and notetaker in each room merged their tally sheets to ensure completeness and consistency. As items were completed, DRC facilitators gave their tally sheets to GaDOE staff for final confirmation of scores and rationales. After the GaDOE had signed off on each tally sheet, the rangefinding set for that item was considered finalized and ready to be used to create handscoring training materials.

In addition to the processes outlined above, PAS and Test Development (TD) staff discussed any edits to the rubrics or scoring guidelines suggested by the rangefinding committees, with GaDOE. Rubric changes approved by GaDOE were incorporated into the scoring guidelines by TD staff. The edited scoring rubrics were used to prepare materials and train raters.

5.4.3 Post-rangefinding

After rangefinding concluded, DRC's rangefinding facilitators assembled field test training materials. GaDOE-approved scoring guidelines and rangefinding sets (i.e., consensus-scored student responses approved by rangefinding committees and/or the GaDOE with their associated scoring rationales) were used to create handscoring training materials, which were subsequently used to train raters for field test scoring. Specifically, SDs assembled responses that were relevant in terms of the scoring concepts they illustrated into anchor sets, training sets, and qualifying sets. At this point, the notes that facilitators took during rangefinding of the committee members' scoring rationales were sourced to write annotations to accompany anchor and training papers.

Anchor Sets—The full range of each score point was clearly represented and annotated in the anchor set for each item. A high example, a middle example, and a low example of each score point were included in each ELA anchor set. In mathematics anchor sets, each score point included examples of different ways the score point could be achieved. The anchor set, the rubric, and the passage for ELA composed the scoring guide for each item. The scoring guide served as raters' primary scoring reference and was used by raters throughout the field test scoring project.

Training and Qualifying Sets—The training and qualifying sets also contained student responses that were consensus-scored by Georgia rangefinding committee members. Raters were instructed on how to apply the scoring guidelines and were required to demonstrate a clear comprehension of each anchor set by performing well on the associated training and qualifying materials. Qualification rater requirements were the same for field test scoring as they were for operational scoring (see Appendix L). Qualification was on a per-item basis; if a rater failed to qualify on a particular item, he or she was not allowed to score any student responses for that item.

Responses were selected for training to show raters the range of each score point (e.g., high, mid, and low 2s). Examples of 0s were also included for all mathematics and ELA items that included a score point of 0. Conducting field test training in this manner helped raters recognize the various ways that a student could respond to earn each score point outlined and defined in the scoring guidelines. Additionally, this training process ensured that the scoring decisions made by the rangefinding committees informed Georgia field test scoring in a direct, tangible manner.

5.5 Field Test Scoring

5.5.1 Staffing

DRC's staffing philosophy for the 2018–2019 Georgia Milestones field test placed a heavy emphasis on consistency of leadership and continuity of the program. SDs were selected for this project based on their previous experience with Georgia Milestones content. SDs selected for the field test had helped oversee some portion of operational scoring within the same content area and at the same or similar grade-level or course during a spring or winter administration. In most circumstances, the SD responsible for training and scoring the field test either facilitated rangefinding for that grade/content area or course or was present as an observer at rangefinding. In instances in which the SD wasn't at rangefinding, there was a constant line of communication open between the rangefinding SD and the scoring SD to ensure committee decisions were fully understood and followed. This consistency of leadership ensured that committee scoring

decisions were implemented as seamlessly as possible during field test scoring. The majority of team leaders selected for field test work reprised their roles from a spring or winter administration. Raters for this project were selected on the basis of their handscoring experience within the content area.

5.5.2 Training and Qualification Process

The training and qualification process for the Georgia Milestones field test closely mirrored the process employed during Georgia Milestones operational scoring. Raters made use of a scoring guide that contained a copy of the question, the passage for ELA, the rubric, and three anchor responses per score point. The SDs thoroughly reviewed the scoring guide with raters, taking and answering their questions, before raters scored a 10-response training set. Upon completion of the training set, SDs gave raters annotated paper copies of the training set and reviewed the correct scores and scoring rationales. Again, raters had the opportunity to ask questions before attempting to qualify on each item. To score student responses, raters had to meet the same accuracy standard that is required during operational administrations on one or both qualifying sets. Specifically, raters had to achieve 80% agreement with true scores, without any non-adjacent scores, on at least one qualifying set. These parameters were used for all ELA and mathematics items field tested in 2019, with the exception of ELA narrative prompts for which a 70% qualifying threshold was employed. After qualifying on an item, raters scored it to completion before training and qualification on the next item. This process was repeated until all items for their designated grade/content area or course had been scored. Each rater had to qualify on each item separately to score it.

5.5.3 Field Test Scoring

Field test scoring for Georgia Milestones was conducted at the following locations:

- Woodbury, Minnesota
- Plymouth, Minnesota
- Cincinnati, Ohio
- Indianapolis, Indiana

Embedded field test scoring began on June 21, 2019, for all Georgia Milestones items and concluded by July 3, 2019. All items were double scored at a rate of 25%, with the first score being the score of record. There were no resolution reads for non-adjacent scores. Validity responses were dealt to every rater scoring an item. The scoring data generated during field test scoring will be used in subsequent steps of the constructed-response item development process, including operational item selection.

5.6 Operational Scoring

Operational handscoring for the 2018–2019 Georgia Milestones assessments occurred within the following time frames.

- EOG spring: April 8–May 24, 2019

- EOC fall mid-month:
  - Window 1: August 9–28, 2018;
  - Window 2: September 12–25, 2018;
  - Window 3: October 11–23, 2018;
  - Window 4: November 7–20, 2018
- EOC winter: November 27, 2018–January 9, 2019
- EOC spring mid-month:
  - Window 1: January 18–30, 2019;
  - Window 2: February 14–26, 2019;
  - Window 3: March 12–26, 2019
- EOC spring: April 23–June 15, 2019
- EOC summer: June 21–July 26, 2019

5.6.1 Rater Recruitment and Qualifications

DRC retains a significant number of raters from year to year; the overall return rate in 2019 was 50%. The pool of experienced Georgia raters was drawn from staff who scored the 2017–2018 Georgia Milestones or other similar assessments. To complete the rater staffing for this project, DRC placed advertisements in local newspapers and utilized a variety of recruiting sites. Open houses were held and applications for rater positions were screened by DRC's recruiting staff. Candidates were personally interviewed by DRC staff. In addition, each candidate was required to provide an on-demand writing sample, an on-demand mathematics sample, references, and proof of a four-year college degree. In this screening process, preference was given to candidates with previous experience scoring large-scale assessments and degrees emphasizing expertise in mathematics or ELA. Thus, the rater pool consisted of educators and other professionals with content-specific backgrounds. These individuals were valued for their content-specific knowledge but were required to set aside their own biases about student performance and accept the scoring standards outlined in the Georgia Milestones training materials. Due to changes in local economies that resulted in hiring shortfalls, staffing partners were used in some locations to augment hiring.

5.6.2 Leadership Recruitment and Qualifications

SDs and team leaders were selected from a pool of employees who displayed expertise as raters and leaders on previous DRC projects. These individuals had strong backgrounds in mathematics or ELA and demonstrated organizational, leadership, communication, and management skills. A majority of SDs and team leaders had at least five years of experience working on large-scale assessments, including Georgia Milestones. All SDs, team leaders, and raters were required to sign confidentiality agreements before handling secure materials.

Each room of raters was assigned an SD. The SD led all handscoring activities for the duration of the project. SDs assisted in rangefinding, worked with supervisors to create training materials, conducted team leader training, and were responsible for training raters. The SD made

sure that handscoring quality control reports (see Section 5.11) were available, and the SD interpreted those reports for raters. The SD also supervised the team leaders. SDs were monitored by the project director and project managers.

Once raters were qualified, team leaders were responsible for maintaining the accuracy and workload of each team member. Accuracy was maintained through a variety of means, including monitoring and the validity process (see Section 5.10), both of which were used to identify individual raters having difficulty scoring accurately. These raters received one-on-one retraining from the team leader and/or the SD. Any rater who could not be successfully retrained had his or her scores purged and was released from the project.

5.7 Operational Training Materials

Operational training materials for all 2018–2019 Georgia Milestones items were pulled forward from each item's previous administration. For the majority of constructed-response operational items, the item's previous use occurred during field testing. If an item's field test training materials needed supplementation (i.e., more examples of student responses in the anchor set), this happened by means of joint phone calls or Skype meetings between DRC and the GaDOE. These calls or meetings about the development of operational materials made use of a process similar to that of the pre-rangefinding calls described in Section 5.3. SDs searched out and sought confirmation on supplemental student responses appropriate for inclusion in anchor, training, and/or qualifying sets. The calls took place from March 20 to 21, 2019, for the EOG spring administration. No supplementation of anchor sets was needed for the EOC operational items in the winter or spring administrations.

Because operational training materials were pulled from a previous administration, the same types of sets were used in operational training as are used in field test handscoring training. These sets included anchor sets (located within the scoring guide), training sets, and qualifying sets. In operational scoring, as in field test scoring, the raters constantly referenced the scoring guide throughout the training and scoring processes. The training sets provided raters with the opportunity to practice using what they learned from the discussion of the anchor sets to score sample responses before attempting to qualify. The qualifying sets were the means by which raters determined whether they could score operational responses. Anchor and training sets were available to be consulted throughout the qualifying process. This helped raters who were trying to qualify recognize the various ways that a student could respond to earn each score point outlined and defined in the scoring guidelines.

5.8 Operational Training Process

Each SD conducted a team leader training session before training raters. This session followed the same procedures as rater training. During team leader training, all Georgia Milestones materials, including pertinent portions of the Georgia scoring philosophy, were reviewed and discussed. Team leaders were required to annotate training materials with rubric language and/or committee rationales generated during rangefinding meetings. To facilitate scoring consistency, it was imperative that all team leaders imparted the same rationale for each response. Once the team leaders were qualified, leadership responsibilities were reviewed and team assignments were given. A ratio of one team leader per 6–11 raters ensured sufficient monitoring rates for team members.

Rater training began with the SD providing an intensive review of the scoring guidelines and anchor papers. Next, raters practiced by independently scoring the responses in the training set. After all the raters in the room completed the training set, the SD led a thorough discussion of the responses.

Once the scoring guidelines, anchor sets, and training sets were thoroughly discussed, each rater was required to demonstrate understanding of the scoring criteria by qualifying on at least one of the qualifying sets. The acceptable level of agreement needed to qualify varied by item type:

- For mathematics extended constructed-response items and constructed-response items and for ELA constructed-response items, the acceptable qualifying rate was at least 80% agreement with true scores, without any non-adjacent scores, on at least one qualifying set.
- For writing prompts and narrative items, the acceptable qualifying rate was at least 70% agreement with true scores, without any non-adjacent scores, on at least one qualifying set.

These qualifying rates and stipulations have been approved by the GaDOE and are considered contractual.

Raters who failed to qualify on the first qualifying set were given additional training. Raters who did not perform at the required level of agreement by the end of the qualifying process were not allowed to score any student responses. These individuals were removed from the pool of potential raters in DRC's imaging system for that item. If a rater failed to qualify on all items scored in a particular room, he or she was released from the project.

5.9 Handscoring Process

Raters scored the imaged Georgia student responses at DRC scoring centers in Cincinnati and Columbus, Ohio; Indianapolis, Indiana; Madison, Wisconsin; Philadelphia, Pennsylvania; and Plymouth and Woodbury, Minnesota. Additional responses were scored at the Georgia Center for Assessment and at temporary scoring centers in Orlando, Florida, and Atlanta, Georgia. In all locations, raters were seated at tables with individual imaging stations. Image distribution was controlled, ensuring that student images were sent only to designated groups of raters qualified to score those items. Imaged student responses were electronically separated for routing to individual raters by item. Raters were provided only with student responses that they were qualified to score. Scores were keyed into DRC's imaging system.

Georgia Milestones assessments utilized three distinct sets of scoring parameters, or rules, in the 2018–2019 administration. The set of parameters used depended on the assessment (i.e., EOG, EOC) and the administration (e.g., spring, fall, mid-month). All Georgia student responses were scored independently by individual raters, regardless of the scoring parameters used. Additionally, each student response was scored within a predetermined turnaround time (TAT). This meant that once a student response entered DRC's handscoring system, that response had to be scored within a specified number of days. Table 5.5 summarizes the three sets of Georgia Milestones scoring parameters (including TATs) and their application in terms of assessment and administration.

Table 5.5: 2018–2019 Georgia Milestones Operational Scoring Parameters

Administration	Score of Record	Percentage Double Read	Handscoring Turnaround Time	Targeted Rescores
EOG spring	First rater's score	Grades 3, 5, and 8: 25 Grades 4, 6, and 7: 10	**Grades 3, 5, and 8:** 9 days **Grades 4, 6, and 7:** 14 days	No
EOC winter, spring	First rater's score	10	**All courses:** 9 days	Yes
EOC summer, spring/ fall mid-month	When first and second scores agree, the agreed-upon score; when first and second scores are adjacent, the higher of the two scores; when first and second scores are non-adjacent, determined by a resolution read	100	**All courses:** 9 days	No

For all administrations of the Georgia Milestones assessments, the second rater's score was used to calculate the exact, adjacent, and non-adjacent agreement rates. The responses selected for a second read were randomly chosen by the imaging system at the item level when the percentage of double reads was below 100%.

To facilitate expedited scoring for the EOC spring administration, all students whose scale score for a given course was below and adjacent to one of the two designated cut scores (i.e., *Developing Learner* or *Proficient Learner*) received a targeted second read by an expert scorer. This targeted second read became the score of record in all cases when it was higher than the previous score of record. Expert scorers were SDs, assistant SDs, team leaders, and raters, who were selected for their scoring accuracy. Targeted second reads occurred within 24 hours of the determination of their necessity.

To ensure that spring EOC examination results were available prior to commencement ceremonies, instances for identified seniors were placed into a dedicated priority category (i.e., Priority 2) beginning on May 15, 2019. SDs completed scoring of all the responses found in this priority category for each item at least once each day, guaranteeing that each senior's responses were scored, locked, and released for reporting by the end of the day following the student's test session.

For all assessments and administrations, additional monitoring was performed by team leaders and SDs to further ensure reliability. Monitoring levels were generally set between 10% and 20% per rater.

To handle possible alerts, DRC's imaging system allows raters to forward responses needing further attention to the SD. These alerts were reviewed by project management staff, who then notified the GaDOE of the occurrences. At no time in the alert process do raters, or other DRC handscoring staff, acquire any knowledge concerning a student's personal identity. Here is a list of the response attributes that result in an alert status:

- Evidence of possible plagiarism or educator interference
- Evidence of a student at risk of physical harm, neglect, suicide, mental or sexual abuse, drug abuse, criminal activity, etc.
- Specific means of direct contact provided by the student

5.10 Handscoring Validity Process

One of the training tools PAS utilized to ensure rater accuracy was the validity process. The goal of the validity process is to ensure that scoring standards are maintained. Specifically, the objective is to make sure that raters score student responses in a manner consistent with the Georgia scoring philosophy and scoring standards both within a single administration of Georgia Milestones testing and across consecutive administrations. During the scoring of the 2018–2019 Georgia Milestones assessments, scoring consistency was maintained, in part, through the validity process.

The validity process was initiated with the onset of "live" scoring (i.e., the scoring of previously unscored student responses). All validity papers were chosen and had scores verified by the SD. The scores on validity papers are considered true scores. The validity papers were then implemented to test rater accuracy. Ten validity papers were fed into each rater's response queue on each full day of scoring. The responses were selected within the imaging system and dispersed intermittently to raters.

Raters were unaware that validity responses dealt to them were pre-scored validity papers and assumed that they were scoring live student responses. This helped bolster the internal validity of the process. It is important to note that all raters who received validity papers had already successfully completed the training and qualifying process.

Next, the scores that raters assigned to the validity papers were compared to the true scores to determine the validity of raters' scores. For each item, the percentage of exact agreement and the percentage of high and low scores were computed. These data were accessed through the Validity Item Detail Report. The same sorts of data were also computed for each specific rater and were accessed through the Validity Reader Detail Report. Both of these may be run as daily or cumulative reports.

The Validity Reader Detail Report was used to identify particular raters for retraining. If a rater on a certain day generated a lower rate of agreement on a group of validity papers, it was immediately apparent in the Validity Reader Detail Report. A lower rate of agreement was defined as anything below the percentage of exact agreement with the true score required to qualify on that particular item. For example, for ELA constructed-response items, any validity agreement rate significantly below 80% would be considered lower than acceptable. Any time a rater's validity agreement rate fell below the acceptable rate, the SD was cued to examine that rater's scoring. First, the SD attempted to ascertain what kind of validity papers the rater was scoring incorrectly. This was done to determine whether there was any sort of a trend (e.g., trending low on the score point 1–2 line). Once the source of the low agreement rate was determined, the rater was retrained.

The Validity Item Detail Report was utilized to identify potential room-wide trends in need of correction. For instance, if a particular validity response with a true score of 3 was given a score of 2 by a significant number of raters within the room, that trend would be revealed in the Validity Item Detail Report. To correct a trend of this sort, the SD would look for student responses similar to the validity paper being scored incorrectly. Once located, these responses would be used in room-wide retraining, usually in the form of an annotated handout.

The exact rater agreement rate generated during the validity process was often slightly higher than the inter-rater agreement rate for the same item. The reason for this discrepancy has to do

with how validity sets are formulated. While validity papers for each item are intended to cover the full breadth of each score point, certain types of responses are generally not included in validity sets. These include line papers (i.e., examples of score points that are so close to an adjacent score point that raters are instructed to consult with a supervisor before assigning a score) and responses that, because of poor word choice or writing, are difficult to understand. The reason for these exclusions is that line papers and confusing or illegible papers often do not impart a teachable lesson. Since these types of papers are usually unique, any potential lesson the response might teach would apply only to that particular paper. Conversely, the papers in validity sets are chosen because they represent common response types and teach lessons that can be applied to other similar papers. Due to this distinction, validity sets often generate a slightly higher agreement rate than is typically generated between raters during operational scoring.

For all 2018–2019 Georgia Milestones operational items, a room-wide cumulative validity rate equal to or greater than the required qualifying percentage was sought and, in the vast majority of cases, achieved (see item-specific cumulative validity rates in the tables in Appendix L).

5.11 Quality Control

Rater accuracy was monitored throughout the scoring session by means of daily and on-demand handscoring quality control reports. These reports ensured that an acceptable level of scoring accuracy was maintained throughout the project. Inter-rater reliability (IRR) was tracked and monitored with multiple quality control reports, which were reviewed by quality assurance analysts. These reports and other quality control documents were generated at the scoring centers, where they were reviewed by the SDs, team leaders, and project managers. The following reports and documents were used during the scoring of the open-ended items:

5.11.1 The Scoring Summary Report

The Scoring Summary Report was used to monitor accuracy during the scoring process. This report incorporates two related reports, both of which can be employed with an individual or room-wide focus.

- The Reader Monitor Report monitored how often raters were in exact agreement with one another and ensured that an acceptable agreement rate was maintained. This report provided daily and cumulative exact and adjacent IRR on all double-read responses. For all Georgia Milestones operational items, a room-wide cumulative IRR rate equal to or greater than the required qualifying percentage was sought and almost always achieved (see item-specific cumulative validity rates in the tables in Appendix L).
- The Score Point Distribution Report monitored the percentage of responses at each score point. For example, daily and cumulative reports for ELA constructed-response items showed what percentages of 0s, 1s, and 2s a rater had given to all the responses scored at the time the report was produced. It also indicated the number of responses read by each rater so that production rates could be monitored.

The Item Status Report monitored the progress of handscoring. This report tracked each response and indicated the status (e.g., “not read,” “complete,” “awaiting supervisor review”).

This report was used to ensure that all responses were scored to completion within the mandated TAT.

The Reader Score Report identified all responses scored by an individual rater. This report was useful if any responses needed rescoring because of possible rater drift.

The Validity Reports (see Section 5.10) tracked how raters performed by comparing pre-scored responses to raters' scores for the same responses. If a rater's score fell below the acceptable percentage-agreement rate for that item, remediation occurred. Raters who did not retrain to the required level of agreement were released from the project.

The Read-Behind Log was used by the team leaders and SDs to monitor individual rater reliability. Team leaders read randomly selected, scored items from each team member. If the team leader disagreed with a rater's score, remediation occurred. This proved to be a very effective type of feedback because it was done with live items recently scored by a particular rater.

Recalibration Sets were employed throughout the scoring sessions to ensure accuracy by comparing each rater's scores with the true scores on a preselected set of responses. Recalibration sets helped refocus raters on particular Georgia scoring standards and response types. This check ensured there was no change in the scoring pattern as the project progressed. Raters failing to achieve an acceptable percentage of agreement with the recalibration true scores were given additional training.

CHAPTER 6:
OPERATIONAL ANALYSES: KEY CHECKING, CALIBRATION, AND SCALING

Chapter 6 of this technical report describes the operational analyses of the 2018–2019 administration of the Georgia Milestones Assessment System (Georgia Milestones). It describes the data collection design, key checking, calibration, equating, and scaling analyses required to construct score conversion tables for each form administered to students in 2019.

Chapter 6 demonstrates adherence to AERA, APA, and NCME (2014) Standards 1.8, 1.10, 2.3, 2.13, 2.14, 2.15, 2.19, 3.6, 4.14, 5.1, 5.2, 5.13, 7.2, and 7.4 in the Georgia Milestones program.

6.1 Data Collection Design

Georgia Milestones is based on the application of pre-equating to support the rapid reporting of student results. Pre-equating involves using existing item parameter estimates to construct scoring conversions between the raw scores and the Georgia Milestones scale scores prior to the administration of the assessment. The parameters used to build the pre-equated score conversions are derived from field testing and updated using state population data after the spring operational administration.

A “common item, non-equivalent groups” design was used for post-equating verification of the pre-equated scoring tables. All items were used as common linking items in post-equating verification calibration. Item parameters for the common items were derived using data collected from two different groups of students. The first set of item parameters was derived using data collected from the group of students who participated in the 2019 operational administration of Georgia Milestones. The second set of item parameters was derived using data collected from the group of students who participated in previous field test analyses from prior administrations. The equivalence of the item parameters and associated test statistics was then thoroughly evaluated within the verification analyses.

6.1.1 Operational Test Design

For the End-of-Grade (EOG) assessments, all data collection activities were conducted in spring 2019. Two operational core forms were administered to students, and each test form included a set of 15 common items used for linking purposes.

For the End-of-Course (EOC) assessments, the data collection activities spanned multiple administrations. In winter 2018, two operational forms were administered and each operational form contained 25 common items used to link the form to a common scale. In addition, common items were placed in test forms in the spring administration to connect all the test forms to a common scale of measurement.

Tables 6.1 through 6.4 present the structure of the test forms administered during the 2018−2019 administrations. Information and associated statistics for the EOG retest administration are reported in Appendix H. Information and associated statistics for the summer EOC test administration are reported in Appendix I.

Table 6.1: Structure of the 2018–2019 Georgia Milestones, English Language Arts

Grade or Course	Administration	Number of Core Forms	Number of Items per Form	Number of Core Points per Form	Number of FT Items per Form	Number of TerraNova Items per Form	Number of Dual-Purpose Items per Form
3	Spring 2019	2	60	55	7	20	10
4	Spring 2019	2	60	55	7	20	10
5	Spring 2019	2	60	55	7	20	10
6	Spring 2019	2	60	55	7	20	10
7	Spring 2019	2	60	55	7	20	10
8	Spring 2019	2	60	55	7	20	10
3	Summer 2019	1	53	55	0	20	10
5	Summer 2019	1	53	55	0	20	10
8	Summer 2019	1	53	55	0	20	10
9LCO	Winter 2018	2	60	55	7	20	10
AMLC	Winter 2018	2	60	55	7	20	10
9LCO	Spring 2019	2	60	55	7	20	10
AMLC	Spring 2019	2	60	55	7	20	10
9LCO	Summer 2019	2	60	55	7	20	10
AMLC	Summer 2019	2	60	55	7	20	10

Note: Dual-purpose items count toward both Georgia Milestones and TerraNova scores.

Table 6.2: Structure of the 2018–2019 Georgia Milestones, Mathematics

Grade or Course	Admin	Number of Core Forms	Number of Items per Form	Number of Core Points per Form	Number of FT Items per Form	Number of TerraNova Items per Form	Number of Dual-Purpose Items per Form
3	Spring 2019	2	73	58	10	20	9
4	Spring 2019	2	73	58	10	20	9
5	Spring 2019	2	73	58	10	20	9
6	Spring 2019	2	73	58	10	20	9
7	Spring 2019	2	73	58	10	20	9
8	Spring 2019	2	73	58	11	20	9
5	Summer 2019	1	55	58	0	20	9
8	Summer 2019	1	56	58	0	20	10
CALG	Winter 2018	2	73	58	9	20	8
AGEO	Winter 2018	2	73	58	10	20	9
ALG1	Winter 2018	2	73	58	10	20	9
GEOM	Winter 2018	2	73	58	11	20	10
CALG	Spring 2019	2	73	58	9	20	8
AGEO	Spring 2019	2	73	58	10	20	9
ALG1	Spring 2019	2	73	58	10	20	9
GEOM	Spring 2019	2	73	58	11	20	10
CALG	Summer 2019	2	73	58	10	20	8
AGEO	Summer 2019	2	73	58	11	20	9
ALG1	Summer 2019	2	73	58	11	20	9
GEOM	Summer 2019	2	73	58	12	20	10

Note: Dual-purpose items count toward both Georgia Milestones and TerraNova scores.

Table 6.3: Structure of the 2018–2019 Georgia Milestones, Science

Grade or Course	Admin	Number of Core Forms	Number of Items per Form	Number of Core Points per Form	Number of FT Items per Form	Number of TerraNova Items per Form	Number of Dual-Purpose Items per Form
5	Spring 2019	2	75	55	10	20	10
8	Spring 2019	2	75	55	7	20	7
BIOL	Winter 2018	2	75	55	8	20	8
PHSC	Winter 2018	2	75	55	6	20	6
BIOL	Spring 2019	2	75	55	8	20	8
PHSC	Spring 2019	2	75	55	6	20	6
BIOL	Summer 2019	2	75	55	8	20	8
PHSC	Summer 2019	2	75	55	6	20	6

Note: Dual-purpose items count toward both Georgia Milestones and TerraNova scores.

Table 6.4: Structure of the 2018–2019 Georgia Milestones, Social Studies

Grade or Course	Admin	Number of Core Forms	Number of Items per Form	Number of Core Points per Form	Number of FT Items per Form	Number of TerraNova Items per Form	Number of Dual-Purpose Items per Form
5	Spring 2019	2	75	55	9	20	9
8	Spring 2019	2	75	55	9	20	9
HIST	Winter 2018	2	75	55	8	20	8
ECON	Winter 2018	2	75	55	5	20	5
HIST	Spring 2019	2	75	55	8	20	8
ECON	Spring 2019	2	75	55	5	20	5
HIST	Summer 2019	2	75	55	8	20	8
ECON	Summer 2019	2	75	55	5	20	5

Note: Dual-purpose items count toward both Georgia Milestones and TerraNova scores.

6.2 Operational Analyses

The scaling process for Georgia Milestones is based on the application of pre-equating with post-equating verification.

6.2.1 Classical Item Analyses

An initial evaluation of item difficulty and item discrimination was conducted for all items within a comprehensive set of classical item analyses. These statistics represent the item characteristics most often used to determine whether an item functioned properly and how a group of students performed on a particular item.

Item difficulty: At the most general level, an item's difficulty is indicated by the mean score achieved by some specified group (e.g., grade-level). The mean score is found using the following formula:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \tag{6.1}$$

In the mean score formula above, the individual item scores (x_i) are summed and then divided by the total number of students with valid responses (n). For multiple-choice items, student scores are represented by 0s and 1s (i.e., 0 = wrong, 1 = right). With dichotomous scoring, the equation above also represents the number of students correctly answering the item divided by the total number of students. Therefore, this result is also the proportion correct for the item, or the *p*-value. In theory, *p*-values can range from 0.00 to 1.00 on the proportion-correct scale. For example, if an item has a *p*-value of 0.89, it means 89% of students answered the item correctly. Additionally, this value might suggest that the item was relatively easy or that the students who attempted the item were relatively high achievers.

For constructed-response items, mean scores can range from the minimum possible score (usually zero) to the maximum possible score (four points in the case of some English language Arts [ELA] and mathematics items). Sometimes a pseudo *p*-value is provided for a constructed-

response item. This pseudo *p*-value is calculated by dividing the mean item score by the maximum possible item score.

Item difficulty is an important consideration in the development of assessments because of the range of achievement levels that classify Georgia students (i.e., *Beginning Learner*, *Developing Learner*, *Proficient Learner*, and *Distinguished Learner*). Items that are either very hard or very easy provide little information about student differences in achievement. An item answered correctly by a very high percentage of students would suggest that the knowledge or skill the item measures has been mastered by most students. Conversely, an item answered correctly by a very low percentage of students would suggest that the knowledge or skill the item measures has been mastered by few students. On a standards-referenced test like Georgia Milestones, one test development goal is to include a wide range of item difficulties.

Tables 6.5 through 6.8 provide summaries of the item difficulty statistics for all items administered, across all administrations and forms, on Georgia Milestones during the 2018–2019 school year. The summaries include the number of unique items across all forms, the mean *p*-value, the standard deviation (SD), the minimum value, the maximum value, and several percentiles (i.e., P_{10}, P_{25}, P_{50}, P_{75}, and P_{90}). For ELA, mean *p*-values ranged from 0.57 to 0.70 and the SD ranged from 0.13 to 0.16. For mathematics, mean *p*-values ranged from 0.50 to 0.64 and the SD ranged from 0.14 to 0.19. For science, mean *p*-values ranged from 0.50 to 0.63 and the SD ranged from 0.14 to 0.16. For social studies, mean *p*-values ranged from 0.61 to 0.64 and the SD ranged from 0.14 to 0.15. The item difficulties seen for this testing program were consistent with the expectations set. Items with extreme values were evaluated by DRC test development staff and the Georgia Department of Education (GaDOE) to ensure there were no problems with the items. In some cases, items with extremely low or high difficulties were retained so that the test matched the test blueprints.

Table 6.5: 2018–2019 Georgia Milestones Item Difficulty Statistics, English Language Arts

Statistic	EOG						EOC	
	3	4	5	6	7	8	9LCO	AMLC
N of Items	66	65	66	67	67	67	102	112
Mean	0.57	0.62	0.63	0.63	0.61	0.67	0.67	0.70
SD	0.16	0.15	0.14	0.15	0.14	0.13	0.14	0.13
Minimum	0.31	0.29	0.33	0.30	0.26	0.37	0.34	0.35
P_{10}	0.34	0.43	0.44	0.42	0.39	0.47	0.48	0.54
P_{25}	0.43	0.50	0.54	0.48	0.53	0.57	0.57	0.62
P_{50}	0.59	0.62	0.65	0.66	0.60	0.71	0.69	0.72
P_{75}	0.69	0.76	0.72	0.76	0.72	0.78	0.79	0.80
P_{90}	0.78	0.82	0.81	0.80	0.81	0.82	0.84	0.87
Maximum	0.83	0.88	0.88	0.83	0.88	0.87	0.94	0.96

Note: The winter and spring EOC data have been combined.

Table 6.6: 2018–2019 Georgia Milestones Item Difficulty Statistics, Mathematics

Statistic	EOG						EOC			
	3	4	5	6	7	8	CALG	AGEO	ALG1	GEOM
N of Items	79	79	79	79	79	79	131	132	131	132
Mean	0.64	0.61	0.59	0.61	0.55	0.52	0.50	0.54	0.54	0.60
SD	0.16	0.19	0.16	0.15	0.17	0.15	0.16	0.14	0.16	0.16
Minimum	0.29	0.27	0.28	0.33	0.18	0.16	0.15	0.23	0.16	0.22
P_{10}	0.44	0.34	0.37	0.40	0.30	0.33	0.29	0.36	0.32	0.36
P_{25}	0.52	0.45	0.46	0.47	0.46	0.39	0.38	0.43	0.43	0.50
P_{50}	0.63	0.63	0.62	0.63	0.56	0.55	0.50	0.53	0.52	0.61
P_{75}	0.77	0.76	0.72	0.74	0.69	0.63	0.62	0.64	0.66	0.72
P_{90}	0.84	0.86	0.81	0.79	0.76	0.72	0.70	0.74	0.76	0.80
Maximum	0.93	0.94	0.89	0.91	0.92	0.83	0.89	0.93	0.92	0.95

Note: The winter and spring EOC data have been combined.

Table 6.7: 2018–2019 Georgia Milestones Item Difficulty Statistics, Science

Statistic	EOG		EOC	
	5	8	BIOL	PHSC
N of Items	86	86	131	132
Mean	0.63	0.58	0.50	0.54
SD	0.14	0.14	0.16	0.14
Minimum	0.26	0.26	0.15	0.23
P_{10}	0.45	0.39	0.29	0.36
P_{25}	0.52	0.45	0.38	0.43
P_{50}	0.64	0.58	0.50	0.53
P_{75}	0.73	0.69	0.62	0.64
P_{90}	0.81	0.75	0.70	0.74
Maximum	0.93	0.92	0.89	0.93

Note: The winter and spring EOC data have been combined.

Table 6.8: 2018–2019 Georgia Milestones Item Difficulty Statistics, Social Studies

Statistic	EOG		EOC	
	5	8	HIST	ECON
N of Items	86	86	143	148
Mean	0.62	0.62	0.64	0.61
SD	0.15	0.14	0.14	0.15
Minimum	0.32	0.31	0.32	0.28
P_{10}	0.42	0.42	0.44	0.40
P_{25}	0.50	0.53	0.54	0.49
P_{50}	0.63	0.65	0.65	0.63
P_{75}	0.71	0.71	0.75	0.74
P_{90}	0.83	0.81	0.81	0.80
Maximum	0.92	0.87	0.92	0.93

Note: The winter and spring EOC data have been combined.

Item discrimination: At the most general level, item discrimination indicates an item's ability to differentiate between high and low achievers. The discrimination index for multiple-choice items is typically referred to as the point biserial correlation coefficient. For constructed-response items, the term *item-test correlation coefficient* is sometimes used. It is expected that students with high ability (i.e., those who perform well on the Georgia Milestones assessments overall) would be more likely to answer any given item correctly, while students with low ability (i.e., those who perform poorly on the Georgia Milestones assessments overall) would be more likely to answer the same item incorrectly. For Georgia Milestones, Pearson's product-moment correlation coefficient between item scores and test scores is used to indicate discrimination. (As is commonly practiced, DRC removes the item score from the total score such that the resulting correlations will not be spuriously high.) The correlation coefficient can range from −1.0 to +1.0. If the aforementioned expectation is met, the correlation between the item score and the total test score will be both positive and noticeably large in its magnitude (i.e., well above zero), meaning the item is a good discriminator between high- and low-ability students. This should be the case for all operational test items.

In summary, the correlation will be positive in value when the mean test score of the students who answered the item correctly is higher than the mean test score of the students who answered the item incorrectly. A positive value indicates that students who chose that response had a higher-than-average mean score; a negative value indicates that students who chose that response had a lower-than-average mean score. In other words, this correlation indicates that students who did well on the test as a whole also tended to do well on the item. However, an interaction can exist between item discrimination and item difficulty. Items that were answered correctly (or incorrectly) by a large proportion of examinees (i.e., items with extreme *p*-values) can have reduced power to discriminate and, thus, can have lower correlations.

Item discrimination is an important consideration, and the use of more discriminating items on a test is associated with more precise and reliable test scores. This means that more accurate achievement level classifications will be made based on scores estimates. Reliability, confidence intervals, and achievement level classifications are further discussed in Section 8.1.

Tables 6.9 through 6.12 provide summaries of the item discrimination statistics for all items administered, across all administrations, on Georgia Milestones during the 2018–2019 school year. The summaries include the number of unique items across all forms, the mean discrimination statistics, the SD, the minimum value, the maximum value, and several percentiles (i.e., P_{10}, P_{25}, P_{50}, P_{75}, and P_{90}). For ELA, the mean discrimination statistics ranged from 0.39 to 0.44 and the SD ranged from 0.11 to 0.13. For mathematics, the mean discrimination statistics ranged from 0.40 to 0.46 and the SD ranged from 0.09 to 0.12. For science, the mean discrimination statistics ranged from 0.37 to 0.40 and the SD ranged from 0.08 to 0.09. For social studies, the mean discrimination statistics ranged from 0.38 to 0.40 and the SD ranged from 0.09 to 0.10. The item discriminations that were seen for this testing program were consistent with the expectations that were set. Items with extreme values were evaluated by DRC test development staff and the GaDOE to ensure there were no problems with the items. In some cases, items with extremely low discriminations were retained so that the test matched the test blueprints.

Table 6.9: 2018–2019 Georgia Milestones Item Discrimination Statistics, English Language Arts

Statistic	EOG						EOC	
	3	4	5	6	7	8	9LCO	AMLC
N of Items	66	65	66	67	67	67	102	112
Mean	0.40	0.40	0.41	0.41	0.39	0.43	0.44	0.41
SD	0.13	0.12	0.11	0.11	0.13	0.12	0.11	0.12
Minimum	0.09	0.18	0.14	0.13	0.11	0.16	0.17	0.09
P_{10}	0.24	0.25	0.30	0.27	0.26	0.27	0.31	0.26
P_{25}	0.31	0.31	0.34	0.33	0.32	0.33	0.37	0.34
P_{50}	0.40	0.39	0.39	0.41	0.37	0.42	0.42	0.39
P_{75}	0.48	0.47	0.47	0.47	0.46	0.51	0.48	0.48
P_{90}	0.57	0.56	0.54	0.53	0.63	0.60	0.61	0.61
Maximum	0.66	0.67	0.69	0.67	0.72	0.68	0.71	0.69

Note: The winter and spring EOC data have been combined.

Table 6.10: 2018–2019 Georgia Milestones Item Discrimination Statistics, Mathematics

Statistic	EOG						EOC			
	3	4	5	6	7	8	CALG	AGEO	ALG1	GEOM
N of Items	79	79	79	79	79	79	131	132	131	132
Mean	0.46	0.43	0.45	0.44	0.41	0.40	0.40	0.41	0.40	0.43
SD	0.10	0.10	0.09	0.10	0.12	0.09	0.10	0.12	0.10	0.12
Minimum	0.27	0.20	0.23	0.19	0.15	0.19	0.17	0.12	0.17	0.14
P_{10}	0.31	0.31	0.34	0.32	0.25	0.27	0.27	0.26	0.27	0.27
P_{25}	0.37	0.36	0.38	0.38	0.35	0.35	0.34	0.32	0.33	0.36
P_{50}	0.47	0.41	0.45	0.45	0.41	0.40	0.40	0.40	0.40	0.42
P_{75}	0.55	0.51	0.53	0.50	0.47	0.45	0.45	0.47	0.46	0.50
P_{90}	0.60	0.56	0.57	0.58	0.59	0.51	0.54	0.54	0.52	0.55
Maximum	0.65	0.68	0.69	0.71	0.73	0.72	0.73	0.79	0.69	0.76

Note: The winter and spring EOC data have been combined.

Table 6.11: 2018–2019 Georgia Milestones Item Discrimination Statistics, Science

Statistic	EOG		EOC	
	5	8	BIOL	PHSC
N of Items	86	86	141	143
Mean	0.40	0.38	0.40	0.37
SD	0.09	0.08	0.09	0.09
Minimum	0.17	0.15	0.19	0.14
P_{10}	0.30	0.27	0.28	0.27
P_{25}	0.33	0.34	0.34	0.31
P_{50}	0.40	0.39	0.41	0.37
P_{75}	0.47	0.44	0.46	0.43
P_{90}	0.51	0.48	0.52	0.48
Maximum	0.56	0.54	0.59	0.57

Note: The winter and spring EOC data have been combined.

Table 6.12: 2018–2019 Georgia Milestones Item Discrimination Statistics, Social Studies

Statistic	EOG		EOC	
	5	8	HIST	ECON
N of Items	86	86	143	148
Mean	0.38	0.39	0.40	0.38
SD	0.09	0.09	0.09	0.10
Minimum	0.20	0.05	0.18	0.10
P_{10}	0.25	0.29	0.29	0.26
P_{25}	0.32	0.34	0.34	0.32
P_{50}	0.39	0.42	0.41	0.39
P_{75}	0.43	0.45	0.47	0.44
P_{90}	0.48	0.49	0.51	0.50
Maximum	0.61	0.64	0.65	0.57

Note: The winter and spring EOC data have been combined.

6.2.2 Calibration Methods

Item calibration is the process of assigning a difficulty parameter estimate to each item on an assessment so that all items are linked to a common scale. This section briefly introduces the Rasch model, reports the results from evaluations of the adequacy of the Rasch assumptions, and summarizes the Rasch item statistics for the Georgia Milestones assessments. All calibration analyses were conducted separately by grade/content area and course.

6.2.2.1 Item Response Theory Model

The Rasch partial credit model, or RPCM (Wright & Masters, 1982), was used to calibrate Georgia Milestones items because both multiple-choice and constructed-response items were part of the assessment. The RPCM extends the Rasch model (Rasch, 1960) to both dichotomous and polytomous item data. Under the RPCM, for item i with m_i score categories, the probability of person n scoring x [x = 0, 1, 2, . . . m_i] is given by

$$P_{ni}(X = x) = \frac{\exp\sum_{j=0}^{x}(\theta_n - D_{ij})}{\sum_{k=0}^{m_i}\exp\sum_{j=0}^{k}(\theta_n - D_{ij})}, \tag{6.2}$$

where θ_n represents a student's achievement level and D_{ij} is the step difficulty of the jth step on item i. For dichotomous items, the RPCM reduces to the standard Rasch model and the single step difficulty is referred to as the item's difficulty. The Rasch model predicts the probability of person n getting item i correct as follows:

$$P_{ni}(X = 1) = \frac{\exp(\theta_n - D_i)}{1 + \exp(\theta_n - D_i)} \tag{6.3}$$

The Rasch model places both student ability and item difficulty (estimated in terms of log-odds, or logits) on the same continuum. When the model assumptions are met, the Rasch model provides estimates of a person's ability that are independent of the items employed in the

assessment and, conversely, provides estimates of item difficulty independently of the sample of examinees.

Tables 6.13 through 6.16 provide summaries of the item difficulty parameter estimates based on the Rasch model for all items administered, across all administrations, on the Georgia Milestones assessments during the 2018–2019 school year. The summaries include the number of unique items across all forms, the mean item difficulties, the SD, the minimum value, the maximum value, and several percentiles (i.e., P_{10}, P_{25}, P_{50}, P_{75}, and P_{90}). The item difficulty parameter estimates were consistent with the expectations set and were well estimated. Specifically, it was expected that the average item difficulty would be zero, the SD would be approximately one, and the item difficulties would fall within three SDs of the mean. Moreover, when the item difficulty estimates are combined into specific test forms, the resulting test characteristic curves, conditional standard errors of measurement, and test information functions match their intended targets as expected.

Table 6.13: 2018–2019 Georgia Milestones Item Difficulty Parameter Estimates, English Language Arts

Statistic	EOG						EOC	
	3	4	5	6	7	8	9LCO	AMLC
N of Items	66	65	66	67	67	67	102	112
Mean	-0.18	-0.20	-0.17	-0.03	0.10	-0.09	0.09	-0.48
SD	0.85	0.87	0.79	0.82	0.78	0.75	0.87	0.83
Minimum	-1.74	-1.95	-2.06	-1.42	-1.76	-1.45	-2.20	-2.90
P_{10}	-1.41	-1.39	-1.22	-1.10	-0.94	-1.12	-1.07	-1.51
P_{25}	-0.89	-0.88	-0.61	-0.70	-0.47	-0.61	-0.51	-1.06
P_{50}	-0.28	-0.14	-0.27	-0.15	0.09	-0.26	0.20	-0.46
P_{75}	0.47	0.52	0.43	0.68	0.57	0.57	0.70	0.08
P_{90}	1.04	0.93	0.87	1.22	1.24	1.03	1.24	0.49
Maximum	1.29	1.87	1.43	1.80	1.97	1.56	1.76	1.78

Note: The winter and spring EOC data have been combined.

Table 6.14: 2018–2019 Georgia Milestones Item Difficulty Parameter Estimates, Mathematics

Statistic	EOG						EOC			
	3	4	5	6	7	8	CALG	AGEO	ALG1	GEOM
N of Items	79	79	79	79	79	79	131	132	131	132
Mean	-0.12	-0.04	-0.09	-0.02	0.04	0.04	0.01	-0.03	0.09	-0.12
SD	1.00	1.14	0.93	0.89	0.93	0.81	0.87	0.75	0.91	0.92
Minimum	-2.59	-2.55	-2.02	-2.11	-2.47	-1.61	-2.42	-2.42	-2.47	-2.95
P_{10}	-1.45	-1.70	-1.34	-0.96	-0.98	-1.06	-1.02	-1.07	-1.16	-1.35
P_{25}	-0.73	-0.82	-0.76	-0.78	-0.61	-0.55	-0.52	-0.55	-0.55	-0.68
P_{50}	-0.04	-0.03	-0.11	-0.03	0.01	-0.00	-0.06	0.04	0.18	-0.01
P_{75}	0.62	0.92	0.65	0.81	0.53	0.73	0.67	0.53	0.66	0.35
P_{90}	1.04	1.45	1.18	1.13	1.42	1.01	1.18	0.95	1.15	1.08
Maximum	2.11	2.09	1.78	1.74	2.37	2.21	2.52	1.55	2.66	1.91

Note: The winter and spring EOC data have been combined.

Table 6.15: 2018–2019 Georgia Milestones Item Difficulty Parameter Estimates, Science

Statistic	EOG		EOC	
	5	8	BIOL	PHSC
N of Items	86	86	141	143
Mean	0.01	-0.00	0.16	0.25
SD	0.81	0.77	0.60	0.75
Minimum	-2.38	-2.43	-1.62	-1.39
P_{10}	-0.97	-0.90	-0.64	-0.77
P_{25}	-0.55	-0.55	-0.22	-0.33
P_{50}	0.00	0.00	0.24	0.32
P_{75}	0.70	0.65	0.57	0.78
P_{90}	0.98	0.94	0.89	1.17
Maximum	2.07	1.60	1.48	1.99

Note: The winter and spring EOC data have been combined.

Table 6.16: 2018–2019 Georgia Milestones Item Difficulty Parameter Estimates, Social Studies

Statistic	EOG		EOC	
	5	8	HIST	ECON
N of Items	86	86	143	148
Mean	0.22	0.22	-0.23	-0.05
SD	0.83	0.75	0.81	0.80
Minimum	-1.90	-1.37	-2.82	-2.55
P_{10}	-0.97	-0.87	-1.22	-1.04
P_{25}	-0.20	-0.22	-0.75	-0.70
P_{50}	0.28	0.14	-0.25	-0.09
P_{75}	0.82	0.76	0.32	0.63
P_{90}	1.27	1.26	0.82	1.09
Maximum	1.87	1.81	1.51	1.71

Note: The winter and spring EOC data have been combined.

6.2.2.2 Software and Estimation Algorithm

For items on the Georgia Milestones assessments using the Rasch model, parameter estimation was implemented using WINSTEPS 3.63 (Linacre, 2005). WINSTEPS uses joint maximum likelihood estimation, as described by Wright and Masters (1982).

6.2.2.3 Checking Rasch Assumptions

Since the Rasch model was the basis of all calibration, scoring, and scaling analyses associated with Georgia Milestones, the validity of the inferences from these results depends on the degree to which the assumptions of the model were met and how well the model fits the test data. Therefore, it is important to check these assumptions. This section evaluates the dimensionality of the data, local independence, and item fit. It should be noted that only operational items were analyzed, since they are the basis of student scores.

6.2.2.4 Unidimensionality

The Rasch model assumes that one dominant dimension determines the differences among students' performances. A principal component analysis (PCA) can be used to assess the unidimensionality assumption. The purpose of the analysis is to verify whether any other dominant components exist among the items. The unidimensionality assumption holds when one dominant dimension explains a substantial amount of variance in student responses.

A PCA was performed for each test form within each grade/content area and course. Tables 6.17 through 6.20 provide a summary of the percentage of variance accounted for by the first two components across all forms and administrations. For example, the proportion of variance accounted for by the first dimension ranges from 32.7% to 37.4% for the ELA assessments, whereas the proportion of variance accounted for by the second dimension is considerably smaller, ranging from 1.2% to 2.1%. The proportion of variance accounted for by additional dimensions is negligible and is not reported. Across all tests administered within Georgia Milestones, the PCA data suggest that there is a clear dominant dimension for all tests.

Table 6.17: 2018–2019 Georgia Milestones Principal Component Analysis Summary, English Language Arts

Grade or Course	Number of Items	Percentage First Dimension	Percentage Second Dimension
3	66	36.3	2.1
4	65	36.6	1.9
5	66	34.6	2.0
6	67	35.2	1.9
7	67	32.8	2.0
8	67	36.1	1.9
9LCO	102	37.4	1.4
AMLC	112	32.7	1.2

Note: The winter and spring EOC data have been combined.

Table 6.18: 2018–2019 Georgia Milestones Principal Component Analysis Summary, Mathematics

Grade or Course	Number of Items	Percentage First Dimension	Percentage Second Dimension
3	79	40.9	1.7
4	79	43.6	1.2
5	79	38.8	1.6
6	79	37.4	1.4
7	79	35.7	1.5
8	79	33.4	1.5
CALG	131	34.1	1.0
AGEO	132	32.1	1.0
ALG1	131	37.6	0.9
GEOM	132	38.8	0.9

Note: The winter and spring EOC data have been combined.

Table 6.19: 2017–2018 Georgia Milestones Principal Component Analysis Summary, Science

Grade or Course	Number of Items	Percentage First Dimension	Percentage Second Dimension
5	86	27.8	1.5
8	86	25.3	1.6
BIOL	141	23.8	1.0
PHSC	143	26.1	0.9

Note: The winter and spring EOC data have been combined.

Table 6.20: 2017–2018 Georgia Milestones Principal Component Analysis Summary, Social Studies

Grade or Course	Number of Items	Percentage First Dimension	Percentage Second Dimension
5	86	26.4	1.4
8	86	27.1	1.5
HIST	143	26.7	1.0
ECON	148	26.6	1.0

Note: The winter and spring EOC data have been combined.

6.2.2.5 Local Independence

Local independence (LI) is a fundamental assumption of item response theory. No relationship should exist between examinees' responses to different items after accounting for the abilities measured by a test. In formal statistical terms, test X that comprises items X_1, X_2, … X_n has LI with respect to the latent variable θ if, for all $\mathbf{x} = (x_1, x_2, \ldots x_n)$ and θ,

$$P\left(\mathbf{X} = \mathbf{x} \mid \theta\right) = \prod_{i=1}^{I} P\left(X_i = x_i \mid \theta\right). \tag{6.4}$$

This formula essentially states that the probability of any pattern of responses across all items ($\mathbf{x}$), after conditioning on the abilities (θ) measured by the test, should be equal to the product of the conditional probabilities across each item (cf. the multiplication rule for independent events where the joint probabilities are equal to the product of the associated marginal probabilities).

The equation above shows the condition after satisfying the strong form of LI. A weak form of local independence (WLI) was proposed by McDonald (1979). The distinction is important because many indicators of local dependency are framed by WLI. The requirement is for the conditional covariances of all pairs of item responses, conditioned on the abilities, to be equal to zero. When this assumption is met, the joint probability of responses to an item pair, conditioned on abilities, is the product of the probabilities of responses to these two items, as shown below. (This is a weaker form because higher-order dependencies among items are allowed.) Based on the WLI, the following equation can be derived:

$$P\left(X_i = x_i, X_j = x_j \mid \theta\right) = P\left(X_i = x_i \mid \theta\right) P\left(X_j = x_j \mid \theta\right). \tag{6.5}$$

Marais and Andrich (2008) pointed out that local item dependence in the Rasch model can occur in two ways that some may not distinguish. The first way occurs when the assumption of unidimensionality is violated. Other nuisance dimensions, besides a dominant dimension, determine student performance. This can be called trait dependence. The second way occurs

when responses to one item depend on responses to another item. This is a violation of statistical independence and can be called response dependence. The assumptions of unidimensionality and LI are frequently treated as one phenomenon, and many believe that once unidimensionality holds, LI also holds. By distinguishing the two sources of local dependence, one can see that while LI can be related to unidimensionality, the two sources are different assumptions and, therefore, require different tests.

Residual item correlations provided in WINSTEPS for each item pair were used to assess the local dependence among Georgia Milestones items. In general, these residuals are computed as follows. First, expected item performance based on the Rasch model is determined using ability and item parameter estimates. Next, the deviations (i.e., residuals) between the examinees' expected and observed performances are determined for each item. Finally, for each item pair, a correlation between the respective deviations is computed.

Raw residual correlations provided in WINSTEPS correspond to Yen's *Q3* index, a popular LI statistic. The expected value for the *Q3* statistic is approximately $-1/(k-1)$ when no local dependence exists, where k is test length (Yen, 1993). Thus, the expected *Q3* values should be approximately −0.02 for the Georgia Milestones assessments (since most of the Georgia Milestones assessments had more than 50 core items). Absolute index values that are greater than 0.20 indicate a degree of local dependence that probably should be examined by test developers (Chen & Thissen, 1997).

Tables 6.21 through 6.24 show the summary statistics—mean residual correlation, SD, minimum value, maximum value, and several percentiles (i.e., P_{10}, P_{25}, P_{50}, P_{75}, and P_{90})—for all the residual correlations for each content area. The total number of item pairs and the number of pairs with residual correlations greater than 0.20 are also reported in these tables. The mean residual correlations are slightly negative, and the values are close to −0.01. The vast majority of the correlations are very small, suggesting LI generally holds for the Georgia Milestones tests. Larger correlations are seen primarily in ELA, where the use of a common reading passage is the likely source of local dependence.

Table 6.21: Summary of Item Residual Correlations for 2018–2019 Georgia Milestones Assessments, English Language Arts

Statistic	EOG						EOC	
	3	4	5	6	7	8	9LCO	AMLC
N of Item Pairs	2,145	2,080	2,145	2,211	2,211	2,211	5,151	6,216
Mean	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
SD	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02
Minimum	-0.12	-0.11	-0.14	-0.14	-0.12	-0.14	-0.18	-0.14
P_{10}	-0.05	-0.04	-0.04	-0.04	-0.04	-0.04	-0.03	-0.03
P_{25}	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.02	-0.01
P_{50}	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.00	-0.00
P_{75}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P_{90}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Maximum	0.82	0.61	0.69	0.73	0.61	0.67	0.58	0.64
$>\lvert 0.20 \rvert$	2	3	2	2	2	2	4	4

Note: The winter and spring EOC data have been combined.

Table 6.22: Summary of Item Residual Correlations for 2018–2019 Georgia Milestones Assessments, Mathematics

Statistic	EOG						EOC			
	3	4	5	6	7	8	CALG	AGEO	ALG1	GEOM
N of Item Pairs	3,081	3,081	3,081	3,081	3,081	3,081	8,515	8,646	8,515	8,646
Mean	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
SD	0.03	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Minimum	-0.19	-0.11	-0.13	-0.11	-0.11	-0.09	-0.14	-0.11	-0.12	-0.16
P_{10}	-0.04	-0.04	-0.04	-0.04	-0.04	-0.03	-0.03	-0.03	-0.03	-0.03
P_{25}	-0.03	-0.02	-0.03	-0.02	-0.02	-0.02	-0.01	-0.01	-0.01	-0.01
P_{50}	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.00
P_{75}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P_{90}	0.01	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00
Maximum	0.42	0.22	0.46	0.19	0.21	0.32	0.14	0.22	0.18	0.23
$>\lvert 0.20 \rvert$	4	1	4	0	1	1	0	1	0	1

Note: The winter and spring EOC data have been combined.

Table 6.23: Summary of Item Residual Correlations for 2018–2019 Georgia Milestones Assessments, Science

Statistic	EOG		EOC	
	5	8	BIOL	PHSC
N of Item Pairs	3,655	3,655	9,870	10,153
Mean	-0.01	-0.01	-0.01	-0.01
SD	0.02	0.02	0.01	0.01
Minimum	-0.08	-0.09	-0.08	-0.09
P_{10}	-0.03	-0.03	-0.02	-0.02
P_{25}	-0.02	-0.02	-0.01	-0.01
P_{50}	-0.01	-0.01	0.00	0.00
P_{75}	0.00	0.00	0.00	0.00
P_{90}	0.00	0.00	0.00	0.00
Maximum	0.22	0.22	0.11	0.12
$>\lvert 0.20 \rvert$	1	1	0	0

Note: The winter and spring EOC data have been combined.

Table 6.24: Summary of Item Residual Correlations for 2018–2019 Georgia Milestones Assessments, Social Studies

Statistic	EOG		EOC	
	5	8	HIST	ECON
N of Item Pairs	3,655	3,655	10,153	10,878
Mean	-0.01	-0.01	-0.01	-0.01
SD	0.02	0.02	0.01	0.01
Minimum	-0.11	-0.09	-0.09	-0.11
P_{10}	-0.03	-0.03	-0.02	-0.02
P_{25}	-0.02	-0.02	-0.01	-0.01
P_{50}	-0.01	-0.01	0.00	0.00
P_{75}	0.00	0.00	0.00	0.00
P_{90}	0.00	0.00	0.00	0.00
Maximum	0.10	0.14	0.25	0.18
$>\lvert 0.20 \rvert$	0	0	1	0

Note: The winter and spring EOC data have been combined.

Tables 6.25 and 6.26 lists the item pairs with residual correlations greater than 0.20. In the ELA assessment, items with large residual correlations typically refer to the same passage or prompt. In the remaining content areas, large residual correlations tend to be between items that reflect a very specific subject matter within a given domain. Future form construction activities will look to further reduce the residual correlations where possible.

Table 6.25: 2019 EOG Item Pairs with Large Residual Correlations (> 0.20)

Content Area	Grade	Item pair	Correlation
English Language Arts	3	Item pair 1	0.8171
		Item pair 2	0.7268
	4	Item pair 1	0.6102
		Item pair 2	0.2524
		Item pair 3	0.5997
	5	Item pair 1	0.6944
		Item pair 2	0.6344
	6	Item pair 1	0.7272
		Item pair 2	0.7279
	7	Item pair 1	0.6113
		Item pair 2	0.6102
	8	Item pair 1	0.6704
		Item pair 2	0.6438
Mathematics	3	Item pair 1	0.3024
		Item pair 2	0.2176
		Item pair 3	0.4155
		Item pair 4	0.2006
	4	Item pair 1	0.2240
	5	Item pair 1	0.4583
		Item pair 2	0.2400
		Item pair 3	0.2527
		Item pair 4	0.2009
	7	Item pair 1	0.2097
	8	Item pair 1	0.3210
Science	5	Item pair 1	0.2239
	8	Item pair 1	0.2210

Table 6.26: 2018–2019 EOC Item Pairs with Large Residual Correlations (> 0.20)

Course	Item pair	Correlation
9LCO	Item pair 1	0.5767
	Item pair 2	0.5787
	Item pair 3	0.5097
	Item pair 4	0.5064
AMLC	Item pair 1	0.5081
	Item pair 2	0.5035
	Item pair 3	0.6360
	Item pair 4	0.5294
AGEO	Item pair 1	0.2215
GEOM	Item pair 1	0.2285
HIST	Item pair 1	0.2486

Note: The winter and spring EOC data have been combined.

6.2.2.6 Item Fit

WINSTEPS provides two item fit statistics (i.e., infit and outfit) for evaluating the degree to which the Rasch model predicts the observed item responses. Each fit statistic can be expressed as a mean square (MnSq) statistic or on a standardized metric (i.e., Zstd, with mean = 0 and variance = 1). MnSq values are more oriented toward practical significance, while Zstd values are more oriented toward statistical significance. Though both are informative, the Zstd values are very likely too sensitive to the large sample sizes observed on the Georgia Milestones assessments. In this situation, it is recommended that the Zstd values be ignored if the MnSq values are acceptable (Linacre, 2014).

Both infit and outfit MnSq statistics are the average of standardized residual variance (i.e., the difference between the observed score and the Rasch estimated score divided by the square root of the Rasch model variance). The outfit statistic, however, gives all examinees equal weight in computing the fit and tends to be affected more by unexpected responses far from the person, item, or rating scale category measure. The infit statistic is weighted by the examinee locations relative to item difficulty and tends to be affected more by unexpected responses close to the person, item, or rating scale category measure. Some consider that extreme infit values are a greater threat to the measurement process than extreme outfit values, since most tests are designed to measure the on-target population rather than extreme outliers.

The expected MnSq value is 1.0 and can range from zero to infinity. Deviation in excess of the expected value can be interpreted as noise or as lack of fit between the items and the model. Values lower than the expected value can be interpreted as item redundancy or overfitting items (i.e., too predictable, too much redundancy), and values greater than the expected value indicate underfitting items (i.e., too unpredictable, too much noise). Rules of thumb regarding "practically significant" MnSq values vary. More conservative practitioners might prefer items with MnSq values that range from 0.8 to 1.2. Others believe reasonable test results can be achieved with values that range from 0.5 to 1.5. In the results below, values outside of the 0.7 to 1.3 range are used to define thresholds for potential significant misfit.

Tables 6.27 through 6.30 present the summary statistics of infit and outfit MnSq statistics for the Georgia Milestones ELA, mathematics, science, and social studies assessments, including the mean, the SD, the minimum values, the maximum values, and several percentiles (i.e., P_{10}, P_{25}, P_{50}, P_{75}, and P_{90}). As can be seen, the mean values for both fit statistics were close to 1.00 for all content areas and courses. Almost all items had infit values falling in the range of 0.7 to 1.3. Though more outfit values fell outside this range than did infit values, relatively few items fell outside this range. All items flagged for potential misfit were reviewed by DRC psychometric staff. Overall, these results indicate that the Rasch model fits the Georgia Milestones item data for all content areas and courses. The model-data fit suggests that the use of the Rasch model provides an appropriate and coherent framework for all scaling, equating, and score reporting activities.

Table 6.27: 2018–2019 Georgia Milestones Summary of Infit and Outfit Mean Square Statistics, English Language Arts

Statistic	EOG												EOC			
	3		4		5		6		7		8		9LCO		AMLC	
	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
N of Items	66	66	65	65	66	66	67	67	67	67	67	67	102	102	112	112
Mean	0.99	1.01	1.00	1.01	1.00	1.02	1.00	1.00	0.99	1.00	0.99	0.99	1.00	1.01	0.99	0.99
SD	0.12	0.20	0.13	0.19	0.12	0.24	0.13	0.25	0.11	0.18	0.13	0.23	0.13	0.24	0.20	0.30
Minimum	0.76	0.60	0.79	0.71	0.76	0.64	0.79	0.65	0.69	0.65	0.80	0.56	0.69	0.50	0.53	0.35
P_{10}	0.86	0.77	0.86	0.79	0.86	0.77	0.87	0.78	0.85	0.78	0.84	0.75	0.83	0.76	0.82	0.67
P_{25}	0.90	0.84	0.90	0.85	0.94	0.87	0.92	0.85	0.92	0.90	0.89	0.82	0.94	0.87	0.89	0.84
P_{50}	0.98	0.98	0.98	0.99	1.00	1.02	0.98	0.95	0.99	0.99	0.97	0.95	1.00	0.98	0.99	0.95
P_{75}	1.05	1.17	1.07	1.16	1.06	1.09	1.08	1.16	1.04	1.11	1.08	1.16	1.07	1.15	1.06	1.12
P_{90}	1.16	1.26	1.18	1.28	1.13	1.25	1.15	1.23	1.11	1.23	1.17	1.29	1.13	1.23	1.17	1.25
Maximum	1.32	1.56	1.44	1.50	1.49	2.32	1.54	2.50	1.28	1.52	1.48	1.94	1.46	2.28	2.01	2.45

Note: The winter and spring EOC data have been combined.

Table 6.28: 2018–2019 Georgia Milestones Summary of Infit and Outfit Mean Square Statistics, Mathematics

Statistic	EOG												EOC							
	3		4		5		6		7		8		CALG		AGEO		ALG1		GEOM	
	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
N of Items	79	79	79	79	79	79	79	79	79	79	79	79	131	131	132	132	131	131	132	132
Mean	1.00	0.99	0.99	1.00	0.99	0.99	1.00	1.01	1.00	1.02	1.00	1.01	1.00	1.00	1.01	1.01	1.00	0.99	1.00	1.00
SD	0.13	0.22	0.11	0.20	0.12	0.18	0.11	0.19	0.11	0.22	0.10	0.16	0.10	0.16	0.10	0.15	0.10	0.15	0.15	0.24
Min	0.79	0.62	0.74	0.66	0.80	0.63	0.81	0.57	0.79	0.66	0.80	0.71	0.79	0.69	0.76	0.62	0.76	0.63	0.72	0.45
P_{10}	0.84	0.70	0.87	0.75	0.86	0.77	0.86	0.78	0.88	0.80	0.89	0.81	0.87	0.79	0.89	0.83	0.89	0.81	0.86	0.77
P_{25}	0.89	0.83	0.90	0.86	0.90	0.85	0.93	0.88	0.92	0.87	0.92	0.89	0.93	0.89	0.93	0.90	0.92	0.87	0.90	0.83
P_{50}	0.98	0.96	0.97	0.96	0.98	0.98	0.99	1.02	0.97	0.97	1.00	1.00	1.00	1.00	0.99	1.01	1.00	0.99	0.96	0.94
P_{75}	1.05	1.14	1.08	1.16	1.06	1.12	1.05	1.10	1.05	1.12	1.05	1.14	1.08	1.12	1.07	1.11	1.07	1.08	1.09	1.11
P_{90}	1.20	1.33	1.15	1.28	1.15	1.27	1.18	1.30	1.14	1.35	1.15	1.22	1.13	1.18	1.13	1.19	1.13	1.18	1.18	1.31
Max	1.35	1.61	1.26	1.50	1.43	1.45	1.36	1.58	1.35	1.89	1.23	1.38	1.27	1.53	1.37	1.64	1.22	1.37	1.64	1.79

Note: The winter and spring EOC data have been combined.

Table 6.29: 2018–2019 Georgia Milestones Summary of Infit and Outfit Mean Square Statistics, Science

Statistic	EOG				EOC			
	5		8		BIOL		PHSC	
	In	Out	In	Out	In	Out	In	Out
N of Items	86	86	86	86	141	141	143	143
Mean	0.99	0.98	1.00	0.99	1.01	1.00	1.01	1.01
SD	0.09	0.16	0.09	0.15	0.10	0.15	0.09	0.15
Minimum	0.82	0.63	0.84	0.64	0.81	0.62	0.82	0.65
P_{10}	0.87	0.75	0.89	0.79	0.88	0.81	0.89	0.82
P_{25}	0.93	0.86	0.93	0.88	0.94	0.92	0.94	0.92
P_{50}	0.99	1.00	0.99	0.99	0.99	0.99	1.01	1.02
P_{75}	1.06	1.09	1.04	1.07	1.07	1.09	1.07	1.10
P_{90}	1.10	1.19	1.12	1.18	1.14	1.20	1.13	1.20
Maximum	1.20	1.32	1.25	1.44	1.23	1.40	1.35	1.41

Note: The winter and spring EOC data have been combined.

Table 6.30: 2018–2019 Georgia Milestones Summary of Infit and Outfit Mean Square Statistics, Social Studies

Statistic	EOG				EOC			
	5		8		HIST		ECON	
	In	Out	In	Out	In	Out	In	Out
N of Items	86	86	86	86	143	143	148	148
Mean	1.00	0.99	1.00	1.00	1.00	0.99	1.00	1.00
SD	0.09	0.16	0.10	0.18	0.11	0.18	0.09	0.16
Minimum	0.81	0.57	0.83	0.64	0.72	0.57	0.75	0.64
P_{10}	0.89	0.77	0.87	0.78	0.86	0.76	0.89	0.80
P_{25}	0.93	0.87	0.93	0.88	0.90	0.86	0.93	0.89
P_{50}	0.99	0.99	0.97	0.98	1.00	0.98	1.00	1.01
P_{75}	1.07	1.10	1.07	1.12	1.08	1.09	1.06	1.08
P_{90}	1.13	1.22	1.16	1.23	1.12	1.23	1.12	1.21
Maximum	1.21	1.32	1.35	1.67	1.52	1.54	1.27	1.52

Note: The winter and spring EOC data have been combined.

6.2.3 Item Analyses Summaries

The item parameter estimates used within the operational assessments come from several sources, including field test items embedded within last year's operational assessments, previously administered operational items including a subset of *TerraNova* items, and unused items from previous field test activities. ELA assessments are built using a fourth source of item parameters obtained from the REBW field test that expanded the item bank with new sets of writing prompt passages and items.

All item parameters were calibrated and scaled using item response theory models described in section 6.2.2. However, the item difficulty estimates obtained within the REBW field test were further adjusted to account for the differences between this test design and the embedded field test design traditionally used to collect data on newly developed items. The REBW field test

involved the administration of items in the fall instead of the within the spring operational administration and therefore required off-grade level testing.

A common set of linking items was carefully selected from prior operational administrations and was administered within the REBW field test along with the newly developed items. This common set of linking items served to link the item parameters for the new items to a common scale of measurement. In addition, the difference in item difficulty estimates for the linking items obtained within the field test analyses and within the operation analyses specifically reflects the different data collection designs. The mean difference between the two sets of item parameters for the linking items was applied to the field test item parameters as a design effect adjustment before they were stored in the item bank.

An extensive evaluation of the items was conducted upon completion of the classical item analyses and Rasch model calibrations. The evaluation centered on statistical properties of examinees' responses to individual items. Specifically, DRC psychometric and test development staff reviewed item difficulty and item discrimination for correct and incorrect responses. DRC psychometric staff also utilized a series of item analyses based on the ability levels of students taking Georgia Milestones to further screen for potential key errors. Patterns of omitted responses that could indicate item layout issues or potential test speededness issues were evaluated. Next, extensive review of the parameter estimation within the Rasch model calibrations was conducted for each item. Items flagged for potential issues, such as being too difficult or too easy or having patterns of statistics that suggested answer key issues, were reviewed with the GaDOE to determine whether any of the items should be removed for scoring.

Summary statistics summarizing all operational items used for scoring and reporting the Georgia Milestones are presented in Tables 6.5 through 6.30. Item level information can be found in Appendix D, Tables D.1.1 through D.1.16, for the EOG assessments and in Appendix E, Tables E.1.1 through E.1.10, for the EOC assessments.

6.2.4 Calibration/Equating Methods

Georgia Milestones assessments are based on the application of pre-equating with post-equating verification analyses. That is, pre-equated scoring tables are generated prior to the administration by using existing item parameters. Then, the pre-equating is verified by a series of post-administration calibration, scaling, and equating analyses based on samples of available students.

The procedures applied to develop the pre-equated scoring tables are explained in Section 6.2.5. The resulting pre-equated scoring tables are discussed in Section 6.4.

The series of post-equating verification analyses is explained in Section 6.2.6. The results found are discussed in Section 6.5.

Additional post-operational calibration was conducted using state population data, and the results were used to update the Rasch difficulty of the operational items. The calibration is explained in Section 6.2.7, and the results are presented in Section 6.5.

6.2.5 Pre-equating

The application of pre-equating is used to support the rapid reporting of student results. The main objective of pre-equating is to produce a raw-to-scale-score conversion table for a test

before the test is administered. This allows the reporting of test results without the need to wait for equating to be conducted after the operational test administration.

DRC constructed the pre-equated scoring tables using a two-step procedure. First, the scoring conversions that translate Georgia Milestones raw scores to the corresponding ability estimates expressed on the theta scale were produced for each grade/content area or course. The raw-score-to-theta conversions, including the corresponding conditional standard error of measurement (CSEM), were produced using WINSTEPS. Second, each theta in the scoring table was re-scaled to the Georgia Milestones reporting metric using the linear transformation defined for each Georgia Milestones assessment, which is discussed in Section 6.3.1.

6.2.6 Post-equating Verification of the Pre-equated Scoring Tables

The primary focus of the post-equating verification involves comparing the pre-equated scoring tables and a corresponding set of post-equated scoring tables. These post-equated scoring tables are based on samples of students, which are discussed in Section 6.5.2. Prior to the verification analyses, a series of item and test analyses was performed using early return data, and additional key checks were performed by test development staff when items were flagged for review.

A series of WINSTEPS calibration and equating analysis cycles was performed using samples of student data gathered throughout the test administration window. An initial set of fixed-parameter equating analyses (i.e., the pre-equated solution) was implemented using parameters from the previous operational administration as horizontal linking items. Another set of WINSTEPS calibration and equating analyses (i.e., the post-equated solution) was performed after dropping anchor items that were flagged for misfit within the first set of analyses. Each analysis cycle was reviewed for problems with parameter estimation and model-data fit. Model-data fit was reviewed using the MnSq infit statistic (for details, see Section 6.2.2.6). Values outside of the 0.7 to 1.3 range were used to flag items with potential significant misfit. Analysis cycles were repeated until at least 50% of the state population was available for verification analyses.

The results of the post-equating verification analysis were used to make a decision as to whether the pre-equated scoring tables or the post-equated scoring tables would be used for scoring and reporting the 2019 Georgia Milestones. The results of the analysis are discussed in Section 6.5.

6.2.7 Local Calibration

Following the 2018–2019 administration, the local calibration analysis was carried out and the data were analyzed without equating. The primary focus of the local calibration analysis was to evaluate and document item characteristics using state population data. Then, the results of the local calibration were used to update the Rasch difficulty of the 2019 operational items.

6.3 Scaling Methods

Once the common scale of measurement was established using the concurrent calibration methods described above, a linear transformation was used to place the scores onto the reporting scale that is used for the Georgia Milestones program.

6.3.1 Scaling

The linear transformation constants—slope and intercept (i.e., A and B, respectively)—were based on the two cut scores that were from the Georgia Milestones standard settings and that were expressed on the theta scale. The linear transformation constants were also based on the pre-specified target points on the reporting scale. The target cut scores are 475 and 525 for the *Developing Learner* and *Proficient Learner* achievement levels, respectively.

To calculate the linear transformation, the following formulas are used:

$$A = (525 - 475)/(\theta_2 - \theta_1) \tag{6.6}$$

and

$$B = 525 - (A \times \theta_2), \tag{6.7}$$

where θ_1 and θ_2 refer to the ability scores corresponding to the two official cut scores and A and B are the slope and intercept, respectively. Note that the linear transformation is then applied to the *Distinguished Learner* cut score on the theta scale. The official Georgia Milestones cut scores in the theta metric, their corresponding linear transformations, the highest obtainable scale score (HOSS) values, and the lowest obtainable scale score (LOSS) values used for pre-equating and post-equating verification are shown in Appendices D and H for the EOG assessments and in Appendices E and I for the EOC assessments.

The theta estimate and associated CSEM were then expressed on the Georgia Milestones scale score (SS) by applying the linear transformation as follows:

$$SS = (\theta \times A) + B \tag{6.8}$$

and

$$SS_CSEM = \theta_CSEM \times A, \tag{6.9}$$

where θ refers to ability estimates associated with each raw score and θ_CSEM refers to the corresponding CSEM produced by WINSTEPS.

Final adjustments were made on the LOSS and HOSS values for the raw scores of zero, for perfect scores, or for any scale scores that fell outside the LOSS and HOSS values. Once these final adjustments were made on the LOSS and HOSS values, the SS_CSEMs associated with the LOSS and HOSS values were computed. For these adjustments, the LOSS and HOSS values were first converted to a theta estimate. Then, the CSEM for this theta estimate was obtained as the reciprocal of the square root of the test information function (for details, see Section 8.1.2). This CSEM was then expressed on the Georgia Milestones reporting scale by applying the linear transformation.

6.4 Scoring Table Production

WINSTEPS provides a conversion table that maps raw scores to logits (i.e., Rasch model ability estimates) for a given set of item parameters. Score conversion tables were produced for each operational form of the Georgia Milestones assessments administered during the 2018–2019 school year. Ability estimates are computed using the linear transformation defined above.

Raw-to-scale-score conversion tables for Georgia Milestones are provided in Appendix D for the spring 2019 EOG assessments; Appendix E, for the winter 2018 and spring 2019 EOC assessments.

6.5 Post-equating Verification Results

The Georgia Milestones program is based on the application of pre-equating to support the rapid reporting of student results. These tests are pre-equated prior to the administration by using existing item parameters. To verify that the pre-equated scoring tables can be used to produce student scores, a series of post-equating analyses was carried out with samples of students who participated early in the winter or spring administration test windows. Analysis cycles are repeated as more data become available throughout the test administration window until at least 50% of the data are available. It is important to note that the samples are not representative of the state demographics but are designed to provide an opportunity to evaluate whether there are any systematic differences between a post-administration equating solution and a pre-equated solution.

The GaDOE was informed of each step in the equating process. The GaDOE also replicated the pre-equating and reviewed the post-equating verification analyses results. In addition, a member of the Technical Advisory Committee (TAC) reviewed the verification results and provided feedback to the GaDOE throughout the analysis window. Based on the results of the post-equating verification analyses, it was decided that pre-equated scoring tables would be used for scoring and reporting all tests except grade 8 ELA, 9th Grade Literature & Composition, and American Literature & Composition.

After the TAC member's review, the GaDOE made the final call to approve the scoring tables. Several elements of the post-equating verification analyses are extracted from the final set of verification analyses for a given test and summarized in the sections that follow. It is important to note that these analyses are repeated as more data become available during a test window and that the information reported in this section is a snapshot taken from a much larger set of analysis activities.

6.5.1 Key Validation

Prior to calibration analyses, a series of item analyses was performed using early return data. The purpose of these analyses was to confirm the answer keys by using classical item analysis statistics. Item statistics were flagged using the statistical criteria defined in Table 6.31. In addition to the criteria listed below, DRC psychometric staff utilized a series of item analyses based on the ability levels of students taking Georgia Milestones to further screen for potential key errors. Lastly, patterns of omitted responses that could indicate item layout issues or potential test speededness issues were evaluated.

DRC test development and psychometric staff subsequently reviewed all flagged items. No key errors were identified during the key verification, and there was no detection of evidence that Georgia Milestones tests were speeded in 2019.

Table 6.31: Classical Item Analysis Flagging Criteria

Classical Item Analysis Flagging Criteria	Indicates
If *p*-value of keyed response < 0.20	overly difficult item
If *p*-value of keyed response > 0.95	overly easy item
If point biserial of keyed response < 0.20	poorly discriminating item
If point biserial of a distractor > 0.05	possible key error

6.5.2 Calibration Sample Size

A series of early return student samples were used in the post-equating verification analyses. The sample size from the early return calibration sample are listed in Tables 6.32 through 6.34.

Table 6.32: Sizes for the EOG Calibration for Early Return Student Sample

Grade	English Language Arts			Mathematics			Science			Social Studies		
	Total	Form A	Form B	Total	Form A	Form B	Total	Form A	Form B	Total	Form A	Form B
3	40,647	21,711	18,936	67,392	39,968	27,424						
4	27,603	15,067	12,536	75,204	46,770	28,434						
5	37,596	21,411	16,185	75,108	42,976	32,132	75,425	42,339	33,086	82,617	51,716	30,901
6	42,217	24,320	17,897	69,739	40,185	29,554						
7	37,052	19,629	17,423	74,195	44,702	29,493						
8	33,408	18,146	15,262	65,281	41,119	31,731	66,002	37,941	28,061	92,414	55,643	36,771

Table 6.33: Sizes for the Winter EOC Calibration for Early Return Student Sample

Course	Total Tested	Form A	Form B
9LCO	21,105	11,181	9,924
AMLC	20,769	10,677	10,092
CALG	1,711	875	836
AGEO	3,048	1,564	1,484
ALG1	14,239	7,306	6,933
GEOM	22,839	11,871	10,968
BIOL	20,955	11,003	9,952
PHSC	11,384	6,016	5,368
HIST	20,230	10,608	9,622
ECON	40,665	21,114	19,551

Table 6.34: Sizes for the Spring EOC Calibration for Early Return Student Sample

Course	Total Tested	Form A	Form B
9LCO	45,215	19,567	25,748
AMLC	42,256	22,169	20,087
CALG	14,163	8,087	6,076
AGEO	11,058	5,788	5,270
ALG1	64,616	34,430	30,186
GEOM	50,551	27,956	22,595
BIOL	73,519	38,215	35,304
PHSC	53,884	28,754	25,130
HIST	73,458	40,781	32,677
ECON	46,316	23,793	22,523

6.5.3 Reliability

Cronbach's alpha reliability coefficient (Cronbach, 1951) is a commonly used measure of internal consistency over the responses to a set of items measuring an underlying, unidimensional trait; values 0.90 or higher are considered excellent. Reliability statistics computed using data from the early return calibration sample ranged from 0.88 to 0.94 as shown in Tables 6.35 through 6.37. The magnitude of reliability statistics suggest that the internal consistency of the assessments is strong. More information regarding the reliability of the Georgia Milestones assessments can be found in Section 8.1.

Table 6.35: Reliability Estimates Based on the EOG Calibration for Early Return Student Sample

Grade	English Language Arts		Mathematics		Science		Social Studies	
	Form A	Form B	Form A	Form B	Form A	Form B	Form A	Form B
3	0.90	0.90	0.94	0.92				
4	0.90	0.89	0.93	0.92				
5	0.91	0.89	0.93	0.93	0.92	0.92	0.92	0.91
6	0.91	0.90	0.93	0.93				
7	0.90	0.89	0.93	0.91				
8	0.92	0.90	0.92	0.91	0.91	0.91	0.92	0.92

Table 6.36: Reliability Estimates Based on the Winter EOC Calibration for Early Return Student Sample

Course	Form A	Form B
9LCO	0.91	0.89
AMLC	0.89	0.88
CALG	0.90	0.90
AGEO	0.93	0.93
ALG1	0.90	0.90
GEOM	0.94	0.93
BIOL	0.91	0.92
PHSC	0.90	0.90
HIST	0.91	0.92
ECON	0.90	0.91

Table 6.37: Reliability Estimates Based on the Spring EOC Calibration Early Return Student Sample

Course	Form A	Form B
9LCO	0.91	0.91
AMLC	0.90	0.89
CALG	0.91	0.89
AGEO	0.89	0.88
ALG1	0.92	0.92
GEOM	0.92	0.92
BIOL	0.93	0.93
PHSC	0.92	0.91
HIST	0.92	0.92
ECON	0.92	0.91

6.5.4 Scoring Tables Comparison

Within each verification analysis cycle, the post-equated scoring tables were produced, compared, and contrasted with the pre-equated scoring tables. Summaries of the differences between the pre-administration and post-administration equating for each administration are provided in Tables 6.38 through 6.40. Each table shows exact agreement in achievement level classifications (shaded green), adjacent agreement in achievement level classifications (shaded yellow), where the proportion of students that do not agree at the cut scores is less than or equal to 5%, and instances (shaded red) where the proportion of students that do not agree at the cut scores is greater than 5%, for all courses.

As shown, 13 of the 16 assessments for the spring EOG administration have exact agreement for both forms, and 2 of the 16 have adjacent agreement with a difference of 5% or less on at least one form. Grade 8 ELA showed differences that exceeded 5% between the pre- and post-equated scoring tables and scoring tables based on post-administration equating were used for this subject area. For the winter EOC administration, 7 of the 10 assessments have exact agreement on both forms and 2 of the 10 assessments have exact agreement for two of the three cut scores on both of the forms and relatively small differences in student classification that result from the different scoring tables. However, the verification analyses for the American Literature tests suggested that a subset of items would be better estimated using the spring 2019 data and that scoring tables based on post-administration equating should be used for score reporting. For the spring EOC administration, 6 of the 10 assessments have exact agreement on both forms and 3 of the 10 assessments have exact agreement for two of the three cut scores on both of the forms and relatively small differences in student classification that result from the different scoring tables. However, within the verification analyses of the ELA courses, a number of items were flagged for model/data misfit, suggesting that a post-equated solution be used for reporting.

It is important to note that the higher potential for using scoring tables for the ELA assessments based on post-equating within the winter 2018 and spring 2019 test administrations was anticipated by DRC and the GaDOE. These administrations marked the first use of item statistics derived from a standalone field test study implemented in fall 2017. Adjustments to the

reporting schedule were made for the ELA assessment to allow DRC to conduct verification analyses prior to score reporting.

Table 6.38: EOG Pre- and Post-equating Achievement Level Summary/Comparison

Grade	English Language Arts		Mathematics		Science		Social Studies	
	Form A	Form B	Form A	Form B	Form A	Form B	Form A	Form B
3	(0,0,3+)	Exact	Exact	Exact				
4	Exact	(0,0,3+)	Exact	Exact				
5	Exact	Exact	Exact	Exact	Exact	Exact	Exact	Exact
6	Exact	Exact	Exact	Exact				
7	Exact	Exact	Exact	Exact				
8	(2+,4+,0)	(2-,0,0)	Exact	Exact	Exact	Exact	Exact	Exact

Note: Grade 8 ELA was post-equated after a series of verification analyses.

Table 6.39: Winter EOC Pre- and Post-equating Achievement Level Summary/Comparison

Course	Form A	Form B
9LCO	(2-,0,0)	(0,4-,0)
AMLC	(5+,8+,3+)	(0,4+,0)
CALG	Exact	Exact
AGEO	Exact	Exact
ALG1	Exact	Exact
GEOM	(0,0,2-)	(2-,0,0)
BIOL	Exact	Exact
PHSC	Exact	Exact
HIST	Exact	Exact
ECON	Exact	Exact

Note: American Literature tests were post-equated after a series of verification analyses.

Table 6.40: Spring EOC Pre- and Post-equating Achievement Level Summary/Comparison

Course	Form A	Form B
9LCO	(1-,2-,4-)	Exact
AMLC	Exact	Exact
CALG	Exact	Exact
AGEO	Exact	Exact
ALG1	Exact	Exact
GEOM	Exact	Exact
BIOL	Exact	Exact
PHSC	(0,0,2+)	Exact
HIST	Exact	(0,3-,0)
ECON	(0,3-,0)	Exact

Note: ELA courses were post-equated after a series of verification analyses.

6.5.5 Descriptive Statistics by Test

Tables 6.41 through 6.44 present a summary of raw score descriptive statistics for all test forms administered in the winter and spring administrations. The tables include the following information: form, number of operational items, number of raw score points, raw score mean, SD, reliability estimates, and the standard error of measurement (SEM). The descriptive statistics suggest that the test forms are of appropriate difficulty and scores are reliable across all forms.

Table 6.41: 2018–2019 Georgia Milestones Test-Level Descriptive Statistics, English Language Arts

Grade or Course	Form	Number of Operational Items	Number of Raw Score Points	Raw Score Mean	SD	Reliability Estimates	SEM
3	A	44	55	28.5	10.1	0.90	3.2
	B	44	55	31.1	9.7	0.90	3.1
4	A	44	55	31.4	10.3	0.91	3.1
	B	44	55	34.6	9.5	0.90	3.1
5	A	44	55	32.4	10.8	0.91	3.1
	B	44	55	34.9	9.5	0.89	3.1
6	A	44	55	32.4	10.4	0.91	3.2
	B	44	55	34.1	9.8	0.90	3.1
7	A	44	55	31.7	10.3	0.90	3.2
	B	44	55	33.9	9.6	0.89	3.2
8	A	44	55	34.4	10.9	0.92	3.0
	B	44	55	36.6	9.6	0.90	3.1
9LCO	Winter A	44	55	35.7	10.5	0.91	3.1
	Winter B	44	55	36.6	9.6	0.89	3.1
	Spring A	44	55	36.4	10.6	0.92	3.1
	Spring B	44	55	38.9	9.9	0.91	3.0
AMLC	Winter A	44	55	36.2	10.0	0.90	3.2
	Winter B	44	55	37.0	9.5	0.89	3.2
	Spring A	44	55	38.1	9.7	0.90	3.1
	Spring B	44	55	38.1	9.1	0.88	3.1

Table 6.42: 2018–2019 Georgia Milestones Test-Level Descriptive Statistics, Mathematics

Grade or Course	Form	Number of Operational Items	Number of Raw Score Points	Raw Score Mean	SD	Reliability Estimates	SEM
3	A	52	58	34.6	13.2	0.94	3.2
	B	52	58	36.8	12.1	0.92	3.3
4	A	52	58	33.4	12.5	0.93	3.4
	B	52	58	35.5	11.4	0.92	3.2
5	A	52	58	32.6	12.8	0.93	3.3
	B	52	58	34.9	12.3	0.93	3.3
6	A	52	58	33.2	12.8	0.93	3.3
	B	52	58	35.6	12.5	0.93	3.3
7	A	52	58	30.6	13.0	0.93	3.5
	B	52	58	32.4	11.6	0.91	3.4
8	A	52	58	28.3	11.8	0.92	3.3
	B	52	58	30.6	11.3	0.91	3.4
CALG	Winter A	52	58	29.1	12.1	0.92	3.4
	Winter B	52	58	29.3	12.1	0.92	3.4
	Spring A	52	58	28.4	11.7	0.91	3.5
	Spring B	52	58	29.4	10.7	0.90	3.4
AGEO	Winter A	52	58	32.7	13.1	0.93	3.4
	Winter B	52	58	34.1	13.2	0.93	3.4
	Spring A	52	58	27.2	11.3	0.90	3.6
	Spring B	52	58	28.5	10.8	0.89	3.5
ALG1	Winter A	52	58	29.7	11.2	0.90	3.5
	Winter B	52	58	30.1	10.8	0.90	3.5
	Spring A	52	58	30.7	11.7	0.91	3.4
	Spring B	52	58	32.2	11.6	0.91	3.5
GEOM	Winter A	52	58	35.8	13.0	0.94	3.3
	Winter B	52	58	36.5	12.5	0.93	3.3
	Spring A	52	58	30.5	11.5	0.91	3.4
	Spring B	52	58	31.7	11.6	0.91	3.4

Table 6.43: 2018–2019 Georgia Milestones Test-Level Descriptive Statistics, Science

Grade or Course	Form	Number of Operational Items	Number of Raw Score Points	Raw Score Mean	SD	Reliability Estimates	SEM
5	A	56	60	36.3	11.9	0.92	3.3
	B	56	60	39.0	11.2	0.92	3.2
8	A	56	60	33.3	11.3	0.91	3.4
	B	56	60	35.2	11.6	0.91	3.4
BIOL	Winter A	56	60	34.3	11.5	0.91	3.4
	Winter B	56	60	35.8	11.7	0.92	3.4
	Spring A	56	60	36.0	12.5	0.93	3.4
	Spring B	56	60	37.1	12.3	0.93	3.3
PHSC	Winter A	56	60	33.6	10.8	0.90	3.5
	Winter B	56	60	35.1	10.9	0.90	3.4
	Spring A	56	60	36.4	11.5	0.91	3.4
	Spring B	56	60	37.8	11.0	0.91	3.3

Table 6.44: 2018–2019 Georgia Milestones Test-Level Descriptive Statistics, Social Studies

Grade or Course	Form	Number of Operational Items	Number of Raw Score Points	Raw Score Mean	SD	Reliability Estimates	SEM
5	A	56	60	36.0	11.7	0.92	3.4
	B	56	60	39.0	10.7	0.90	3.3
8	A	56	60	37.5	11.6	0.92	3.3
	B	56	60	39.4	11.3	0.92	3.3
HIST	Winter A	56	60	36.1	11.3	0.91	3.4
	Winter B	56	60	37.4	11.7	0.92	3.3
	Spring A	56	60	37.8	11.9	0.92	3.3
	Spring B	56	60	38.5	11.7	0.92	3.3
ECON	Winter A	56	60	35.8	10.9	0.90	3.4
	Winter B	56	60	37.5	11.0	0.91	3.3
	Spring A	56	60	36.8	11.6	0.92	3.3
	Spring B	56	60	38.3	11.1	0.91	3.3

6.5.6 Additional Scores

Students receive a TerraNova national percentile rank score in addition to their regular scale score for Georgia Milestones. In addition, students receive a Lexile measure along with their regular scale score for the ELA assessment. Lexile measures provide a tool that links a student's reading ability with the difficulty of text material. Domain-mastery indicators are also provided to students.

6.5.7 TerraNova

TerraNova is a nationally normed achievement test that measures higher-order thinking skills and basic and applied skills in reading, language, mathematics, science, and social studies. These assessments generate norm-referenced achievement scores, criterion-referenced objective mastery scores, and achievement level scores.

Each Georgia Milestones assessment includes a subset of 20 items from TerraNova to provide some information about the performance of students in Georgia compared to other students nationally. This type of standardized assessment allows for the interpretation of test scores in relation to a specified national reference group, which usually includes other students from the same grade. All students taking the Georgia Milestones assessments receive a national percentile rank score based on their performance on the TerraNova items.

National percentile rank scores range from 1 to 99 and are commonly used for reporting TerraNova assessment results to students, their parents or guardians, and educators. A percentile rank score may be interpreted as the percentage of students in a national sample whose scores are less than or equal to a given student's scale score. For example, if a student's scale score converts to a national percentile rank score of 71, the student scored higher than approximately 71% of the students in the national reference group. It is important to note that all 20 TerraNova items contribute to a student's norm-referenced score. However, only those TerraNova items (i.e., 10 or fewer items depending on the specific assessment) that align to state content standards (as determined by Georgia educators) serve a dual purpose by also contributing to the Georgia Milestones score.

Note that a specific level of TerraNova is used for each Georgia Milestones assessment. Table 6.45 lists the specific content areas of TerraNova that are embedded within the Georgia Milestones assessments.

Table 6.45: Content Area and Grade-Level of TerraNova Items Embedded within 2018–2019 Georgia Milestones

Assessment	Georgia Milestones Content Area or Course		TerraNova Content Area	Grade-Level
EOG	English Language Arts		Reading	3–8
	Mathematics		Mathematics	3–8
	Science		Science	5 and 8
	Social Studies		Social Studies	5 and 8
EOC	English Language Arts	9LCO	Language	9
		AMLC	Language	11
	Mathematics	CALG	Mathematics	9
		AGEO	Mathematics	10
		ALG1	Mathematics	9
		GEOM	Mathematics	10
	Science	PHSC	Science	10
		BIOL	Science	9
	Social Studies	HIST	Social Studies	11
		ECON	Social Studies	12

TerraNova items are included in the Georgia Milestones assessments to provide additional information to students, parents or guardians, and educators. The TerraNova assessments are based on content standards that differ from the Georgia state-mandated content standards. Performance on the TerraNova assessments provides information about how students in Georgia compare to a national sample of students who took TerraNova, whereas performance on Georgia Milestones provides criterion-referenced information about students and their peers enrolled within the state of Georgia. Comprehensive technical documentation of the TerraNova assessments and associated normative information is available from the publisher upon request.

6.5.8 Lexile Measures

A Lexile measure, developed by MetaMetrics, is a standard score that matches a student's reading ability with the difficulty of text material. A student's Lexile measure provides a tool for teachers to use in targeting reading material for students and a tool for parents or guardians to use in selecting reading material for their children.

A Lexile measure can be interpreted as the level of book that a student can read with a 75% comprehension rate. Experts have identified the 75% comprehension rate as a level that offers the reader a certain amount of comfort while still providing a challenge. The Lexile framework map shows Lexile measures ranging between approximately 200L and 1700L.

The GaDOE and MetaMetrics conducted a research study to link Georgia Milestones scores to the Lexile scale in the summer of 2015. For this study, approximately 2,000 students took a parallel Lexile test prior to the administration of Georgia Milestones. The relationship between Lexile measures and Georgia Milestones was determined by matching the scores from the parallel Lexile test to the performance on the subsequent operational test. Conversion tables between Georgia Milestones results and Lexile measures were built as a result. DRC used these conversion tables to report a Lexile measure along with a Georgia Milestones Reading Status indicator. The conversion tables are included in Appendix D, Tables D.6.1 through D.6.6, for the

EOG assessments and in Appendix E, Tables E.10.1 and E.10.2, for the EOC assessments. Note that a code of "BR" in the reported Lexile stands for "Beginning Reading." This code is used for any text or student ability that has a Lexile measure of zero or below.

6.5.9 Domain Scores

To provide more information about student performance on Georgia Milestones, an indication of domain mastery is reported. Domain-mastery indicators provide information about a student's strengths and areas of need for different aspects of test content. Domain mastery is determined by classifying the likelihood of student proficiency on the overall assessment given student performance on the domain. Domain mastery on Georgia Milestones is reported with respect to three levels:

- Remediate Learning—A student who achieves Remediate Learning is performing well below mastery on a domain and should consider additional study or instructional opportunities on that domain. In particular, the domain performance suggests that the student has less than a 40% chance of being at or above the *Proficient Learner* cut score.
- Monitor Learning—A student who achieves Monitor Learning has not consistently demonstrated mastery-level performance on a domain, and thus, additional information should be gathered to further evaluate mastery of the domain. In particular, the domain performance suggests that the student has between a 40% and 89% chance of being at or above the *Proficient Learner* cut score.
- Accelerate Learning—A student who achieves Accelerate Learning has demonstrated achievement on the domain that is consistent with the performances of students who were at or above the *Proficient Learner* cut score on the assessments and, as such, reflects an area of strength. In particular, the domain performance suggests that the student has a 90% or greater chance of being at or above the *Proficient Learner* cut score.

Item response theory is used to compute the likelihood of student proficiency given domain performance. After calibration and equating analyses are completed, all items on a given form will have final difficulty parameter estimates as defined by the Rasch model in Section 6.2.2.1. A domain-specific raw-score-to-theta conversion table is built using these parameters. The theta estimates and their corresponding CSEMs are then translated into the SS metric.

To estimate a student's probability of being above the *Proficient Learner* cut score (i.e., 525), the *z* score associated with the cut score must first be calculated as

$$z\ score = \frac{SS - Cut\ score}{SS_CSEM}. \tag{6.10}$$

Assuming a normal distribution, the probability of being above the *Proficient Learner* cut score (i.e., P[SS > 525]) equals the probability that a standard normal deviation is larger than *z*, for which standard lookup tables are available. For example, a domain-specific raw-to-scale-score conversion table indicates that a student with a scale score of 505 has a corresponding CSEM equal to 20. The (standardized) *z* score = (505 − 525)/20 = −1.0. The lookup table provides the cumulative probability of *z* under the standard normal distribution, which is 0.16. In this

example, the student's mastery classification would be "Remediate Learning" because the student's likelihood of performing at or above the *Proficient Learner* cut score given the domain performance is below 0.40.

Domain-mastery scoring tables are provided in Appendices J and K for EOG and EOC test forms, respectively. Included within each table are the corresponding raw score, theta, scale score, and domain-mastery classification for each domain in each operational form administered. However, only the domain-mastery classification is reported to students. It is important to note that domain-mastery results are less reliable when domains are measured with fewer items. Thus, when only a few items are used to measure a domain, other measures (e.g., observations, homework) should be used to confirm the results reported. The percentage of points allocated to each domain can be found in Table 2.1.

CHAPTER 7:
TEST RESULTS

Chapter 7 of this technical report contains information on the results of the multiple administrations of the Georgia Milestones Assessment System (Georgia Milestones) assessments during the 2018–2019 school year: the End-of-Grade (EOG) assessment results for spring 2019 and the End-of-Course (EOC) assessment results for winter 2018, spring 2019, and summer 2019. The scale score results, based on the scoring tables derived as described in Chapter 6, are presented here. Achievement level information is also provided. Presenting the results by achievement level translates the quantitative scale provided through scale scores into a qualitative, meaningful, and transparent description of student performance. The achievement levels are *Beginning Learner*, *Developing Learner*, *Proficient Learner*, and *Distinguished Learner*. While the scale score provides an essential quantitative reference to student performance, the achievement level information speaks directly to parents and guardians, students, and educators. When combined, scale scores, achievement levels, TerraNova norm-referenced test scores, and Lexile measures provide a comprehensive set of tools for assessing Georgia student performance by grade/content area and course. Chapter 7 also provides descriptions of score reports, data structures, and the *Interpretive Guide for Score Reports* (DRC, 2018). The AERA, APA, and NCME (2014) Standards addressed in Chapter 7 are 5.10, 5.11, 6.1, 6.2, 6.3, 6.4, 6.10, 6.12, 13.15, and 13.19.

Results presented in this chapter are based on population data in the General Research File. The results presented here may differ slightly from the official state summary report of the entire student population due to the ongoing resolution of test materials and slight differences in the application of exclusion rules. Official final results typically use school-level information with more detail for generating state summary reports than for conducting research analyses. The results in the following tables are presented as evidence of the reliability and validity of the Georgia Milestones assessments. Mid-month tests are repeated forms from a previous administration; their data are not included in this report.

7.1 Current Administration Data

Tables 7.1 through 7.4 provide summaries of the scale scores based on the state population for the EOG assessments within Georgia Milestones. Tables 7.5 and 7.6 provide summaries of the EOC assessments for the winter 2018 and spring 2019 administrations. These tables present the numbers of students, means, SDs, and percentages of students at the different achievement levels. Scale score frequency distributions can be found in Appendix D, Tables D.5.1 through D.5.16, for EOG. Scale score frequency distributions for EOC can be found in Appendix E, Tables E.8.1 through E.8.10, for the winter 2018 administration and in Appendix E, Tables E.9.1 through E.9.10, for the spring 2019 administration. Corresponding information about the EOG retest administration can be found in Appendix H, Tables H.4.1 through H.4.5. The scale score frequency distribution tables for the EOC summer administration can be found in Appendix I, Tables I.5.1 through I.5.10.

Table 7.1: 2019 Georgia Milestones Summary Statistics for EOG English Language Arts

Grade	*N* Count	Mean	SD	Percentage of Students in Each Achievement Level			
				Beginning Learner	*Developing Learner*	*Proficient Learner*	*Distinguished Learner*
3	129,231	510.8	63.7	29%	29%	28%	14%
4	133,547	513.9	56.2	25%	32%	27%	16%
5	136,513	517.3	57.0	24%	31%	34%	11%
6	136,673	516.0	67.0	26%	28%	35%	11%
7	133,259	510.9	58.1	28%	33%	31%	8%
8	124,745	518.7	53.6	20%	33%	35%	12%

Table 7.2: 2019 Georgia Milestones Summary Statistics for EOG Mathematics

Grade	*N* Count	Mean	SD	Percentage of Students in Each Achievement Level			
				Beginning Learner	*Developing Learner*	*Proficient Learner*	*Distinguished Learner*
3	129,156	525.3	51.6	18%	31%	39%	13%
4	133,486	525.3	52.3	18%	33%	36%	14%
5	136,458	517.2	56.6	24%	35%	27%	13%
6	136,626	516.5	51.3	22%	39%	27%	12%
7	132,796	521.4	55.1	22%	35%	28%	15%
8	103,388	508.3	48.2	27%	38%	27%	8%

Table 7.3: 2019 Georgia Milestones Summary Statistics for EOG Science

Grade	*N* Count	Mean	SD	Percentage of Students in Each Achievement Level			
				Beginning Learner	*Developing Learner*	*Proficient Learner*	*Distinguished Learner*
5	136,269	514.2	65.8	30%	27%	30%	13%
8	94,788	498.7	61.7	38%	30%	25%	7%

Table 7.4: 2019 Georgia Milestones Summary Statistics for EOG Social Studies

Grade	*N* Count	Mean	SD	Percentage of Students in Each Achievement Level			
				Beginning Learner	*Developing Learner*	*Proficient Learner*	*Distinguished Learner*
5	136,207	507.0	38.8	22%	47%	19%	11%
8	130,182	515.5	50.3	22%	38%	27%	14%

Table 7.5: Georgia Milestones Summary Statistics for EOC Winter 2018

Course	*N* Count	Mean	SD	Percentage of Students in Each Achievement Level			
				Beginning Learner	*Developing Learner*	*Proficient Learner*	*Distinguished Learner*
9LCO	23,448	524.1	54.0	16%	32%	40%	12%
AMLC	24,260	514.0	55.8	22%	34%	34%	9%
CALG	2,362	511.6	61.9	30%	32%	28%	10%
AGEO	3,368	538.1	77.3	24%	25%	28%	24%
ALG1	15,169	508.5	55.2	29%	36%	28%	7%
GEOM	24,339	548.5	81.0	19%	24%	29%	28%
BIOL	25,134	519.1	68.6	30%	24%	35%	10%
PHSC	12,906	506.1	61.9	33%	31%	30%	6%
HIST	21,863	518.8	56.0	23%	34%	32%	11%
ECON	44,992	524.6	66.8	23%	29%	36%	12%

Table 7.6: Georgia Milestones Summary Statistics for EOC Spring 2019

Course	*N* Count	Mean	SD	Percentage of Students in Each Achievement Level			
				Beginning Learner	*Developing Learner*	*Proficient Learner*	*Distinguished Learner*
9LCO	110,892	534.7	54.9	13%	26%	44%	17%
AMLC	95,511	519.1	57.7	19%	33%	39%	9%
CALG	17,669	506.7	55.8	31%	36%	26%	7%
AGEO	15,237	506.3	57.8	32%	33%	27%	7%
ALG1	103,477	515.3	62.8	27%	32%	30%	11%
GEOM	82,809	515.2	64.6	28%	30%	30%	11%
BIOL	103,172	527.7	75.5	28%	22%	35%	14%
PHSC	73,826	524.1	68.4	25%	28%	34%	13%
HIST	91,514	526.5	60.2	21%	30%	34%	15%
ECON	56,926	529.0	71.6	25%	27%	35%	14%

7.2 Reports

Table 7.7 presents the spring test report types and their delivery methods to school systems and the Georgia Department of Education (GaDOE). The individual score reports are designed to show classroom teachers how well students have learned the knowledge and skills outlined in the state-adopted content standards and to help parents or guardians understand their children's individual strengths and weaknesses in relation to the expectations of the state-adopted content standards. The summary score reports help the school and system staff understand the strengths and weaknesses of the school's or system's curriculum and instruction. Reports are available for certain school and system personnel to access electronically through a secure and protected site (i.e., eDIRECT and the MyGaDOE portal) and/or via paper copies.

Table 7.7: Summary of 2018–2019 Georgia Milestones Test Report Types and Delivery Methods to Systems

Spring Test Report Type/Data File	eDIRECT	MyGaDOE Portal	Paper
Individual Student Report (ISR)	X		X
Student Labels			X
Class Roster	X		
School Content Area Summary Report		X	
System Content Area Summary Report		X	
State Content Area Summary Report		X	
School Summary of All Student Populations		X	
System Summary of All Student Populations		X	
State Summary of All Student Populations		X	
Remediation and Retest Roster (.xlsx) (EOG Only)	X		
Local Coding Error (LCE) Roster (.xlsx) (EOG Only)	X		
Student Data File—System (.txt and .xlsx)	X	X	

Note: A state-level Student Data File is delivered directly to the GaDOE through a secure FTP site.

7.2.1 Description of Each Type of Report

This section provides descriptions of Individual Student Reports (ISRs), student labels, class rosters, Content Area Summary Reports, Summary Reports of All Student Populations, Remediation and Retest Rosters, and Local Coding Error (LCE) Reports. These reports provide various types of scores, including scale scores, norm-referenced scores, and Lexile measures, with the corresponding SD if the mean scale score is provided. In addition, achievement level classification, domain mastery level, and reading status are provided.

7.2.1.1 Individual Student Reports (ISR)

ISRs present students' results for each test taken and provide both an overview of students' performance and detailed information such as domain mastery level and a comparison at the school, system, and state levels. Schools receive two printed copies of the ISR for each student—one to keep at the school and one to send home to parents or guardians. Select school staff can also download electronic versions of the ISRs from eDIRECT.

ISRs provide the following information:

- Student demographic information
- Achievement level
- Scale score
- Achievement level descriptors—Brief descriptions of all four Georgia Milestones achievement levels that allow students and parents to see the full continuum of expectations
- Grade conversion score (for EOC)
- Domain mastery categories and achievement levels

- Comparison chart illustrating the student's score compared to the mean scale scores within the school, system, and state
- Comparison of the student's performance to the TerraNova norm-referenced tests
- Reading status (English language arts [ELA] only)
- Lexile information (ELA only)

A portion of a sample report is provided in Appendix F, Figure F.1.1, for EOG and Appendix G, Figure G.1.1, for EOC.

7.2.1.2 Student Labels

Student labels are designed so that each student's test results can be placed in his or her permanent record. A label was provided for every student in grades 3 through 8 and high school who participated in the 2018–2019 Georgia Milestones administrations. Each label has an adhesive backing so that it can be peeled from the sheet and placed in the student's cumulative school record. The label presents a snapshot of the student's results on Georgia Milestones. It shows the scale scores for each content area or course. For ELA, it also lists a student's reading status and Lexile score. In addition, the label indicates whether the student is a *Beginning Learner*, *Developing Learner*, *Proficient Learner*, or *Distinguished Learner* for each content area or course. If the student did not attempt the test (DNA), was present but did not attempt the test (PTNA), or had test results invalidated for any reason (IV or PIV), the appropriate code is reported in place of a scale score.

Student labels provide the following information:

- Student demographic information
- Scale score
- Grade conversion score (for EOC)
- Achievement level
- TerraNova national percentile
- Reading status (ELA only)
- Lexile score (ELA only)

Clear guidance is provided in the *Interpretive Guide for Score Reports* (2019) (see Section 2.3.5). DRC provided one label per student per grade for EOG and one label per student per course for EOC. The labels were provided in print format only. A sample label is provided in Appendix F, Figure F.1.2, for EOG and in Appendix G, Figure G.1.2, for EOC.

7.2.1.3 Class Rosters

Class rosters are generated for all EOG and EOC assessments. These reports contain demographic data and test results for each student listed on the roster. Class rosters are distributed via eDIRECT only and are accessible to system test coordinators. These reports are not produced in paper format.

For EOG assessments, class rosters are produced for each grade-level, with students listed alphabetically within the class, and provide the following information:

- Grade/content area
- Class name
- Student demographic information
- Scale scores
- Achievement levels
- Domain mastery levels
- TerraNova national percentile
- Reading status (ELA only)
- Lexile measures (ELA only)
- Writing performance (ELA only)—The numbers of points earned on the extended writing task and narrative writing response

For EOC assessments, class rosters are produced for each course, with students listed alphabetically within the class, and provide the following information:

- Course
- Class name
- Student demographic information
- Scale scores
- Achievement levels
- Grade conversion score
- Domain mastery levels
- TerraNova national percentile
- Reading status (ELA only)
- Lexile measures (ELA only)
- Writing performance (ELA only)—The numbers of points earned on the extended writing task and narrative writing response

The class roster report consists of two sections. The first section is the list of students and their test performances. Listing every student in a class may require several pages. A sample report is provided in Appendix F, Figure F.1.3, for EOG and Appendix G, Figure G.1.3, for EOC. The last section of the class roster is the summary page.

Class roster summaries provide the following information:

- Summary by content area (for EOG) or summary by course (for EOC)
  - This section provides the number of students with scores, the mean scale score, the SD, and the percentage of students scoring in each achievement level. It also provides details on students with no scores and those who are excluded based on testing purposes. Finally, this section provides the median national percentile for the class and the mean normal curve equivalent for the class.
- Domain performance
  - This section displays the percentage of students in each mastery category for each domain. The domain mastery information for ELA includes the Reading and Vocabulary domain and the Writing and Language domain. Lexile stretch band information is also included.

A sample summary page of a class roster is provided in Appendix F, Figure F.1.4, for EOG and Appendix G, Figure G.1.4, for EOC.

7.2.1.4 Content Area Summary Reports

Content Area Summary Reports are generated at the state, system, and school levels for each grade/content area and course. Each of these reports contains similar information, but comparison data are presented at different levels of aggregation. The School Content Area Summary Report provides overall performance data and domain mastery data for the school, system, Regional Education Service Agency (RESA), and state. Similarly, the System Content Area Summary Report provides overall performance data and domain mastery data for the system, RESA, and state. The State Content Area Summary Report simply provides these data at the overall state level.

Content Area Summary Reports provide the following information:

- Overall performance
- Percentage of students in the *Proficient Learner* and *Distinguished Learner* achievement levels
- Percentage of students in each achievement level
- Norm-referenced performance
- Summary data exclusions
- Performance by reading status (ELA only)
- Lexile distribution (ELA only)
- Percentage of students in each domain mastery category
- Percentage of students with each score point for the extended writing tasks (ELA only)
- Number of students with a writing condition code for the extended writing tasks (ELA only)

A sample of a Content Area Summary Report is provided in Appendix F, Figure F.1.5, for EOG and Appendix G, Figure G.1.5, for EOC.

7.2.1.5 Summary Reports of All Student Populations

Summary Reports of All Student Populations are generated at the school, system, and state levels. The reports are generated by grade/content area and course. They present summary statistics for all students and for particular groups of students.

Summary Reports of All Student Populations provide the following information:

- Demographic breakout for all students (e.g., gender and ethnicity/race)
- Number of students
- Mean scale score
- Percentage of students scoring in each achievement level
- Purpose of testing (for EOC)—Reasons for EOC students to take the Georgia Milestones assessments, including as their final exam in a course, as a retest of an assessment, or to test out of a course
- Students who receive a code instead of a scale score

A sample of a State Summary Report of All Student Populations is provided in Appendix F, Figure F.1.6, for EOG and Appendix G, Figure G.1.6, for EOC.

7.2.1.6 Remediation and Retest Rosters (EOG Only)

Remediation and Retest Rosters are generated at the school level for all students who tested in grades 3, 5, and 8. These reports indicate whether a student should receive remediation in ELA and/or mathematics and whether the student should be provided the opportunity to retest during the EOG retest administration. To be eligible for a retest in ELA, students in grades 3, 5, and 8 must have a reading status designation of *Below Grade Level*. To be eligible for a retest in mathematics, students in grades 5 and 8 must have an achievement level designation of *Beginning Learner*. Remediation and Retest Rosters are provided in an Excel format in eDIRECT and are not produced in paper format.

Remediation and Retest Rosters provide the following information:

- Class name
- Student demographic information
- Reading status—For ELA, students in grades 3, 5, and 8 receive a reading status of either *Below Grade Level* or *Grade Level or Above*.
- Mathematics status—For mathematics, students in grades 5 and 8 reach an achievement level of *Beginning Learner*, *Developing Learner*, *Proficient Learner*, or *Distinguished Learner*.

A sample of a Remediation and Retest Roster is provided in Appendix F, Figure F.1.7.

7.2.1.7 Local Coding Error (LCE) Roster (EOG Only)

LCE Rosters are generated at the system level and include all students in the system who have a local coding error. These are records that reflect a mismatch between a student's Irregularity Status—IR, IV, PIV—and the associated five-digit Irregularity Code. All LCE codes must be investigated by the system test coordinator and corrected in eDIRECT prior to final reporting.

The LCE Roster Report provides the following information:

- Class name
- Student demographic information
- Subject

A sample of an LCE Roster is provided in Appendix F, Figure F.1.8.

7.3 Student Data Files

Student data files contain demographic information for each student and student performance data for each content area and were provided to the GaDOE and each system. The EOG file contains one record for each student and contains the data for all content areas. The EOC file contains one record for every answer document or online submission. DRC provides a state-level text file to the GaDOE. DRC also provides a text file and an Excel data file for each system, which the GaDOE posts for the systems to access.

7.3.1 Student Data File—State

The layout for a state-level student data file is included in Appendix F, Table F.1.1, for EOG and Appendix G, Table G.1.1, for EOC.

7.3.2 Student Data File—System

The layout for a system-level student data file is included in Appendix F, Table F.1.2, for EOG and Appendix G, Table G.1.2, for EOC.

7.4 Score Conversion Tables between Georgia Milestones and Lexile Measures

Score conversion tables between Georgia Milestones and Lexile measures were built as a result of a linking study conducted by MetaMetrics. A description of the Lexile measures is provided in Section 6.5.8, and the conversion tables are included in Appendix D, Tables D.6.1 through D.6.6, for EOG and Appendix E, Tables E.10.1 and E.10.2, for EOC.

7.5 Interpretive Guide for Score Reports

The *Interpretive Guide for Score Reports* (2019) is written for Georgia teachers and administrators who receive score reports for the 2019 administration of Georgia Milestones. More details about this guide can be found in Section 2.3.5.

CHAPTER 8:
RELIABILITY AND VALIDITY EVIDENCE

Chapter 8 of this technical report provides evidence of the reliability and validity of the Georgia Milestones Assessment System (Georgia Milestones). The first half of Chapter 8 builds on the preceding chapter by providing estimates of the reliability of the test results. Several measures of reliability are discussed here. The chapter thus demonstrates adherence to AERA, APA, and NCME (2014) Standards 2.3, 2.7, 2.11, 2.13, 2.14, and 2.15.

The second half of Chapter 8 presents additional information to use in evaluating the validity of the Georgia Milestones program. This technical report has covered the phases of the testing cycle and highlighted the meaning and significance of the procedures, processes, and results in terms of validity and their relationship to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). The validity section covers topics in validity, demonstrating adherence to AERA, APA, and NCME (2014) Standards 1.1, 1.9, 1.11, 3.5, 3.6, 3.16, 4.0, 4.6, 4.7, 4.8, 4.10, 4.12, 7.2, and 8.4.

8.1 Reliability

The *Standards for Educational and Psychological Testing* states the following about reliability:

> The term *reliability* has been used in two ways in the measurement literature. First, the term has been used to refer to the reliability coefficients of classical test theory, defined as the correlation between scores on two equivalent forms of the test, presuming that taking one form has no effect on performance on the second form. Second, the term has been used in a more general sense, to refer to the consistency of scores across replications of a testing procedure, regardless of how this consistency is estimated or reported (e.g., in terms of standard errors, reliability coefficients per se, generalizability coefficients, error/tolerance ratios, item response theory [IRT] information functions, or various indices of classification consistency). (AERA, APA, & NCME, 2014, p. 33)

A reliable assessment is one that would produce stable scores if the same group of students were to take the same test repeatedly without any fatigue or memory of the test. As detailed below, the reliability of the 2018–2019 Georgia Milestones assessments was estimated in four ways:

- Internal consistency was assessed for all multiple-choice items using Cronbach's alpha.
- The standard error of measurement (SEM) was assessed for raw scores.
- The conditional standard error of measurement (CSEM), as the reciprocal of the square root of the test information function, was assessed for theta and the scale score of the three cuts.
- Classification consistency and accuracy were estimated for performance classification.

Combined, Cronbach's alpha, the SEM, the CSEM, classification consistency, and classification accuracy provide several ways of looking at the reliability of Georgia Milestones.

Cronbach's alpha and the SEM operate at the content level and provide estimates of reliability for examinee scores on a test, such as the grade 4 English language arts (ELA) or mathematics EOG assessments. The CSEM, classification consistency, and classification accuracy provide important information related to the Georgia Milestones achievement level classifications. These are of particular interest in the context of ensuring that students are college- and career-ready.

8.1.1 Measures of Internal Consistency

Cronbach's alpha reliability coefficient is a frequently used measure of internal consistency over the responses to a set of items measuring an underlying, unidimensional trait. The reliability coefficient alpha expresses the consistency of test scores as the ratio of true score variance to total score (i.e., observed) variance (i.e., true score variance plus error variance). A larger index would indicate that test scores were less influenced by random sources of error. The reliability coefficient is a "unitless" index, which can be compared from test to test and ranges from 0 to 1, where 0.80 is typically considered the minimal acceptable level of reliability for assessments like Georgia Milestones. While sensitive to random error associated with content sampling variability, the index is not sensitive to other types of errors, such as temporal stability or variability in performance that might occur across different testing occasions. Cronbach's alpha is computed as

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_j^2}{\sigma_X^2}\right), \tag{8.1}$$

where k is the number of items, σ_X^2 is the total score variance, and σ_j^2 is the variance of item j (Crocker & Algina, 1986).

The reliability coefficients for the forms administered within the winter and spring main Georgia Milestones assessments are summarized for all students and for demographic subgroups, and they are provided in Tables 8.1 through 8.4. The demographic subgroups include gender, ethnicity, accommodated/non-accommodated, special education, and English Learner (EL) subgroups. Reliability results for the special education and EL students are also reported in terms of accommodated/non-accommodated subgroups.

Form-specific descriptive statistics, including the number of students and the number of items, can be found in Appendix D, Tables D.1.1 through D.1.16, for the EOG assessments and Appendix E, Tables E.1.1 through E.1.10, for the EOC assessments. Looking at all examinees together in the "All" column in Tables 8.1 through 8.4, reliability ranges from 0.88 to 0.92 across grades and courses for ELA, from 0.89 to 0.94 across grades and courses for mathematics, from 0.90 to 0.93 across grades and courses for science, and from 0.90 to 0.92 across grades and courses for social studies. An alpha of 0.90 is considered to be an appropriate target for high stakes summative assessments, and the reliability for most tests exceeds that criterion. The reliability estimates for some grades of ELA and one EOC course are slightly lower than this criterion. However, the reliability estimates for all grades, content areas, and courses are well above 0.80, which is typically considered the minimum acceptable level of reliability.

Looking across the subgroup reliability statistics in Tables 8.1 through 8.4, some differences by subgroups in all grades, content areas, and courses were observed. Reliability

across male and female subgroups was above the criterion of 0.80 for both groups at all grade-levels, with the maximum difference in reliability across groups being 0.02. Reliability was well above the criterion of 0.80 for each ethnic group, with the maximum difference in reliability across all ethnic groups being 0.07. Similarly, reliability estimates were typically above the 0.80 criterion for EL students, with the average difference between all students and all EL students being 0.02. Reliability statistics were above the 0.80 criterion for special education students, with the maximum difference across groups being 0.09. Reliability statistics for accommodated students were above the 0.80 criterion, with the maximum difference in reliability across groups being 0.06.

The SEM is an index of the random variability in test scores in raw score units and is defined as follows:

$$SEM = SD\sqrt{1 - \hat{\alpha}}\,, \tag{8.2}$$

where SD represents the standard deviation of the raw score distribution and $\hat{\alpha}$ represents Cronbach's alpha, as expressed in Equation 8.1.

The overall SEM is expressed in raw score units and is a test-level statistic. The corresponding SEM information for all students and for demographic subgroups (e.g., gender, ethnicity) is summarized in Tables 8.5 through 8.8.

There were some observable differences in the SEMs by subgroups in all grade/content areas and courses. For the most part, the SEMs for female students and male students were similar. When differences existed, they were inconsistent in the direction, with the maximum difference between the groups being 0.1. Looking at the SEMs by ethnicity, the SEM was smaller for Asian students than for students of all other ethnicities, with the average difference between Asians and all other ethnicities being 0.30. In almost all grades and courses in ELA, science, and social studies, the SEMs were the same or smaller for non-accommodated students than for accommodated students, the same or larger for special education students than for all students, and the same or larger for EL students than for all students, with the average difference being 0.2 between all groups. For mathematics, the SEMs were very similar between non-accommodated and accommodated students, special education and all students, and EL students and all students, with the average difference being 0.0 between all groups.

Table 8.1: 2018–2019 Georgia Milestones Reliability Statistics by Subgroup, English Language Arts

Grade or Course	Form	All	Gender		Ethnicity				Accommodation		Special Education			English Learner		
			Female	Male	African American	Asian	Hispanic	White	Non-Accom	Accom	Non-Accom	Accom	All	Non-Accom	Accom	All
3	A	0.90	0.90	0.90	0.89	0.91	0.88	0.89	0.89	0.85	0.89	0.82	0.87	0.83	0.83	0.84
	B	0.90	0.90	0.90	0.88	0.87	0.88	0.88	0.90	0.88	0.89	0.87	0.90	0.84	0.85	0.85
4	A	0.91	0.91	0.91	0.90	0.92	0.89	0.90	0.90	0.87	0.90	0.84	0.87	0.85	0.86	0.86
	B	0.90	0.89	0.90	0.89	0.86	0.88	0.87	0.90	0.89	0.89	0.90	0.91	0.84	0.86	0.85
5	A	0.91	0.91	0.92	0.90	0.92	0.91	0.91	0.89	0.88	0.91	0.85	0.87	0.83	0.84	0.85
	B	0.89	0.89	0.89	0.89	0.87	0.87	0.88	0.89	0.87	0.90	0.87	0.89	0.81	0.81	0.81
6	A	0.91	0.90	0.91	0.89	0.92	0.90	0.90	0.88	0.87	0.88	0.84	0.86	0.85	0.82	0.83
	B	0.90	0.89	0.90	0.89	0.90	0.89	0.88	0.89	0.89	0.90	0.88	0.90	0.84	0.83	0.84
7	A	0.90	0.89	0.90	0.88	0.92	0.89	0.89	0.88	0.84	0.88	0.81	0.83	0.82	0.79	0.79
	B	0.89	0.88	0.89	0.88	0.90	0.88	0.88	0.89	0.88	0.89	0.87	0.88	0.81	0.82	0.82
8	A	0.92	0.92	0.92	0.91	0.93	0.92	0.91	0.90	0.89	0.91	0.87	0.88	0.88	0.85	0.86
	B	0.90	0.89	0.90	0.90	0.90	0.89	0.87	0.89	0.91	0.90	0.89	0.90	0.85	0.85	0.86
9LCO	Winter A	0.91	0.90	0.91	0.90	0.93	0.91	0.91	0.88	0.89	0.89	0.87	0.88	0.88	0.86	0.87
	Winter B	0.89	0.88	0.90	0.89	0.90	0.88	0.88	0.89	0.90	0.91	0.89	0.90	0.87	0.87	0.87
	Spring A	0.92	0.91	0.92	0.91	0.91	0.91	0.90	0.89	0.90	0.92	0.88	0.89	0.88	0.86	0.87
	Spring B	0.91	0.90	0.91	0.91	0.91	0.91	0.88	0.90	0.92	0.92	0.91	0.91	0.89	0.89	0.89
AMLC	Winter A	0.90	0.89	0.90	0.88	0.91	0.89	0.89	0.87	0.88	0.88	0.86	0.87	0.79	0.83	0.83
	Winter B	0.89	0.89	0.89	0.88	0.91	0.88	0.87	0.88	0.89	0.90	0.87	0.88	0.87	0.79	0.81
	Spring A	0.90	0.89	0.90	0.88	0.91	0.88	0.88	0.87	0.89	0.90	0.86	0.87	0.84	0.82	0.83
	Spring B	0.88	0.88	0.89	0.87	0.89	0.88	0.86	0.87	0.89	0.89	0.88	0.88	0.82	0.81	0.81

Note: The reliability statistics are included in the table when the *N* count is 15 or greater.

Table 8.2: 2018–2019 Georgia Milestones Reliability Statistics by Subgroup, Mathematics

Grade or Course	Form	All	Gender		Ethnicity				Accommodation		Special Education			English Learner		
			Female	Male	African American	Asian	Hispanic	White	Non-Accom	Accom	Non-Accom	Accom	All	Non-Accom	Accom	All
3	A	0.94	0.94	0.94	0.93	0.94	0.93	0.93	0.93	0.93	0.93	0.91	0.93	0.91	0.92	0.92
	B	0.92	0.92	0.93	0.92	0.90	0.91	0.91	0.92	0.92	0.92	0.93	0.93	0.90	0.91	0.90
4	A	0.93	0.92	0.93	0.91	0.93	0.91	0.92	0.92	0.90	0.92	0.88	0.91	0.89	0.89	0.90
	B	0.92	0.91	0.92	0.90	0.91	0.90	0.90	0.92	0.91	0.92	0.92	0.93	0.89	0.88	0.89
5	A	0.93	0.93	0.94	0.91	0.94	0.92	0.93	0.93	0.90	0.93	0.87	0.89	0.88	0.88	0.88
	B	0.93	0.92	0.93	0.91	0.92	0.91	0.92	0.93	0.92	0.93	0.91	0.93	0.88	0.88	0.88
6	A	0.93	0.93	0.93	0.91	0.94	0.92	0.93	0.92	0.89	0.92	0.87	0.89	0.89	0.87	0.88
	B	0.93	0.93	0.93	0.91	0.93	0.91	0.92	0.93	0.92	0.93	0.91	0.92	0.90	0.89	0.89
7	A	0.93	0.93	0.93	0.90	0.94	0.91	0.92	0.92	0.87	0.92	0.83	0.86	0.88	0.85	0.86
	B	0.91	0.91	0.92	0.89	0.92	0.89	0.90	0.91	0.89	0.92	0.87	0.90	0.87	0.88	0.88
8	A	0.92	0.92	0.92	0.90	0.94	0.91	0.92	0.91	0.88	0.92	0.86	0.87	0.89	0.87	0.88
	B	0.91	0.91	0.91	0.89	0.92	0.90	0.90	0.91	0.89	0.91	0.88	0.89	0.88	0.87	0.88
CALG	Winter A	0.92	0.92	0.92	0.89	0.93	0.89	0.91	0.92	0.87	0.93	0.82	0.87	0.93	0.69	0.83
	Winter B	0.92	0.92	0.92	0.89	0.90	0.89	0.91	0.92	0.92	0.84	0.88	0.88	0.87	0.85	0.87
	Spring A	0.91	0.90	0.91	0.87	0.93	0.90	0.91	0.90	0.84	0.92	0.80	0.83	0.87	0.82	0.84
	Spring B	0.90	0.89	0.90	0.86	0.92	0.88	0.89	0.89	0.86	0.91	0.84	0.86	0.81	0.81	0.81
AGEO	Winter A	0.93	0.92	0.94	0.90	0.95	0.93	0.93	0.93	0.91	0.90	0.84	0.85	0.68	0.83	0.81
	Winter B	0.93	0.93	0.94	0.91	0.95	0.93	0.92	0.93	0.92	0.93	0.89	0.91	0.85	0.86	0.86
	Spring A	0.90	0.89	0.91	0.87	0.93	0.88	0.90	0.90	0.84	0.88	0.80	0.81	0.83	0.82	0.83
	Spring B	0.89	0.89	0.90	0.86	0.93	0.86	0.89	0.89	0.87	0.85	0.85	0.85	0.87	0.84	0.86
ALG1	Winter A	0.90	0.90	0.90	0.88	0.92	0.89	0.90	0.90	0.86	0.89	0.83	0.85	0.83	0.87	0.87
	Winter B	0.90	0.89	0.90	0.87	0.92	0.88	0.90	0.89	0.88	0.88	0.84	0.85	0.87	0.83	0.85
	Spring A	0.91	0.91	0.92	0.89	0.93	0.90	0.91	0.91	0.86	0.91	0.82	0.85	0.89	0.85	0.86
	Spring B	0.91	0.91	0.92	0.89	0.92	0.90	0.90	0.91	0.89	0.90	0.86	0.88	0.89	0.88	0.88
GEOM	Winter A	0.94	0.93	0.94	0.91	0.91	0.92	0.93	0.93	0.91	0.91	0.86	0.88	0.91	0.88	0.89
	Winter B	0.93	0.93	0.93	0.91	0.90	0.91	0.92	0.93	0.92	0.93	0.87	0.90	0.91	0.89	0.90
	Spring A	0.91	0.91	0.92	0.88	0.92	0.90	0.91	0.91	0.87	0.91	0.82	0.84	0.90	0.85	0.86
	Spring B	0.91	0.91	0.92	0.89	0.91	0.90	0.91	0.91	0.90	0.90	0.87	0.88	0.89	0.88	0.88

Note: The reliability statistics are included in the table when the *N* count is 15 or greater.

Table 8.3: 2018–2019 Georgia Milestones Reliability Statistics by Subgroup, Science

Grade or Course	Form	All	Gender		Ethnicity				Accommodation		Special Education			English Learner		
			Female	Male	African American	Asian	Hispanic	White	Non-Accom	Accom	Non-Accom	Accom	All	Non-Accom	Accom	All
5	A	0.92	0.92	0.93	0.90	0.93	0.91	0.92	0.91	0.89	0.92	0.88	0.89	0.86	0.86	0.86
	B	0.92	0.91	0.92	0.90	0.90	0.90	0.90	0.92	0.91	0.92	0.91	0.92	0.86	0.84	0.86
8	A	0.91	0.90	0.92	0.89	0.93	0.90	0.91	0.90	0.87	0.90	0.85	0.86	0.86	0.81	0.82
	B	0.91	0.91	0.92	0.90	0.93	0.90	0.91	0.91	0.90	0.91	0.89	0.90	0.89	0.85	0.86
BIOL	Winter A	0.91	0.91	0.92	0.89	0.92	0.90	0.91	0.91	0.85	0.89	0.82	0.84	0.88	0.76	0.79
	Winter B	0.92	0.91	0.92	0.90	0.93	0.90	0.91	0.91	0.90	0.92	0.89	0.90	0.84	0.83	0.84
	Spring A	0.93	0.92	0.93	0.91	0.93	0.91	0.92	0.92	0.88	0.92	0.84	0.86	0.86	0.81	0.83
	Spring B	0.93	0.92	0.93	0.91	0.93	0.92	0.92	0.92	0.91	0.92	0.88	0.90	0.83	0.82	0.83
PHSC	Winter A	0.90	0.89	0.90	0.87	0.92	0.87	0.90	0.89	0.85	0.88	0.83	0.84	0.83	0.81	0.82
	Winter B	0.90	0.90	0.91	0.88	0.92	0.89	0.90	0.90	0.88	0.87	0.83	0.85	0.89	0.82	0.85
	Spring A	0.91	0.91	0.92	0.89	0.92	0.90	0.90	0.90	0.87	0.92	0.85	0.87	0.88	0.82	0.84
	Spring B	0.91	0.90	0.92	0.89	0.91	0.90	0.89	0.91	0.90	0.92	0.88	0.90	0.87	0.85	0.86

Note: The reliability statistics are included in the table when the *N* count is 15 or greater.

Table 8.4: 2018–2019 Georgia Milestones Reliability Statistics by Subgroup, Social Studies

Grade or Course	Form	All	Gender		Ethnicity				Accommodation		Special Education			English Learner		
			Female	Male	African American	Asian	Hispanic	White	Non-Accom	Accom	Non-Accom	Accom	All	Non-Accom	Accom	All
5	A	0.92	0.91	0.92	0.90	0.93	0.90	0.91	0.91	0.87	0.92	0.86	0.88	0.85	0.84	0.84
	B	0.90	0.90	0.91	0.89	0.89	0.89	0.89	0.90	0.89	0.92	0.90	0.92	0.85	0.84	0.85
8	A	0.92	0.91	0.92	0.90	0.92	0.91	0.91	0.90	0.88	0.92	0.86	0.88	0.86	0.81	0.83
	B	0.92	0.91	0.92	0.91	0.91	0.91	0.90	0.91	0.90	0.92	0.89	0.90	0.88	0.85	0.86
HIST	Winter A	0.91	0.90	0.92	0.89	0.92	0.91	0.91	0.90	0.88	0.90	0.87	0.87	0.90	0.80	0.83
	Winter B	0.92	0.92	0.92	0.91	0.93	0.91	0.91	0.92	0.91	0.93	0.91	0.92	0.88	0.83	0.85
	Spring A	0.92	0.92	0.93	0.90	0.93	0.91	0.92	0.92	0.90	0.92	0.88	0.89	0.84	0.83	0.84
	Spring B	0.92	0.92	0.93	0.91	0.93	0.91	0.91	0.92	0.92	0.93	0.91	0.91	0.87	0.86	0.86
ECON	Winter A	0.90	0.90	0.91	0.88	0.92	0.89	0.90	0.89	0.86	0.90	0.83	0.85	0.87	0.79	0.81
	Winter B	0.91	0.91	0.91	0.90	0.92	0.90	0.90	0.90	0.91	0.90	0.89	0.89	0.88	0.84	0.86
	Spring A	0.92	0.92	0.92	0.90	0.92	0.91	0.92	0.91	0.89	0.91	0.86	0.87	0.87	0.84	0.85
	Spring B	0.91	0.91	0.92	0.89	0.92	0.90	0.90	0.91	0.91	0.91	0.89	0.90	0.89	0.86	0.88

Note: The reliability statistics are included in the table when the *N* count is 15 or greater.

Table 8.5: 2018–2019 Georgia Milestones SEM by Subgroup, English Language Arts

Grade or Course	Form	All	Gender		Ethnicity				Accommodation		Special Education			English Learner		
			Female	Male	African American	Asian	Hispanic	White	Non-Accom	Accom	Non-Accom	Accom	All	Non-Accom	Accom	All
3	A	3.2	3.2	3.2	3.2	3.1	3.2	3.1	3.1	3.2	3.1	3.2	3.2	3.2	3.2	3.2
	B	3.1	3.1	3.1	3.2	2.9	3.1	3.0	3.1	3.2	3.1	3.2	3.1	3.2	3.2	3.2
4	A	3.1	3.1	3.1	3.2	2.9	3.1	3.0	3.0	3.2	3.0	3.2	3.2	3.1	3.2	3.2
	B	3.1	3.1	3.1	3.2	2.8	3.1	3.0	3.1	3.2	3.0	3.2	3.1	3.2	3.2	3.2
5	A	3.1	3.1	3.1	3.2	2.9	3.2	3.0	3.0	3.2	3.1	3.2	3.2	3.2	3.3	3.3
	B	3.1	3.0	3.1	3.2	2.8	3.1	3.0	3.1	3.2	3.1	3.2	3.2	3.3	3.3	3.3
6	A	3.2	3.2	3.2	3.2	2.9	3.2	3.1	3.1	3.2	3.2	3.2	3.2	3.3	3.3	3.3
	B	3.1	3.1	3.1	3.2	2.8	3.2	3.0	3.1	3.3	3.1	3.3	3.2	3.3	3.3	3.3
7	A	3.2	3.2	3.3	3.3	3.0	3.3	3.2	3.2	3.3	3.3	3.3	3.3	3.3	3.3	3.3
	B	3.2	3.1	3.2	3.2	2.9	3.2	3.1	3.1	3.3	3.2	3.3	3.3	3.3	3.3	3.3
8	A	3.0	3.0	3.1	3.1	2.7	3.1	2.9	2.9	3.2	3.1	3.2	3.2	3.3	3.2	3.2
	B	3.1	3.0	3.1	3.2	2.8	3.1	2.9	3.0	3.2	3.1	3.2	3.2	3.3	3.3	3.3
9LCO	Winter A	3.1	3.1	3.2	3.2	3.0	3.2	3.0	3.0	3.3	3.2	3.3	3.3	3.3	3.3	3.3
	Winter B	3.1	3.1	3.2	3.2	2.9	3.2	3.0	3.1	3.3	3.2	3.3	3.3	3.4	3.3	3.3
	Spring A	3.1	3.0	3.1	3.2	2.8	3.1	2.9	3.0	3.2	3.2	3.2	3.2	3.3	3.2	3.2
	Spring B	3.0	2.9	3.0	3.1	2.6	3.1	2.8	3.0	3.2	3.2	3.2	3.2	3.3	3.3	3.3
AMLC	Winter A	3.2	3.2	3.3	3.3	3.1	3.3	3.1	3.1	3.4	3.4	3.4	3.4	3.5	3.4	3.4
	Winter B	3.2	3.1	3.2	3.3	3.0	3.2	3.0	3.1	3.3	3.3	3.3	3.3	3.4	3.4	3.4
	Spring A	3.1	3.1	3.1	3.2	2.8	3.2	3.0	3.1	3.3	3.3	3.3	3.3	3.3	3.4	3.3
	Spring B	3.1	3.0	3.1	3.2	2.8	3.1	2.9	3.0	3.3	3.3	3.3	3.3	3.3	3.3	3.3

Note: The SEM estimates are included in the table when the *N* count is 15 or greater.

Table 8.6: 2018–2019 Georgia Milestones SEM by Subgroup, Mathematics

Grade or Course	Form	All	Gender		Ethnicity				Accommodation		Special Education			English Learner		
			Female	Male	African American	Asian	Hispanic	White	Non-Accom	Accom	Non-Accom	Accom	All	Non-Accom	Accom	All
3	A	3.2	3.3	3.2	3.3	2.9	3.3	3.2	3.2	3.3	3.2	3.3	3.3	3.2	3.3	3.3
	B	3.3	3.3	3.3	3.4	2.9	3.4	3.2	3.3	3.4	3.3	3.4	3.4	3.4	3.4	3.4
4	A	3.4	3.4	3.4	3.4	3.0	3.4	3.3	3.4	3.4	3.4	3.3	3.3	3.4	3.4	3.4
	B	3.2	3.2	3.2	3.3	2.8	3.3	3.2	3.2	3.3	3.2	3.3	3.3	3.3	3.3	3.3
5	A	3.3	3.3	3.3	3.4	2.9	3.4	3.2	3.3	3.4	3.3	3.4	3.4	3.4	3.4	3.4
	B	3.3	3.3	3.3	3.4	2.8	3.3	3.3	3.3	3.4	3.3	3.4	3.3	3.4	3.3	3.3
6	A	3.3	3.3	3.3	3.4	2.8	3.4	3.2	3.3	3.4	3.3	3.4	3.4	3.4	3.4	3.4
	B	3.3	3.3	3.3	3.4	2.7	3.4	3.2	3.3	3.4	3.3	3.4	3.4	3.4	3.4	3.4
7	A	3.5	3.5	3.5	3.5	3.0	3.5	3.4	3.5	3.4	3.5	3.4	3.4	3.5	3.4	3.4
	B	3.4	3.4	3.4	3.4	3.0	3.4	3.3	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.4
8	A	3.3	3.4	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.2	3.3	3.3	3.3	3.3
	B	3.4	3.4	3.4	3.4	3.2	3.4	3.3	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.4
CALG	Winter A	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.3	3.3	3.2	3.2	3.4	3.3	3.3
	Winter B	3.4	3.4	3.4	3.4	3.6	3.5	3.4	3.4	3.4	3.6	3.2	3.3	3.6	3.3	3.3
	Spring A	3.5	3.5	3.5	3.5	3.4	3.5	3.5	3.5	3.4	3.3	3.3	3.3	3.4	3.4	3.4
	Spring B	3.4	3.5	3.4	3.5	3.4	3.5	3.3	3.4	3.4	3.5	3.4	3.4	3.4	3.4	3.4
AGEO	Winter A	3.4	3.4	3.4	3.5	3.2	3.5	3.3	3.4	3.4	3.5	3.3	3.4	3.5	3.4	3.4
	Winter B	3.4	3.4	3.4	3.5	3.2	3.4	3.2	3.4	3.4	3.4	3.3	3.3	3.4	3.4	3.4
	Spring A	3.6	3.6	3.5	3.5	3.5	3.5	3.6	3.6	3.4	3.3	3.3	3.3	3.4	3.4	3.4
	Spring B	3.5	3.5	3.5	3.5	3.4	3.5	3.5	3.5	3.4	3.5	3.4	3.4	3.4	3.5	3.4
ALG1	Winter A	3.5	3.5	3.5	3.6	3.3	3.5	3.5	3.5	3.5	3.6	3.5	3.5	3.5	3.5	3.5
	Winter B	3.5	3.5	3.5	3.5	3.3	3.5	3.4	3.5	3.4	3.4	3.4	3.4	3.4	3.4	3.4
	Spring A	3.4	3.5	3.4	3.4	3.1	3.5	3.4	3.4	3.4	3.4	3.3	3.4	3.4	3.4	3.4
	Spring B	3.5	3.5	3.4	3.5	3.1	3.5	3.4	3.5	3.4	3.5	3.4	3.4	3.4	3.4	3.4
GEOM	Winter A	3.3	3.3	3.3	3.3	2.6	3.3	3.2	3.3	3.3	3.3	3.2	3.3	3.4	3.4	3.4
	Winter B	3.3	3.3	3.3	3.4	2.7	3.4	3.3	3.3	3.4	3.3	3.4	3.4	3.4	3.4	3.4
	Spring A	3.4	3.4	3.3	3.4	3.2	3.4	3.3	3.4	3.3	3.3	3.3	3.3	3.3	3.3	3.3
	Spring B	3.4	3.4	3.4	3.4	3.2	3.5	3.4	3.4	3.4	3.4	3.4	3.4	3.5	3.4	3.4

Note: The SEM estimates are included in the table when the *N* count is 15 or greater.

Table 8.7: 2018–2019 Georgia Milestones SEM by Subgroup, Science

Grade or Course	Form	All	Gender		Ethnicity				Accommodation		Special Education			English Learner		
			Female	Male	African American	Asian	Hispanic	White	Non-Accom	Accom	Non-Accom	Accom	All	Non-Accom	Accom	All
5	A	3.3	3.3	3.3	3.4	2.9	3.4	3.2	3.2	3.5	3.2	3.5	3.5	3.5	3.5	3.5
	B	3.2	3.3	3.2	3.4	2.8	3.3	3.1	3.2	3.4	3.2	3.4	3.3	3.5	3.5	3.5
8	A	3.4	3.4	3.4	3.5	3.2	3.5	3.3	3.4	3.5	3.4	3.5	3.5	3.5	3.5	3.5
	B	3.4	3.4	3.4	3.5	3.1	3.5	3.3	3.4	3.5	3.5	3.5	3.5	3.5	3.5	3.5
BIOL	Winter A	3.4	3.4	3.4	3.5	3.3	3.5	3.3	3.4	3.6	3.5	3.6	3.6	3.5	3.6	3.6
	Winter B	3.4	3.4	3.4	3.5	3.2	3.4	3.3	3.4	3.5	3.4	3.5	3.5	3.6	3.6	3.6
	Spring A	3.4	3.4	3.4	3.5	2.9	3.5	3.3	3.3	3.6	3.5	3.6	3.6	3.6	3.6	3.6
	Spring B	3.3	3.3	3.3	3.5	2.9	3.4	3.2	3.3	3.5	3.5	3.5	3.5	3.6	3.6	3.6
PHSC	Winter A	3.5	3.5	3.4	3.6	3.4	3.5	3.4	3.4	3.6	3.6	3.6	3.6	3.6	3.6	3.6
	Winter B	3.4	3.4	3.4	3.5	3.3	3.4	3.3	3.4	3.5	3.5	3.6	3.6	3.5	3.6	3.5
	Spring A	3.4	3.4	3.3	3.5	3.0	3.5	3.3	3.3	3.5	3.4	3.5	3.5	3.5	3.6	3.6
	Spring B	3.3	3.3	3.3	3.5	2.8	3.4	3.2	3.3	3.5	3.4	3.5	3.5	3.6	3.6	3.6

Note: The SEM estimates are included in the table when the *N* count is 15 or greater.

Table 8.8: 2018–2019 Georgia Milestones SEM by Subgroup, Social Studies

Grade or Course	Form	All	Gender		Ethnicity				Accommodation		Special Education			English Learner		
			Female	Male	African American	Asian	Hispanic	White	Non-Accom	Accom	Non-Accom	Accom	All	Non-Accom	Accom	All
5	A	3.4	3.4	3.3	3.5	3.0	3.4	3.2	3.3	3.5	3.3	3.5	3.5	3.5	3.5	3.5
	B	3.3	3.3	3.3	3.5	2.8	3.3	3.2	3.3	3.5	3.2	3.5	3.4	3.5	3.6	3.5
8	A	3.3	3.3	3.3	3.4	3.0	3.4	3.2	3.2	3.6	3.4	3.6	3.6	3.6	3.6	3.6
	B	3.3	3.3	3.3	3.4	2.8	3.4	3.1	3.3	3.5	3.4	3.5	3.5	3.6	3.6	3.6
HIST	Winter A	3.4	3.4	3.3	3.5	3.2	3.4	3.3	3.3	3.5	3.5	3.5	3.5	3.5	3.6	3.5
	Winter B	3.3	3.4	3.3	3.4	3.1	3.4	3.2	3.3	3.5	3.4	3.5	3.5	3.5	3.6	3.6
	Spring A	3.3	3.3	3.3	3.5	3.0	3.4	3.1	3.2	3.5	3.5	3.5	3.5	3.6	3.6	3.6
	Spring B	3.3	3.3	3.2	3.4	3.0	3.3	3.1	3.2	3.5	3.4	3.5	3.5	3.6	3.6	3.6
ECON	Winter A	3.4	3.4	3.4	3.5	3.1	3.4	3.3	3.3	3.5	3.5	3.5	3.5	3.5	3.6	3.6
	Winter B	3.3	3.3	3.3	3.4	3.0	3.4	3.2	3.3	3.5	3.5	3.5	3.5	3.5	3.5	3.5
	Spring A	3.3	3.3	3.3	3.4	3.0	3.3	3.1	3.2	3.5	3.4	3.5	3.5	3.5	3.5	3.5
	Spring B	3.3	3.3	3.3	3.4	3.0	3.4	3.1	3.3	3.5	3.5	3.5	3.5	3.5	3.5	3.5

Note: The SEM estimates are included in the table when the *N* count is 15 or greater.

8.1.2 Conditional Standard Error of Measurement and Test Information

Within the context of IRT, the precision with which ability can be estimated is called the standard error of the estimate and is analogous to the SEM described in the preceding section. The standard error of the estimate expresses the degree of measurement error in scale score units and is conditional on the ability of the student. The standard error of the estimate, referred to as the CSEM within WINSTEPS, is defined as the reciprocal of the square root of the test information function and can be estimated across all points of the ability continuum (Hambleton & Swaminathan, 1985):

$$\theta\text{_CSEM} = \text{CSEM}(\theta_i) = \frac{1}{\sqrt{I(\theta_i)}}, \tag{8.3}$$

where θ refers to ability estimates associated with each raw score and θ_CSEM refers to the corresponding conditional standard of error produced by WINSTEPS. $I(\theta_i)$ is the test information function as a sum of the item information function obtained as

$$I(\theta_i) = \sum_j \frac{p'_{ij}(\theta_i)^2}{p_{ij}(\theta_i)\, q_{ij}(\theta_i)} + \sum_k [\sum_{j=0}^{mk} j^2 P_{nk}(X = j) - (\sum_{j=0}^{mk} j P_{nk}(X = j))^2], \tag{8.4}$$

where in the first term, $p'_{ij}(\theta_i)$ is the first derivative of $p_{ij}(\theta_i)$ and $q_{ij}(\theta_i) = 1 - p_{ij}(\theta_i)$ for multiple-choice items, and the second term is the information for constructed-response items.

Then θ_CSEM is expressed on the Georgia Milestones reporting scale by applying the linear transformation as expressed in Equation 6.9 (for further details, see Section 6.3.1):

$$\text{SS_CSEM} = \theta\text{_CSEM} \times \text{A} \tag{8.5}$$

Note that the CSEMs vary in magnitude across the entire range of student ability estimates (i.e., scale scores) and are smaller in the middle of the score distribution and higher at the tail ends. This pattern is seen for all Georgia Milestones CSEMs and is to be expected when IRT methods are used. The CSEMs at the three cut scores that define the Georgia Milestones achievement levels are presented in Tables 8.9 through 8.12 and range from 10 to 27 scale score points. Please note that there are slight variations in CSEMs at the cut scores for different test forms. The CSEMs at all scale score points for Georgia Milestones are reported in the scoring tables, as shown in Appendix D, Tables D.2.1 through D.4.16, for the EOG assessments; Appendix E, Tables E.2.1 through E.4.10, for the winter 2018 EOC assessments; and Appendix E, Tables E.5.1 through E.7.10, for the spring 2019 EOC assessments. The CSEMs reported for the 2019 Georgia Milestones suggest that the scores reported to students in 2019 are well estimated and provide an accurate picture of student performance.

Table 8.9: 2018–2019 Georgia Milestones CSEM at Cut Scores for English Language Arts Forms

Grade or Course	Form	Cut Scores			CSEM		
		Developing Learner	***Proficient Learner***	***Distinguished Learner***	***Developing Learner***	***Proficient Learner***	***Distinguished Learner***
3	A	475	525	581	18	19	21
	B	475	525	581	18	19	21
4	A	475	525	574	15	16	19
	B	475	525	574	15	17	20
5	A	475	525	587	15	16	21
	B	475	525	587	15	16	20
6	A	475	525	599	18	19	25
	B	475	525	599	18	19	25
7	A	475	525	592	16	17	23
	B	475	525	592	16	18	23
8	A	475	525	581	14	16	20
	B	475	525	581	14	15	20
9LCO	Winter A	475	525	587	14	15	21
	Winter B	475	525	587	14	15	20
	Spring A	475	525	587	14	15	20
	Spring B	475	525	587	14	15	21
AMLC	Winter A	475	525	590	15	17	24
	Winter B	475	525	590	15	17	24
	Spring A	475	525	590	16	17	26
	Spring B	475	525	590	15	17	27

Table 8.10: 2018–2019 Georgia Milestones CSEM at Cut Scores for Mathematics Forms

Grade or Course	Form	Cut Scores			CSEM		
		Developing Learner	***Proficient Learner***	***Distinguished Learner***	***Developing Learner***	***Proficient Learner***	***Distinguished Learner***
3	A	475	525	580	12	11	16
	B	475	525	580	12	11	16
4	A	475	525	585	13	12	16
	B	475	525	585	13	12	15
5	A	475	525	580	12	13	17
	B	475	525	580	13	13	16
6	A	475	525	580	11	12	17
	B	475	525	580	11	12	16
7	A	475	525	580	13	13	16
	B	475	525	580	14	13	16
8	A	475	525	579	13	13	16
	B	475	525	579	13	13	16
CALG	Winter A	475	525	594	16	15	18
	Winter B	475	525	594	16	15	19
	Spring A	475	525	594	15	15	19
	Spring B	475	525	594	15	15	19
AGEO	Winter A	475	525	596	17	16	20
	Winter B	475	525	596	17	17	20
	Spring A	475	525	596	17	16	20
	Spring B	475	525	596	17	17	19
ALG1	Winter A	475	525	594	16	15	20
	Winter B	475	525	594	16	16	19
	Spring A	475	525	594	16	16	20
	Spring B	475	525	594	16	16	19
GEOM	Winter A	475	525	596	17	17	20
	Winter B	475	525	596	17	17	20
	Spring A	475	525	596	17	17	20
	Spring B	475	525	596	17	16	21

Table 8.11: 2018–2019 Georgia Milestones CSEM at Cut Scores for Science Forms

Grade or Course	Form	Cut Scores			CSEM		
		Developing Learner	*Proficient Learner*	*Distinguished Learner*	*Developing Learner*	*Proficient Learner*	*Distinguished Learner*
5	A	475	525	595	16	17	23
	B	475	525	595	16	17	23
8	A	475	525	593	16	18	23
	B	475	525	593	16	17	23
BIOL	Winter A	475	525	609	18	19	25
	Winter B	475	525	609	18	18	25
	Spring A	475	525	609	18	18	26
	Spring B	475	525	609	18	18	24
PHSC	Winter A	475	525	604	17	18	25
	Winter B	475	525	604	17	18	25
	Spring A	475	525	604	17	18	23
	Spring B	475	525	604	17	18	24

Table 8.12: 2018–2019 Georgia Milestones CSEM at Cut Scores for Social Studies Forms

Grade or Course	Form	Cut Scores			CSEM		
		Developing Learner	*Proficient Learner*	*Distinguished Learner*	*Developing Learner*	*Proficient Learner*	*Distinguished Learner*
5	A	475	525	555	10	11	14
	B	475	525	555	10	11	14
8	A	475	525	572	13	14	18
	B	475	525	572	12	14	18
HIST	Winter A	475	525	590	14	15	20
	Winter B	475	525	590	14	15	20
	Spring A	475	525	590	14	15	20
	Spring B	475	525	590	14	15	20
ECON	Winter A	475	525	610	18	18	24
	Winter B	475	525	610	18	18	24
	Spring A	475	525	610	18	19	24
	Spring B	475	525	610	18	18	25

8.1.3 Classification Consistency and Accuracy

Classification consistency is defined as the extent of agreement between two classifications of a single student from two independent administrations of the same test (or two parallel forms of the test). However, it is difficult to obtain data from repeated administrations of the same form because of the associated costs and time, and two parallel forms are often not available. For these reasons, the common practice is to estimate classification consistency from a single administration. In conjunction with internal consistency, classification consistency is an important type of reliability. As a form of reliability, classification consistency represents how reliably students can be classified into performance categories.

A straightforward approach to classification consistency estimation can be expressed in terms of a contingency table representing the probability of a particular classification outcome under specific scenarios. For example, Table 8.13 is a contingency table of (H+1) × (H+1), where H is the number of cut scores, so three cut scores yield a 4×4 contingency table.

Table 8.13: Example of Contingency Table with Three Cut Scores

	Level 1	Level 2	Level 3	Level 4	Sum
Level 1	P_{11}	P_{21}	P_{31}	P_{41}	$P_{.1}$
Level 2	P_{12}	P_{22}	P_{32}	P_{42}	$P_{.2}$
Level 3	P_{13}	P_{23}	P_{33}	P_{43}	$P_{.3}$
Level 4	P_{14}	P_{24}	P_{34}	P_{44}	$P_{.4}$
Sum	$P_{1.}$	$P_{2.}$	$P_{3.}$	$P_{4.}$	1.0

To report classification consistency, Swaminathan, Hambleton, and Algina (1974) suggested using Cohen's kappa (1960). Cohen's kappa is a statistic used to measure the agreement for categorical classifications that accounts for the possibility of agreement occurring by chance. Cohen's kappa is expressed as

$$\text{kappa} = \frac{P - P_c}{1 - P_c}, \tag{8.6}$$

where P is defined as the sum of diagonal values of the contingency table (shaded above) and P_c is the chance probability of a consistent classification under two completely random assignments. This probability, P_c, is the sum of the probabilities obtained by multiplying the marginal probability of the first administration and the corresponding marginal probability of the second administration:

$$P_c = (P_{1.} \times P_{.1}) + (P_{2.} \times P_{.2}) + (P_{3.} \times P_{.3}) + (P_{4.} \times P_{.4}) \tag{8.7}$$

Classification accuracy is defined as the agreement between the actual classifications using observed cut scores and true classifications based on known true cut scores (Livingston & Lewis, 1995). In other words, classification *consistency* refers to the agreement between two observed scores, while classification *accuracy* refers to the agreement between the observed score and the true score. It is common to estimate classification accuracy by utilizing a psychometric model to find true scores corresponding to observed scores.

Kolen and Kim (2005) suggested a method for estimating both consistency and accuracy that involves the generation of item responses using item parameters based on the IRT model (see also Kim, Choi, Um, & Kim, 2006, and Kim, Barton, & Kim, 2007). Two sets of item responses are generated using a set of item parameters and an examinee's ability distribution from a single test administration. These two sets of item responses are used in place of an examinee's responses on two administrations of the same form. The procedure is described below and is implemented with KKCLASS software (Kim, 2005).

- Step 1: Obtain item parameters (**I**) and the scale score distribution weight $\left(\hat{g}(\theta)\right)$ at each score point from a single test. The number of quadrature points used for these analyses is equal to the number of possible score points on the assessment.
- Step 2: Compute two raw scores at each quadrature point. At a given quadrature point, θ_i, generate two sets of item responses using the item parameters from a test form, assuming that the same test form was administered twice to an examinee with the true ability θ_i.
- Step 3: Construct a classification matrix at each quadrature point. Determine the joint event for the cells in Table 8.13 using the raw scores obtained from Step 2.
- Step 4: Repeat the previous two steps 250 times and get average values over the replications.
- Step 5: Multiply the distribution weight $\left(\hat{g}(\theta)\right)$ by the average values in Step 4 for each quadrature point, and sum across all quadrature points. From this final contingency table, agreement rates and kappa statistics are computed.
- Step 6: Because examinees' abilities are estimated at each quadrature point, this quadrature point can be considered the true score. Therefore, classification accuracy is computed using both examinees' estimated abilities (i.e., observed score) and quadrature point (i.e., true score).

Tables 8.14 through 8.17 show the combined results across all administrations for the 2019 Georgia Milestones classification analyses. Classification consistency (i.e., kappa), and classification accuracy results are presented across all test forms administered in winter and spring. As can be seen, classification consistency ranged from 0.57 to 0.69 and classification accuracy ranged from 0.77 to 0.84. Kappa statistics between 0.61 and 0.80 are considered substantially strong rates of agreement and greater than 0.80 are considered extremely high (Landis & Koch, 1977). The magnitude of classification consistency and accuracy measures is influenced by key features of the test design, including the number of items, the number of cut scores, and the reliability and associated SEMs. The classification consistency and accuracy results established for 2019 suggest that consistent and accurate achievement level classifications are being made for students in Georgia based on the Georgia Milestones assessments.

Table 8.14: 2018–2019 Georgia Milestones Classification Consistency and Accuracy, English Language Arts Forms

Grade or Course	Form	Number of Items	Kappa	Classification Accuracy
3	A	44	0.60	0.79
	B	44	0.58	0.77
4	A	44	0.59	0.79
	B	44	0.57	0.77
5	A	44	0.62	0.81
	B	44	0.58	0.78
6	A	44	0.61	0.80
	B	44	0.60	0.79
7	A	44	0.60	0.80
	B	44	0.58	0.79
8	A	44	0.61	0.80
	B	44	0.58	0.79
9LCO	Winter A	44	0.61	0.80
	Winter B	44	0.59	0.80
	Spring A	44	0.61	0.81
	Spring B	44	0.60	0.80
AMLC	Winter A	44	0.58	0.78
	Winter B	44	0.57	0.78
	Spring A	44	0.59	0.79
	Spring B	44	0.57	0.78

Table 8.15: 2018–2019 Georgia Milestones Classification Consistency and Accuracy, Mathematics Forms

Grade or Course	Form	Number of Items	Kappa	Classification Accuracy
3	A	52	0.69	0.84
	B	52	0.65	0.83
4	A	52	0.67	0.83
	B	52	0.64	0.82
5	A	52	0.67	0.83
	B	52	0.66	0.82
6	A	52	0.68	0.84
	B	52	0.66	0.83
7	A	52	0.66	0.82
	B	52	0.63	0.81
8	A	52	0.66	0.83
	B	52	0.62	0.81
CALG	Winter A	52	0.67	0.83
	Winter B	52	0.65	0.82
	Spring A	52	0.64	0.82
	Spring B	52	0.60	0.80
AGEO	Winter A	52	0.64	0.81
	Winter B	52	0.67	0.82
	Spring A	52	0.62	0.81
	Spring B	52	0.60	0.80
ALG1	Winter A	52	0.62	0.81
	Winter B	52	0.61	0.80
	Spring A	52	0.63	0.81
	Spring B	52	0.62	0.81
GEOM	Winter A	52	0.66	0.82
	Winter B	52	0.65	0.82
	Spring A	52	0.63	0.81
	Spring B	52	0.62	0.80

Table 8.16: 2018–2019 Georgia Milestones Classification Consistency and Accuracy, Science Forms

Grade or Course	Form	Number of Items	Kappa	Classification Accuracy
5	A	56	0.65	0.82
	B	56	0.62	0.80
8	A	56	0.63	0.82
	B	56	0.63	0.81
BIOL	Winter A	56	0.63	0.81
	Winter B	56	0.63	0.81
	Spring A	56	0.65	0.82
	Spring B	56	0.64	0.81
PHSC	Winter A	56	0.61	0.81
	Winter B	56	0.63	0.80
	Spring A	56	0.64	0.81
	Spring B	56	0.61	0.80

Table 8.17: 2018–2019 Georgia Milestones Classification Consistency and Accuracy, Social Studies Forms

Grade or Course	Form	Number of Items	Kappa	Classification Accuracy
5	A	56	0.64	0.82
	B	56	0.59	0.80
8	A	56	0.63	0.80
	B	56	0.63	0.81
HIST	Winter A	56	0.63	0.81
	Winter B	56	0.63	0.81
	Spring A	56	0.64	0.81
	Spring B	56	0.64	0.82
ECON	Winter A	56	0.60	0.80
	Winter B	56	0.61	0.80
	Spring A	56	0.63	0.81
	Spring B	56	0.61	0.80

8.2 Validity

The *Standards for Educational and Psychological Testing* defines validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests" (AERA, APA, & NCME, 2014, p. 11). The purpose of test score validation is not to validate the test itself but to validate interpretations of the test scores for particular purposes or uses. Test score validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the entire assessment process. Every aspect of an assessment provides evidence that either supports or challenges its validity, including

design, content specifications, item development, psychometric quality, and inferences made from the results.

The validity of score interpretations for Georgia Milestones is supported by multiple sources of evidence. Chapter 1 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) specifies the following sources of validity evidence that are important to gather and document to support validity claims for an assessment:

- Test content
- Response processes
- Internal test structure
- Relation to other variables
- Consequences of test use

It is important to note that these categories are not mutually exclusive. One source of validity evidence often falls into more than one category, as discussed in more detail in this chapter. The process of gathering evidence of the validity of score interpretations is best characterized as ongoing throughout test development, administration, scoring, reporting, and beyond. As this technical report has progressed, it has covered the different phases of the testing cycle. Each part of this technical report has detailed the procedures and processes applied in Georgia Milestones and the results. Each part has also highlighted the meaning and significance of the procedures, processes, and results in terms of validity and their relationship to specific sections of the *Standards.* The current section now addresses these final aspects of validity: test content, response processes, internal test structure, relation to other variables, and consequences of test use.

8.2.1 Evidence Based on Test Content

According to the *Standards*, "evidence based on test content can include logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretation of test scores" (AERA, APA, & NCME, 2014, p. 14). Hence, documentation of the content domains, how the content is sampled and represented, and alignment of items to the content were articulated in Chapter 2. It showed how test specification documents, derived from earlier developmental activities, guided the final phases of test development and ultimately yielded the test forms that were administered to students.

Chapter 2 also showed that the participation of Georgia educators in that process provided a solid rationale for the credibility of the content and design of Georgia Milestones as a tool from which to derive valid inferences about Georgia student performance. The use of classroom teachers also brings into the process the enacted curriculum perspective and the written curriculum perspective. The test development process and the involvement of Georgia educators in that process formed an important part of the validity of the entire Georgia Milestones program. Through their knowledge, expertise, and professional judgment, Georgia educators ultimately ensured that the content of Georgia Milestones formed an adequate and representative sample of appropriate content, and they ensured that the content formed a legitimate basis upon which to derive conclusions about student performance.

8.2.2 Evidence Based on Response Processes

According to the *Standards*, evidence based on response processes "generally comes from analyses of individual responses" (AERA, APA, & NCME, 2014, p. 15). Hence, the best opportunity to detect and eliminate potential sources of invalidity occurs during the test development process (U.S. Department of Education, 2015). As indicated in Chapter 2, all items for Georgia Milestones were carefully reviewed through multiple cycles of the item development process for ambiguity, bias, sensitivity, irrelevance, and inaccuracy to ensure a fit between the construct and the nature of the actual performance.

8.2.3 Evidence Based on Internal Test Structure

According to the *Standards*, evidence based on internal structure reflects "the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA, APA, & NCME, 2014, p. 16). Three important sources of internal structure evidence have been addressed within this technical report: measurement invariance, dimensionality, and reliability. Evidence of measurement invariance is provided in Section 8.2.3.1 on differential item functioning (DIF) and in Section 8.1.1, which reports the subgroup reliability and standard errors of measurement estimates. Dimensionality is addressed in Sections 6.2.2.3, 6.2.2.4, and 8.2.3.2. Section 8.1.1 also provides supporting evidence of test reliability, indicating that the reported test scores are consistent across repeated administrations.

8.2.3.1 Differential Item Functioning

The topic of test bias is addressed through an analysis of DIF. It is possible for items to function differently among different subpopulation groups. It is also possible that results for an item do not reflect student ability but instead reflect irrelevant information influenced by demographic factors. The DIF analyses provided in this section serve to determine whether those biases occurred and to what degree, item by item, for each gender and ethnicity group. The *Standards* considers the lack of DIF as evidence based on internal structure.

The 2019 Georgia Milestones assessments were developed using procedures to minimize item and test bias. Expertise in this area is not, however, a substitute for statistical analyses of the items. Thus, an empirical DIF approach was used to examine potential item bias. The approach applied included systematic item analyses to determine whether examinees with the same underlying level of ability had the same probability of correctly responding to the item. Items identified with DIF were examined to determine whether item performance differences between identifiable subgroups of the population were due to extraneous or construct-irrelevant information, making the items unfairly difficult for one of the subgroups.

DIF analyses were conducted for two grouping factors: gender (i.e., male and female) and ethnicity (i.e., White, African American, and Hispanic). Two DIF statistics, the Mantel-Haenszel (1959) and the delta statistic (Holland & Thayer, 1988) were used to evaluate dichotomous items. Similarly, the Mantel-Haenszel and the standardized mean difference (SMD) statistics were used to evaluate DIF for multipoint items.

(1) Mantel-Haenszel

The Mantel-Haenszel statistic is computed as

$$\text{Mantel } \chi^2 = \frac{\left(\sum_k F_k - \sum_k E(F_k)\right)^2}{\sum_k \text{Var}(F_k)}, \tag{8.8}$$

where F_k is the sum of scores for the focal group at the kth level of the matching variable (Mantel & Haenszel, 1959; Zwick, Donoghue, & Grima, 1993). Note that the Mantel-Haenszel (1959) statistic is sensitive to N counts such that larger sample sizes increase the value of chi-square statistics.

In addition to the Mantel-Haenszel (1959) chi-square statistic, the delta statistic (MH-D DIF) was computed for all dichotomous items (Holland & Thayer, 1988). To compute delta, alpha (i.e., the odds ratio) is first computed as

$$\alpha_{MH} = \frac{\sum_{k=1}^{K} N_{r1k} N_{f0k} / N_k}{\sum_{k=1}^{K} N_{f1k} N_{r0k} / N_k}, \tag{8.9}$$

where N_{r1k} is the number of correct responses in the reference group at ability level k, N_{f0k} is the number of incorrect responses in the focal group at ability level k, N_k is the total number of responses, N_{f1k} is the number of correct responses in the focal group at ability level k, and N_{r0k} is the number of incorrect responses in the reference group at ability level k.

MH-D DIF is then computed as

$$\text{MH-D DIF} = -2.35 \ln \alpha_{MH}. \tag{8.10}$$

A negative MH-D DIF value indicates an item on which the focal group has a lower mean than the reference group. A positive MH-D DIF value indicates an item on which the reference group has a lower mean than the focal group. Thus, positive values of MH-D DIF indicate items that favor the focal group (i.e., African American, Hispanic, female, or accommodated), whereas negative values of MH-D DIF indicate items that favor the reference group (i.e., White, male, or non-accommodated).

To identify whether Georgia Milestones items exhibit strong, weak, or no DIF (referred to as C, B, and A, respectively), items were flagged based on the combination of the Mantel-Haenszel (1959) chi-square statistic and the delta statistic. The definitions of the DIF categories for multiple-choice items are shown in Table 8.18.

Table 8.18: DIF Categories for Multiple-Choice Items

Category	Description	Criterion
A	No DIF	Mantel-Haenszel chi-square statistic not significantly different from zero (p>0.05)
B	Weak DIF	Significant Mantel-Haenszel chi-square statistic (p<0.05) and |MH D-DIF|<1.5
C	Strong DIF	Significant Mantel-Haenszel chi-square statistic (p<0.05) and |MH D-DIF|>1.5

(2) Standardized Mean Difference

The SMD is an effect size index of DIF and is relatively easy to interpret (Dorans & Holland, 1992; Zwick et al., 1993). The SMD compares the means of the reference and focal groups, adjusting for the distribution of reference and focal group members on the conditioning variable (Zwick et al., 1993), which for these analyses is the Georgia Milestones raw score. SMD is computed as

$$\mathrm{SMD} = p_{Fk}\left(\sum_{k} m_{Fk} - \sum_{k} m_{Rk}\right), \qquad (8.11)$$

where p_{Fk} is the proportion of the focal group members at the kth level of the matching variable, m_{Fk} is equal to $1/N_{F1k}$, and m_{Rk} is equal to $1/N_{R1k}$ (Zwick et al., 1993).

A negative SMD value indicates an item on which the focal group has a lower mean than the reference group. A positive SMD value indicates an item on which the reference group has a lower mean than the focal group. Thus, positive values of SMD indicate items that favor the focal group (i.e., African American, Hispanic, female, or accommodated), whereas negative values of SMD indicate items that favor the reference group (i.e., White, male, or non-accommodated). The SMD statistic is utilized to flag DIF for multipoint items, and the corresponding DIF categories are defined in Table 8.19.

Table 8.19: DIF Categories for Constructed-Response Items

Category	Description	Criterion
A	No DIF	Mantel-Haenszel chi-square statistic not significantly different from zero (p≥0.05)
B	Weak DIF	Significant Mantel-Haenszel chi-square statistic (p<0.05) and 0.17≤|SMD|<0.25
C	Strong DIF	Significant Mantel-Haenszel chi-square statistic (p<0.05) and |SMD|≥0.25

(3) Results

Tables 8.20 and 8.21 provide the number of operational items flagged for “Strong DIF” for each grade/content area and course. Note that a single item can be flagged for multiple groupings, such as for African American students and for Hispanic students. For the DIF analysis by gender, the reference group is male, meaning that the results for female students are considered with reference to male student performance. For ethnicity, the reference group is White. This means that the performance of other ethnic groups is considered with reference to the performance of White students.

The summary flag information in the DIF tables is always expressed with reference to the focal group. That means that a negative flag (such as B- or C-) indicates that an item favors the reference group, such as male or White students. A positive flag (such as B+ or C+) indicates that the item favors the focal group, such as female, African American, or Hispanic students.

The DIF analysis results obtained for field test items were used in data review (see Section 2.2.7 for details). The selection of operational items took place after the data review, and the item pool reflects those items that have passed the final stage of the education committee review.

Table 8.22 shows that 8 out of 398 items were flagged for "Strong DIF" in tests across all EOG ELA forms. Out of 474 items, 3 were flagged in the EOG mathematics forms, and out of 172 items, 1 was flagged in the EOG science forms. Out of 172 items, 2 were flagged in the EOG social studies forms. Table 8.23 shows that 4 out of 265 items were flagged in EOC ELA courses across all test administrations. Out of 632 items, 11 were flagged in forms in the EOC mathematics courses across all test administrations. Out of 340 items, 1 was flagged in EOC science courses across all administrations. Out of 347 items, 4 were flagged in EOC social studies courses. The flagging rate within the EOG and EOC assessments was well below the nominal rate, and overall, the flagging is bidirectional, with items favoring the focal and reference groups.

Table 8.20: 2019 Georgia Milestones Operational EOG Number of Strong (C) DIF Flags

Content	Grade	Number of Operational Items	Number of Items with DIF Information Available	Number of African American C DIF Flags	Number of Hispanic C DIF Flags	Number of Female C DIF Flags	Number of Accommodation C DIF Flags
English Language Arts	3	66	66	0	2	2	1
	4	65	65	0	0	0	0
	5	66	66	1	0	0	0
	6	67	67	0	0	0	0
	7	67	67	0	0	1	0
	8	67	67	1	0	0	0
Mathematics	3	79	79	0	0	0	0
	4	79	79	0	0	0	0
	5	79	79	0	0	0	0
	6	79	79	0	0	2	0
	7	79	79	0	0	0	0
	8	79	79	0	0	1	0
Science	5	86	86	0	0	0	1
	8	86	86	0	0	0	0
Social Studies	5	86	86	0	0	0	0
	8	86	86	1	1	0	0

Table 8.21: 2018–2019 Georgia Milestones Operational EOC Number of Strong (C) DIF Flags

Course	Administration	Number of Operational Items	Number of Items with DIF Information Available	Number of African American C DIF Flags	Number of Hispanic C DIF Flags	Number of Female C DIF Flags	Number of Accommodation C DIF Flags
9LCO	Winter	64	64	0	1	1	0
	Spring	67	67	0	1	1	0
AMLC	Winter	67	67	0	0	0	0
	Spring	67	67	0	0	0	0
CALG	Winter	79	79	0	4	1	2
	Spring	79	79	0	0	0	0
AGEO	Winter	79	79	1	0	0	0
	Spring	79	79	0	0	0	0
ALG1	Winter	79	79	0	0	0	0
	Spring	79	79	1	0	2	0
GEOM	Winter	79	79	0	0	1	0
	Spring	79	79	0	0	1	0
BIOL	Winter	85	85	0	0	0	0
	Spring	86	86	0	0	0	0
PHSC	Winter	85	85	0	0	1	0
	Spring	84	84	0	0	0	0
HIST	Winter	87	87	1	1	0	0
	Spring	87	87	1	0	1	0
ECON	Winter	87	87	0	0	0	0
	Spring	86	86	0	0	0	0

Table 8.22: 2019 Georgia Milestones EOG List of Operational DIF Flags

Content	Grade	Item	Focal Group	Reference Group	MH	Delta	SMD	DIF Flag
English Language Arts	3	Flagged item 1	Female	Male	0.00	-1.81	-0.13	C-
		Flagged item 2	Female	Male	0.00	-1.80	-0.11	C-
		Flagged item 3	Hispanic	White	0.00	-1.96	-0.09	C-
		Flagged item 4	Hispanic	White	0.00	-3.43	-0.19	C-
		Flagged item 4	Accommodated	Non-Accommodated	0.00	-1.94	-0.14	C-
	5	Flagged item 5	African American	White	0.00	1.50	0.10	C+
	7	Flagged item 6	Female	Male	0.00	-1.93	-0.15	C-
	8	Flagged item 7	African American	White	0.00	-1.70	-0.12	C-
Mathematics	6	Flagged item 8	Female	Male	0.00	-2.41	-0.14	C-
		Flagged item 9	Female	Male	0.00	-1.95	-0.11	C-
	8	Flagged item 10	Female	Male	0.00	-1.60	-0.13	C-
Science	5	Flagged item 11	Accommodated	Non-Accommodated	0.00	1.66	0.05	C+
Social Studies	8	Flagged item 12	African American	White	0.00	-1.72	-0.14	C-
		Flagged item 12	Hispanic	White	0.00	-1.73	-0.14	C-

Table 8.23: 2018–2019 Georgia Milestones EOC List of Operational DIF Flags

Course	Administration	Item	Focal Group	Reference Group	MH	Delta	SMD	DIF Flag
9LCO	Winter	Flagged item 1	Female	Male	0.00	-1.59	-0.08	C-
	Winter	Flagged item 1	Hispanic	White	0.00	1.78	0.09	C+
	Spring	Flagged item 1	Female	Male	0.00	-1.59	-0.08	C-
	Spring	Flagged item 1	Hispanic	White	0.00	1.78	0.09	C+
CALG	Winter	Flagged item 2	Female	Male	0.00	-1.80	-0.13	C-
	Winter	Flagged item 3	Hispanic	White	0.01	-3.05	-0.11	C-
	Winter	Flagged item 4	Hispanic	White	0.00	1.67	0.16	C+
	Winter	Flagged item 5	Hispanic	White	0.00	1.73	0.16	C+
	Winter	Flagged item 6	Hispanic	White	0.01	-1.55	-0.15	C-
AGEO	Winter	Flagged item 7	African American	White	0.00	1.65	0.12	C+
ALG1	Spring	Flagged item 8	Female	Male	0.00	-1.74	-0.08	C-
	Spring	Flagged item 9	Female	Male	0.00	-1.94	-0.10	C-
	Spring	Flagged item 10	African American	White	0.00	-2.04	-0.11	C-
GEOM	Winter	Flagged item 11	Female	Male	0.00	-1.52	-0.10	C-
	Spring	Flagged item 11	Female	Male	0.00	-1.52	-0.10	C-
PHSC	Winter	Flagged item 12	Female	Male	0.00	-2.01	-0.17	C-
HIST	Winter	Flagged item 13	African American	White	0.00	-1.64	-0.12	C-
	Winter	Flagged item 14	Hispanic	White	0.00	1.68	0.10	C+
	Spring	Flagged item 15	Female	Male	0.00	-1.71	-0.13	C-
	Spring	Flagged item 16	African American	White	0.00	1.67	0.07	C+

8.2.3.2 Dimensionality Analysis

Achievement tests are typically designed to measure student proficiency on a single continuum (or unidimensional construct). For a test to be scalable and adequately analyzed using a unidimensional IRT model, such as what was used for Georgia Milestones, the test should be essentially unidimensional. The *Standards* considers unidimensionality as evidence based on internal structure.

The WINSTEPS program includes a principal component analysis of the residual variation that is used to assess the unidimensionality assumption. That is, it is expected that the first dimension associated with the Rasch model will account for the majority of score variation. If the unidimensionality assumption is met, an analysis of the residual variation will indicate that variation associated with secondary factors is negligible.

Table 8.24 shows the results of the principal component analyses conducted as part of the concurrent calibration of all items for each grade/content area and course of the Georgia Milestones assessments. Across all content areas, the percentages of variation accounted for by a first dimension ranged from 23.8% to 43.6%, while the percentages of variation accounted for by a second dimension ranged from 0.9% to 2.1%. Such small variation accounted for by the second factor was considered minimal and indicated that the unidimensionality assumption holds for each Georgia Milestones assessment. That is, there is confidence that using the IRT model for scaling, equating, and score reporting was appropriate for the Georgia Milestones program because the key assumption of unidimensionality was tenable.

Table 8.24: 2018–2019 Principal Component Analyses of Georgia Milestones Items—Percentage of Variation Accounted for by First and Second Dimensions

Content Area	Grade or Course	Number of Items	Percentage First Dimension	Percentage Second Dimension
English Language Arts	3	66	36.3	2.1
	4	65	36.6	1.9
	5	66	34.6	2.0
	6	67	35.2	1.9
	7	67	32.8	2.0
	8	67	36.1	1.9
	9LCO	102	37.4	1.4
	AMLC	112	32.7	1.2
Mathematics	3	79	40.9	1.7
	4	79	43.6	1.2
	5	79	38.8	1.6
	6	79	37.4	1.4
	7	79	35.7	1.5
	8	79	33.4	1.5
	CALG	131	34.1	1.0
	AGEO	132	32.1	1.0
	ALG1	131	37.6	0.9
	GEOM	132	38.8	0.9
Science	5	86	27.8	1.5
	8	86	25.3	1.6
	BIOL	141	23.8	1.0
	PHSC	143	26.1	0.9
Social Studies	5	86	26.4	1.4
	8	86	27.1	1.5
	HIST	143	26.7	1.0
	ECON	148	26.6	1.0

Note: The winter and spring EOC data have been combined.

8.2.4 Evidence Based on Relations to Other Variables

The *Standards* indicates that analyses of test scores in relationship to variables external to the test address questions about the degree to which the relationships with other variables are consistent with the construct underlying the proposed test interpretation.

There are various ways to express evidence of the relation of test scores to a relevant criterion. Two designs, often called *predictive* and *concurrent*, have been distinguished for evaluating test-criterion relationships. A predictive study indicates how accurately test data can predict criterion scores that are obtained at a later time, whereas a concurrent study obtains predictor and criterion information at the same time.

The Georgia Milestones program includes the administration of both Georgia Milestones and TerraNova items. While some TerraNova items serve as dual-purpose items, in that they also count toward the Georgia Milestones test scores, the assessments were developed to serve different purposes and have distinct test blueprints and psychometric specifications, and therefore, the correlation of student performance on both assessments provides concurrent validation evidence. The correlations between student scores on Georgia Milestones and TerraNova items administered during the 2018–2019 school year are presented in Tables 8.25 and 8.26. The correlations between the assessments are relatively strong, ranging from 0.66 to 0.83 and averaging 0.77. Given the overlap between the content standards covered by both assessments and the fact that some items count toward total scores on both assessments, the strong relationship between Georgia Milestones and TerraNova is to be expected.

It is important to note that the 2018–2019 school year served as the fifth year of the Georgia Milestones testing program. Previous work predicting Georgia Milestones proficiency rates based on previous scores suggests that the Georgia Milestones assessments are providing coherent information. This work is presented in prior technical reports and was not repeated in 2018–2019.

Table 8.25: 2019 Correlation between Georgia Milestones EOG Scale Scores and TerraNova Scale Scores

Content	Grade	*N*	Correlation
English Language Arts	3	129,231	0.77
	4	133,547	0.78
	5	136,513	0.79
	6	136,673	0.80
	7	133,259	0.77
	8	124,745	0.82
Mathematics	3	129,156	0.81
	4	133,486	0.79
	5	136,458	0.79
	6	136,626	0.81
	7	132,796	0.80
	8	103,388	0.74
Science	5	136,269	0.80
	8	94,788	0.74
Social Studies	5	136,207	0.75
	8	130,182	0.76

Table 8.26: 2018–2019 Correlation between Georgia Milestones EOC Scale Scores and TerraNova Scale Scores

Course	Administration	*N*	Correlation
9LCO	Winter 2018	23,448	0.81
	Spring 2019	110,892	0.82
	Summer 2019	24,260	0.80
AMLC	Winter 2018	95,511	0.80
	Spring 2019	2,362	0.76
	Summer 2019	17,669	0.75
CALG	Winter 2018	3,368	0.80
	Spring 2019	15,237	0.74
	Summer 2019	15,169	0.73
AGEO	Winter 2018	103,477	0.76
	Spring 2019	24,339	0.83
	Summer 2019	82,809	0.81
ALG1	Winter 2018	25,134	0.74
	Spring 2019	103,172	0.77
	Summer 2019	12,906	0.66
GEOM	Winter 2018	73,826	0.72
	Spring 2019	21,863	0.76
	Summer 2019	91,514	0.77
BIOL	Winter 2018	44,992	0.73
	Spring 2019	56,926	0.73
	Summer 2019	23,448	0.81
PHSC	Winter 2018	110,892	0.82
	Spring 2019	24,260	0.80
	Summer 2019	95,511	0.80
HIST	Winter 2018	2,362	0.76
	Spring 2019	17,669	0.75
	Summer 2019	3,368	0.80
ECON	Winter 2018	15,237	0.74
	Spring 2019	15,169	0.73
	Summer 2019	103,477	0.76

8.2.5 Evidence Based on Consequences of Test Use

The *Standards* incorporates the intended and unintended consequences of test use into the concept of validity. It indicates that information about the consequences of testing does not in and of itself detract from the validity of intended test interpretations (AERA, APA, & NCME, 2014, p. 19). Rather, according to the *Standards*, a more searching inquiry into the sources of those consequences given the intended purposes of an assessment is a basis for evaluating the quality of the validity evidence. The test data alone do not provide sufficient verification of this type of evidence. For this reason, it is not straightforward to measure/collect evidence on the consequential aspects of validity.

To address the intended consequences of the Georgia Milestones assessments, the purposes of Georgia Milestones must be specified. The GaDOE has carefully articulated the intended

purposes of Georgia Milestones as a driving feature of the development and implementation of the testing program. The specific purposes associated with the Georgia Milestones testing program are as follows:

- Georgia Milestones measures how well students have acquired the knowledge and skills across the full achievement continuum as described in the Georgia-mandated content standards.
- Georgia Milestones provides a consistent and coherent signal about student preparedness for the next level, be it the next grade, the next course, college, or a career.
- Georgia Milestones informs state and federal accountability, including educator effectiveness, at the school, district, and state levels.
- Georgia Milestones provides a consistent and coherent signal about student achievement both within the system (i.e., across grades and courses) and with external measures (e.g., NAEP, PSAT, SAT, ACT).
- Georgia Milestones is fair for all students, including those with disabilities or limited English proficiency at all levels of achievement.

The GaDOE will be implementing a usability study that will provide evidence regarding the intended and unintended uses of the Georgia Milestones assessments. In particular, a survey will explore and confirm the extent to which important stakeholder groups (i.e., parents, teachers, and district staff) are using the results from the assessments as intended. Surveys are under development for each stakeholder group to gather this information. The surveys will measure both accessibility and actual use of the exam results. The *Standards* indirectly addresses reporting, but specific standards that mention reporting procedures span multiple chapters (i.e., 6.10–6.16, 7.13, 8.5–8.8, and 12.18–12.19). Likert-type questions will largely be used in a confirmatory sense to measure the extent to which stakeholders are using the results as intended. Open-ended questions will be used to measure the ways in which stakeholders are using the results more generally, which will capture unintended uses that will be uncovered through thematic analysis via inductively emerging themes (and content) from the data.

8.2.6 Developing the Validity Argument

The argument-based approach to test validation (Cronbach, 1988; House, 1980; Kane, 1992, 2006, 2013a, 2013b; Shepard, 1993) is designed to focus validation efforts on the specific purposes for which a test is intended. Test validation requires the collection and evaluation of evidence, organized relative to the intended purposes of the test results and the intended interpretations being made about the test results. The process of developing a validity argument involves this systematic collection, organization, and evaluation of evidence.

The GaDOE established guiding principles for the development of Georgia Milestones that articulate the primary purposes and goals for the assessment system. The intended purposes associated with the Georgia Milestones testing program are stated below and are followed by information describing the multiple sources of evidence supporting each.

- Georgia Milestones measures how well students have acquired the knowledge and skills across the full achievement continuum as described in the Georgia-mandated content standards.
- Georgia Milestones provides a consistent and coherent signal about student preparedness for the next level, be it the next grade, the next course, college, or a career.
- Georgia Milestones informs state and federal accountability, including educator effectiveness, at the school, district, and state levels.
- Georgia Milestones provides a consistent and coherent signal about student achievement both within the system (i.e., across grades and courses) and with external measures (e.g., NAEP, PSAT, SAT, ACT).
- Georgia Milestones is fair for all students, including those with disabilities or limited English proficiency, at all levels of achievement.

Georgia Milestones has multiple sources of evidence to support its use in measuring how well students have acquired the knowledge and skills described in the Georgia-mandated content standards. This report has presented detailed examples regarding the item and test development process, including the alignment of the items with the content standards and the correspondence of the test blueprint and the content standards. This evidence is further supported by the item development protocols defined by the item specifications and the involvement of Georgia educators in the approval, revision, and selection of all items used on the Georgia Milestones assessments. Detailed information regarding the item and test development process is contained throughout, and this process included extensive opportunity for educator feedback that was systematically documented. The test assembly protocol was reviewed and documented, and the evidence for the requisite assumptions for the estimations of IRT models was reviewed and documented. The processes by which the achievement levels were developed and the associated cuts were established were recapped in this technical report and described in detail in the *Georgia Milestones Standard Setting Technical Report* (2015) and the *Georgia Milestones Standard Setting Technical Report for End-of-Grade and End-of-Course Science and Social Studies* (2018). Documentation of the extensive item and test analyses is provided throughout this technical report to confirm that the student item responses were correctly scored and total test scores were reliable.

A number of sources provide evidence in support of the claim that Georgia Milestones provides a consistent and clear signal about student preparedness for the next level. Georgia Milestones was designed to be an integrated assessment system that replaces multiple testing programs, and the extensive documentation of its implementation is relevant. For example, the standard setting included specific design elements to facilitate the coherency of the assessment system. Participants were initially organized in grade/course pairs to facilitate articulation of the achievement levels and the associated expectations for students as expressed by the cut scores. Subsequent cross-group activities were designed to facilitate vertical articulation of the achievement levels and the associated cut scores for all grades and courses. The articulation of the expectations for students across grades and courses within a content area and across the entire assessment program was then evaluated by relevant stakeholders in a policy review meeting. The body of evidence in support of this claim is thoroughly documented within the *Georgia Milestones Standard Setting Technical Report* (2015) and the *Georgia Milestones*

Standard Setting Technical Report for End-of-Grade and End-of-Course Science and Social Studies (2018).

Georgia Milestones provides effective achievement information that can be used to support state and federal accountability, including educator effectiveness at the school, district, and state levels. Documentation of the item and test development process is described within this report and supported by an extensive set of standard operating procedures that guided the development of the assessments, including routine internal and external feedback gathering from stakeholders. Psychometric analyses, including classical item analyses, reliability assessment, dimensionality, and model/data fit evaluation provided convergent evidence. Critical feedback about the increasing transition to online administrations for all students at all grades has been routinely gathered, and questions about the score comparability of the assessments across mode have been addressed in detail. A series of mode comparability analyses over four years have indicated negligible differences, empirically supporting the equivalent validity of both modes of assessment. As Georgia Milestones transitions to 100% online administrations, more Georgia students and teachers have determined it to be the appropriate mode of assessment to fit their learning environment.

A number of sources provide evidence in support of the claim that Georgia Milestones provides a consistent and clear signal about student achievement within the system and with external measures. From the outset, Georgia Milestones was designed to be an integrated assessment that replaced multiple testing programs to provide a coherent picture of student achievement. The standard setting included specific design elements to facilitate the internal and external coherency of the assessment system. Participants were initially organized in grade/course pairs to facilitate articulation of the achievement levels and the associated expectations for students as expressed by the cut scores. Subsequent cross-group activities were designed to facilitate vertical articulation of the achievement levels and the associated cut scores for all grades and courses. The standard setting sessions also explicitly incorporated external performance benchmarks to facilitate coherency of the Georgia Milestones testing program with external assessments. Specifically, bookmarks consistent with student performance on the NAEP, the PSAT, or the Lexile assessments were included within the ordered item booklets that are central to the method used to establish the Georgia Milestones standards. The body of evidence in support of this claim is thoroughly documented within the standard setting technical report. A recent study linking Georgia Milestones and ACCUPLACER, a college placement assessment used by high schools and community colleges, suggests that the Georgia Milestones assessments are providing coherent information about student preparedness for technical schools, and thus, the study provides convergent evidence in support of this claim.

Georgia Milestones has multiple sources of evidence to support the claim that the assessments are fair for all students, including those with disabilities or limited English proficiency, at all levels of achievement. This report presents detailed examples regarding the inclusion of universal design elements in the item and test development process. This evidence is further supported by the item development protocols defined by the item specifications and the involvement of Georgia educators' item and data review meetings. Detailed information regarding the item and test development process is contained throughout and reflects extensive opportunity for educator feedback, which was systematically documented; Georgia educators are involved in the approval, revision, and selection of all items used on the Georgia Milestones assessments. Psychometric analyses, including classical item analyses, reliability assessment,

dimensionality, and model/data fit evaluation, provided convergent evidence. DIF statistics are calculated for every item, and all items flagged for differences by group were also flagged and subsequently reviewed for bias. Documentation of the extensive item and test analyses is provided throughout to confirm that the student item responses were correctly scored and total test scores were reliable for all students.

Kane (2013a) articulated the argument-based approach to validity in terms of two arguments. First, the interpretation and use argument is designed to facilitate specification of the assumptions and inferences necessary to use a test score for a given decision or purpose. Second, the validity argument is centered on the degree to which the interpretation and use argument can be supported and effectively provides the guidance for evidence collection. Sireci (2013) argued that an effective way to make the validity argument framework immediately accessible for practitioners is to simply crosscheck the intended purposes for a test with the following sources of validity evidence as stated in the *Standards*:

- Test content
- Response processes
- Internal test structure
- Relation to other variables
- Consequences of test use

Kane (2013b) acknowledged that distinguishing between the two arguments was designed to facilitate explicit statements of the claims inherent in the interpretations and associated uses of test scores and the evaluations of these claims, but articulation of both arguments may not be necessary when the expressed purposes for an assessment are carefully articulated such that they can be evaluated. Given that the intended purposes for the Georgia Milestones assessments encompass the interpretations and uses of the scores, Kane's argument-based approach is embedded throughout the validation work implemented in support of the assessments. The purposes for Georgia Milestones have been carefully articulated, and the sources of evidence as provided by this technical report, which documents all phases of the test development and implementation cycle, are provided below.

As for validation evidence based on test content, Chapter 2 presents item and test development evidence: alignment of the standards and the assessment; item development based on test specifications, which identify the number of items for each content domain (also used for score reporting category); item specifications, which provide item writers with clear models of phrasing, formatting, and graphical presentations for acceptable test items; and the participation of Georgia educators in the approval, revision, and selection of all test items. The careful review of items for any ambiguity, irrelevant clues, and inaccuracy provided evidence based on the response process.

Evidence of validity based on the internal structure of Georgia Milestones was obtained through DIF. The Georgia Milestones program formally assessed the possibility of test bias through an analysis of DIF, in order to verify that items included did not function differently within different population groups. This analysis also addresses whether results for an item reflect student ability instead of irrelevant information influenced by demographic factors. For

each item, the DIF analysis serves to determine whether that possibility occurred and to what degree it could have occurred with respect to ethnicity and gender groupings.

Evidence of validity based on internal structure was also obtained through dimensionality analyses. For a test to be scalable and adequately analyzed using a unidimensional IRT model, as was used for Georgia Milestones, the test should be essentially unidimensional. That is, it is expected that the first dimension associated with the Rasch model will account for the majority of score variation. If the unidimensionality assumption is met, an analysis of the residual variation will indicate that variation associated with secondary factors is negligible. The results of the principal component analysis indicated that the unidimensionality assumption holds for each Georgia Milestones assessment. An additional evaluation of the extent to which the Rasch model calibrations and scale scores can be explained under a confirmatory factor analytic framework also indicated the unidimensionality assumption holds for the assessments. That is, there is confidence that using an IRT model for scaling, equating, and score reporting is appropriate for the Georgia Milestones program because the key assumption of unidimensionality was tenable.

Additional evidence of validity related to the internal test structure is obtained through an ongoing evaluation of item fit, person fit, and test summary statistics. The item fit of IRT models is routinely evaluated within field testing and within each operational administration. Items with poor fit are removed from the item bank before they are ever administered; they are typically rewritten and field tested again. Similarly, the person fit of IRT models is routinely evaluated and the data are expected to fit well regardless of gender, ethnicity, or level of performance. Comparability of each test form is evaluated in terms of the test characteristics curves, test information function, and CSEM. Regular maintenance of item fit, person fit, and test summary statistics within a test and across test forms provides validity evidence that supports the Georgia Milestones assessments.

Evidence of validity related to response processes was obtained through a series of cognitive laboratory studies. The purpose of the cognitive laboratory was to explore how Georgia students interacted with both the graphing family of items and drag-and-drop family of items. Observers were primarily looking at item functionality and documenting any barriers to student access or understanding with these item types. The results of the study indicated that the graphing family and drag-and-drop family of item types appear to be quite intuitive for students. To ensure access to all of the content presented through these technology-enhanced item types, students will benefit from opportunities to interact with the items in advance of any testing events, such as through the Experience Online Testing Georgia, which is a tools-focused practice test available to all students.

As for evidence based on the relationship to other variables and measures, such research is ongoing. Extensive analyses of model data fit for items and persons indicate that the IRT models fit the data well for all test administrations. Research analyses evaluating relationships between the Georgia Milestones scores and the TerraNova scores and relationships between the Georgia Milestones scores and the Lexile scores have already been completed and documented. A subsequent study examining the relationship between Georgia Milestones and the ACCUPLACER has been completed.

This technical report provides extensive evidence in support of the primary purposes of the Georgia Milestones assessments. A usability study that focuses on the use of Georgia Milestones

scores will be conducted to gather additional validation evidence. An extensive survey of teachers, parents, and district staff is planned.

It is important to note that, given the interrelated nature of the goals and the evidence offered, a validity matrix that specifically crosschecks the primary purposes of the testing program and the sources of evidence has been compiled in separate documentation. Additional investigations will concentrate on fully articulating the validity arguments to support the expressed purposes and intended consequences of the Georgia Milestones testing program.

CHAPTER 9:
CONCLUSION AND DISCUSSION

The *Georgia Milestones Assessment System 2019 Operational Technical Report* documents the processes and procedures implemented to support the Georgia Milestones Assessment System (Georgia Milestones) by DRC under the supervision of the GaDOE. This technical report shows how the applied processes, procedures, and results relate to the issues of validity and reliability and to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).

Chapter 9 provides a synthesis of this technical report, with an emphasis on the conclusions about reliability and validity of the Georgia Milestones scores.

9.1 Summary of the Technical Report

Georgia Milestones began with item development and form-assembly activities for several administrations within the 2018–2019 school year. Data gathered from the initial test administrations in the 2014–2015 school year were used to build a scale of measurement that served as the basis for score reporting and test form construction. Scoring tables based on the scale of measurement and incorporating the achievement level cut scores were used to report student test scores. The reliability and validity of all the applied processes and procedures and the results were evaluated. A brief summary of the contents of this technical report is provided below:

- Item and Test Development (Chapter 2)
  - Item Sources
    - Three initial sources of items were available for form construction:
      - Items from the existing Georgia item bank, including items previously developed for Georgia Milestones, and from previous operational testing programs (i.e., CRCT and EOCT)
      - Items from the TerraNova norm-referenced achievement test
      - Newly developed items that provide new item types and/or address any gaps in the pool of items from the first two sources of items
    - Items were reviewed at content and bias review meetings in 2018.
    - Items were selected and were field tested in spring 2018.
    - Student responses to the field test items from the spring 2018 administration were analyzed.
    - Items that were field tested in spring 2017 and their respective statistics were reviewed at the data review meetings involving Georgia educators in 2017.
    - Items for the spring 2019 operational forms were selected for all grade/content areas and courses based on various statistics and the pre-equated scoring tables.

- 2019 Form Assembly
  - Using items from the three sources, two forms were built for the spring End-of-Grade (EOG) administration and two forms were built for the winter and spring End-of-Course (EOC) administrations (four total EOC forms). Each core form met the test blueprint and the specified psychometric targets.
  - The test blueprint included items from TerraNova and from Georgia Milestones. Typically, ten of the twenty TerraNova items served a dual purpose because they were used to report criterion-referenced scores and norm-referenced scores.
  - Embedded field test items were also included on the operational test forms. The number of embedded field test items on the operational forms ranged from five to twelve, depending on the grade/content area, course, and administration.

- Standards, Standard Setting, and Standards Validation (Chapter 3)
  - Chapter 3 provides a brief summary of the standard setting that took place during the 2014–2015 baseline year of Georgia Milestones and a standards validation that took place during the 2017–2018 school year. The standard setting meeting established the cut scores that distinguish four achievement levels that have been used since the initial administration of Georgia Milestones. The standards validation confirmed the ongoing use of the cut scores in science and social studies despite some changes to the test blueprints associated with the transition to the Georgia Standards of Excellence.

- Test Administration (Chapter 4)
  - To ensure that the Georgia Milestones assessments were administered in accordance with the GaDOE's mandates, the GaDOE conducted several training sessions specifically geared toward system test coordinators.
  - The following test administration ancillary materials for Georgia Milestones were developed and distributed: administration manuals, answer documents, return material forms and guidelines, security forms, and the *Interpretive Guide for Score Reports* (2019).
  - Several test security measures were implemented for Georgia Milestones. Test security procedures were discussed extensively throughout the EOG and EOC *Test Administration Manual*s (Georgia Department of Education, 2018–2019).
  - The test administration window for the EOG assessments was April 8 through May 17, 2019. For the EOC assessments, the administration windows were November 26, 2018, through January 4, 2019, for the winter administration; April 22 through May 31, 2019, for the spring administration; and June 17 through July 19, 2019, for the summer administration. Small mid-month administrations were held for the EOC assessments in each month between the main administrations for additional opportunities to test.
  - For the EOG assessments, two core forms were administered in four content areas. In English language arts (ELA) and mathematics, the tests were

administered in grades 3–8, and in science and social studies, the tests were administered in grades 5 and 8. Two core forms were administered within ten courses for each main EOC administration. Each test contained the set of operational items used to report student test performance and contained a different set of embedded field test items.

- Performance Scoring (Chapter 5)
  - Handscoring activities included an extensive set of pre-rangefinding, rangefinding, and materials development activities designed to support the scoring of operational constructed-response items and field tested constructed-response items that will be used for future form construction.
- Operational Analyses: Key Checking, Calibration, and Scaling (Chapter 6)
  - Georgia Milestones was based on the application of pre-equating with post-equating verification analyses.
  - The main objective of pre-equating was to produce a raw-score-to-scale-score conversion table for a test before the test was administered. This allowed the reporting of test results without the need to wait for equating to be conducted after the operational test administration.
  - The primary focus of post-equating verification analysis was to compare the pre-equated scoring tables and a corresponding set of post-equated scoring tables by using early return samples.
  - Post-administration calibration and equating analysis results confirmed the accuracy of the pre-equated tables. Based on the findings of the verification analyses, pre-equated scoring tables were used for scoring and reporting for the 2019 Georgia Milestones, except in grade 8 ELA, 9th Grade Literature & Composition, and American Literature & Composition, where the post-equated solution was used.
  - Local calibration of Georgia Milestones was performed. The primary focus of the local calibration was to evaluate and document item characteristics without equating. Then the results were used to update the Rasch difficulty of the 2019 operational items.
- Test Results (Chapter 7)
  - Descriptive summary statistics for the reported scale scores were reported.
  - The percentage of students at each achievement level was analyzed.
  - Both electronic and printed reports were provided for various types of reports.
  - Data files containing the demographic information of each student and scale score data for each content area were provided to the GaDOE and each system.
  - The *Interpretive Guide for Score Reports* (2019) was provided for Georgia teachers and administrators who received score reports from the 2019 administration of Georgia Milestones.

- Reliability and Validity Evidence (Chapter 8)
  - The reliability of the 2019 Georgia Milestones assessments was estimated in four ways:
    - Internal consistency using Cronbach's alpha
    - The standard error of measurement (SEM)
    - The conditional standard error of measurement (CSEM), as the reciprocal of the square root of the test information function
    - Classification consistency and accuracy
  - This technical report provides detailed documentation concerning the different phases of the testing cycle and highlights the meaning and significance of the procedures, processes, and results in terms of validity and their relationship to the *Standards*. The following final issues of validity are addressed in Chapter 8:
    - Test content
    - Response processes
    - Internal test structure
    - Relation to other variables
    - Consequences of test use

9.2 Conclusions about Reliability and Validity

9.2.1 Reliability

According to the *Standards*, reliability refers first "to the reliability coefficients of classical test theory, defined as the correlation between scores on two equivalent forms of the test. . . . Second, the term has been used in a more general sense, to refer to the consistency of scores across replications of a testing procedure" (AERA, APA, & NCME, 2014, p. 33). A reliable assessment is one that would produce stable scores if the same group of students were to take the same test repeatedly without any fatigue or memory of the test. As detailed below, the reliability of the 2018–2019 Georgia Milestones assessments was estimated in four ways:

- Internal consistency was assessed for test items by using Cronbach's alpha.
- The SEM was assessed for raw scores.
- The CSEM, as the reciprocal of the square root of the test information function, was assessed for the theta and the scale score of the three cuts.
- Classification consistency and accuracy were estimated for performance classification.

Combined, Cronbach's alpha, the SEM, the CSEM, classification consistency, and classification accuracy provide several ways of looking at the reliability of Georgia Milestones. Cronbach's alpha and the SEM operate at the content level and provide estimates of reliability for examinee scores on a test, such as the grade 4 ELA or mathematics EOG assessments. The CSEM, classification consistency, and classification accuracy operate on the associated

achievement level classifications. The consistency and accuracy of the classification of students into these performance categories is of particular interest given federal requirements stipulated in the Every Student Succeeds Act (ESSA).

Standards 2.3 and 2.13 advise providing reliability estimates and the SEM for reported scores. Accordingly, Chapters 6 and 8 present both a reliability analysis (using Cronbach's alpha) and the SEM. Reliability and the SEM are presented by content area, by total group of examinees, and by specific subgroups, thereby speaking to Standard 2.11, which advises that reliability should be assessed for all population groups.

The reliabilities and SEMs for the 2019 Georgia Milestones assessments suggest that the Georgia Milestones assessments are sufficiently reliable for their intended purpose. That is, the reliability estimates obtained for the 2019 Georgia Milestones assessments suggest that scores reported to students in 2018–2019 are well estimated and provide a reliable picture of student performance.

Standard 2.16 advises that when testing measures are used to make categorical decisions, the reliability of those decisions should be estimated. In the present context, Standard 2.16 applies specifically to achievement level determinations, such as *Proficient Learner* or *Developing Learner*. The Georgia Milestones program adhered to this standard through a detailed analysis of classification consistency and classification accuracy—two related measures with bearing upon the reliability of the achievement level classifications. This analysis also addresses Standard 2.14 by providing a CSEM for the cut scores, which separate the achievement levels.

The classification consistency and accuracy results established for 2019 suggest that consistent and accurate achievement level classifications are being made for Georgia students based on the Georgia Milestones assessments. Moreover, the CSEMs reported for the 2019 Georgia Milestones program are consistent with the previous operational testing programs (i.e., CRCT and EOCT) and further indicate that the scores reported to students in 2019 are well estimated and provide an accurate picture of student performance.

9.2.2 Validity

The *Standards for Educational and Psychological Testing* defines validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests" (AERA, APA, & NCME, 2014, p. 11). The purpose of test score validation is not to validate the test itself but to validate interpretations of the test scores for particular purposes or uses. Test score validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the entire assessment process. Every aspect of an assessment provides evidence that either supports or challenges its validity, including design, content specifications, item development, psychometric quality, and inferences made from the results.

This technical report addresses all phases of the testing cycle. Each part of this technical report details the procedures, processes, and results applied in Georgia Milestones. Each chapter also highlights the meaning and significance of the procedures, processes, and results in terms of validity and their relationship to the *Standards*. Below is a brief review:

Chapter 2 of this technical report, "Item and Test Development," describes the involvement of Georgia educators in the item and test development process. As indicated in Chapter 2, the test

development process and the involvement of Georgia educators in that process formed an important part of the validity of the entire Georgia Milestones program. Through their knowledge, expertise, and professional judgment, Georgia educators ultimately ensured that the content of the Georgia Milestones assessments formed an adequate and representative sample of appropriate content, and they ensured that the content formed a legitimate basis upon which to derive valid conclusions about student performance. This chapter thus addresses Standard 4.6 of the *Standards* (AERA, APA, & NCME, 2014). It shows that the test design process and the participation of Georgia educators in that process provided a solid rationale for the credibility of the content and design of Georgia Milestones as a tool from which to derive valid inferences about Georgia student performance. This chapter also addresses AERA, APA, and NCME (2014) Standards 1.1, 1.11, 4.0, 4.1, 4.2, 4.12, 7.2, 8.4, 12.4, and 12.8.

Chapter 2 further shows how test specification documents that were derived from earlier developmental activities guided the final phases of test development and ultimately yielded the test booklets that students used. This chapter thus addresses AERA, APA, and NCME (2014) Standards 3.1, 3.2, 3.5, 4.6, 4.7, 4.8, and 4.10.

Chapter 3 of this technical report, "Standards, Standard Setting, and Standards Validation," provides background on the standard setting activities and functions to address Standards 5.21, 5.22, and 5.23 of the *Standards* (AERA, APA, & NCME, 2014). Detailed information regarding standard setting activities is available in two separate reports: *Georgia Milestones Standard Setting Technical Report* (DRC, 2015) and *Georgia Milestones Standard Setting Technical Report for End-of-Grade and End-of-Course Science and Social Studies* (DRC, 2018).

Chapter 4 of this technical report, "Test Administration," describes the processes, procedures, and policies that guided the administration of Georgia Milestones, including accommodations, security, and written procedures provided to test administrators and school personnel. It addresses AERA, APA, and NCME (2014) Standards 4.15, 4.16, 6.1, 6.2, 6.3, 6.4, 6.6, 6.7, and 6.10.

Chapter 5 of this technical report, "Performance Scoring," describes the processes and activities implemented to ensure consistent and accurate standardized test handscoring procedures for all students. This chapter includes information on how the performance scoring procedures implemented for the handscored portions of the Georgia Milestones assessments work together to maximize scoring accuracy and consistency both within and across administrations in a pre-equated testing model. It also demonstrates adherence to AERA, APA, and NCME (2014) Standards 6.8, 6.9, 7.10, 12.14, and 12.15 in the Georgia Milestones program.

Chapter 6 of this technical report, "Operational Analyses: Key Checking, Calibration, and Scaling," describes the data used for calibration and scaling. It shows that the data used to determine the validity of calibration and scaling were sufficiently representative of Georgia students. Raw score results and a classical item analysis were provided, and these served as a foundation for subsequent analyses. This chapter also describes the calibration and scaling processes, procedures, and results. Some references to introductory and advanced discussions of item response theory are provided. Several axes upon which to evaluate the calibration and scaling procedures are discussed, such as the data used, the software applied, and the successful estimation of parameters, item fit, and the SEM. This chapter thereby demonstrates adherence to AERA, APA, and NCME (2014) Standards 1.8, 1.10, 2.3, 2.13 2.14, 2.15, 2.19, 3.6, 4.14, 5.1, 5.2, 5.13, 7.2, and 7.4.

Chapter 7 of this technical report, "Test Results," presents scale score results, achievement level information, and Lexile scores. Scale score results provide a basic quantitative reference to student performance as derived through the Rasch model that was applied. The achievement level information spoke directly to the achievement level requirements of the No Child Left Behind and ESSA policy environments and to parents and guardians, students, and educators. The Lexile scores then further provide a tool to match a student's reading ability with the difficulty of text material. The combination of scale scores, achievement levels, TerraNova national percentiles, and Lexile scores provided a comprehensive set of information tools to assess Georgia student performance by content area, grade-level, ethnicity, and gender. This chapter thus addresses AERA, APA, and NCME (2014) Standards 5.10, 5.11, 6.1, 6.2, 6.3, 6.4, 6.10, 6.12, 13.15, and 13.19.

Chapter 8 of this technical report, "Reliability and Validity Evidence," spends its first half demonstrating adherence to the AERA, APA, and NCME (2014) *Standards* through several analyses of the reliability of the 2019 Georgia Milestones assessments. It presents a reliability analysis using Cronbach's alpha, SEM results, CSEM results, and a detailed analysis of classification consistency and accuracy. The 2019 Georgia Milestones assessments thereby addresses AERA, APA, and NCME (2014) Standards 2.3, 2.7, 2.11, 2.13, 2.14, and 2.15.

The second half of Chapter 8 addresses the collection of validity evidence and specifically addresses test content, response processes, issues of bias, dimensionality analysis, relations to other tests, and consequences of test use. It demonstrates adherence to AERA, APA, and NCME (2014) Standards 1.1, 1.9, 1.11, 3.5, 3.6, 3.16, 4.0, 4.6, 4.7, 4.8, 4.10, 4.12, 7.2, and 8.4. This chapter ends with a section addressing the development of validity arguments for Georgia Milestones.

The GaDOE and DRC maintained an unwavering focus on the gathering of validity evidence in support of Georgia Milestones throughout the development, administration, analysis, and reporting of the 2019 Georgia Milestones administration. Subsequent documentation will be further gathered to augment validation evidence in support of Georgia Milestones.

REFERENCES

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

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. doi:10.2307 /1165285

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. doi:10.1177/001316446002000104

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Group/Thomson Learning.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3−17). Hillsdale, NJ: Erlbaum.

Dale, E., & O'Rourke, J. (1976). *The living word vocabulary: A national vocabulary inventory*. Unknown binding.

Data Recognition Corporation. (2015). *Georgia Milestones standard setting technical report*. Maple Grove, MN: Author.

Data Recognition Corporation. (2015). *Linking the Georgia milestones end-of-grade and end-of-course reading and vocabulary with the lexile framework for reading*. Maple Grove, MN: Author.

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice, and a proposed framework. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). New York, NY: Routledge.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233.

Georgia Department of Education. (2018). *Accommodations manual: A guide to selecting, administering, and evaluating the use of test administration accommodations for students with disabilities*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Assessment guide: Algebra I.* Atlanta, GA: Author.

Georgia Department of Education. (2018). *Assessment guide: American literature and composition*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Assessment guide: Analytic geometry*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Assessment guide: Biology*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Assessment guide: Coordinate algebra*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Assessment guide: Economics*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Assessment guide: Geometry*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Assessment guide: Grade 3*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Assessment guide: Grade 4*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Assessment guide: Grade 5*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Assessment guide: Grade 6*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Assessment guide: Grade 7*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Assessment guide: Grade 8*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Assessment guide: Ninth grade literature and composition*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Assessment guide: Physical science*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Assessment guide: United states history*. Atlanta, GA: Author.

Georgia Department of Education. (2015–2019). *End-of-grade (EOG) interpretive guide for score reports for use with score reports from spring and summer 2019 administrations*. Atlanta, GA: Author.

Georgia Department of Education. (2015–2019). *End-of-course (EOC) interpretive guide for score reports for use with score reports from winter 2018–summer 2019 administrations*. Atlanta, GA: Author.

Georgia Department of Education. (2017). *The Georgia milestones style guide*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Georgia student assessment program student assessment handbook 2018–2019*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Winter 2018; spring, summer, and fall mid-month 2019 end-of-course test administration manual for system test coordinators, school test coordinators, and examiners*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Winter 2018; spring, summer, and fall mid-month 2019 end-of-course paper-and-pencil test administration supplement for system test coordinators, school test coordinators, and examiners*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Spring and summer retest 2019 end-of-grade test administration manual for system test coordinators, school test coordinators, and examiners*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Spring and summer retest 2019 end-of-grade paper-and-pencil test administration supplement for system test coordinators, school test coordinators, and examiners*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Study/resource guide for students and parents: Algebra I*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Study/resource guide for students and parents: American literature and composition*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Study/resource guide for students and parents: Analytic geometry*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Study/resource guide for students and parents: Biology*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Study/resource guide for students and parents: Coordinate algebra*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Study/resource guide for students and parents: Economics*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Study/resource guide for students and parents: Geometry*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Study/resource guide for students and parents: Grade 3*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Study/resource guide for students and parents: Grade 4*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Study/resource guide for students and parents: Grade 5*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Study/resource guide for students and parents: Grade 6*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Study/resource guide for students and parents: Grade 7*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Study/resource guide for students and parents: Grade 8*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Study/resource guide for students and parents: Ninth grade literature and composition*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Study/resource guide for students and parents: Physical science*. Atlanta, GA: Author.

Georgia Department of Education. (2018). *Study/resource guide for students and parents: United states history*. Atlanta, GA: Author.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Hingham, MA: Kluwer-Nijhoff Publishing.

Harris, A. J., & Jacobson, M. D. (1982). *Basic reading vocabularies*. New York, NY: Macmillan Publishing Company.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

House, E. R. (1980). *Evaluating with validity*. Beverly Hills, CA: Sage.

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice*, *25*(1), 4–12. doi:10.1111/j.1745-3992.2006.00047.x

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. (2013a). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kane, M. (2013b). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement*, *50*(1), 115–122.

Kim, D. (2005). KKCLASS [Computer software]. Unpublished.

Kim, D., Barton, K., & Kim, J. (2007). *Estimating classification consistency and classification accuracy with pattern scoring*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Kim, D., Choi, S. W., Um, K. R., & Kim, J. (2006). *A comparison of methods for estimating classification consistency*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Kolen, M. J., & Kim, D. (2005). Personal correspondence.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The bookmark standard setting procedure. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 225–253). New York, NY: Routledge.

Linacre, J. M. (2005). WINSTEPS Rasch measurement [Computer software]. Chicago, IL: Winsteps.com

Linacre, J. M. (2014). WINSTEPS (Version 3.81.0) [Computer software]. Retrieved from http://www.winsteps.com

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*(2), 179–197. doi:10.1111/j.1745-3984.1995.tb00462.x

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*(4), 719–748. doi:10.1093 /jnci/22.4.719

Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement*, *9*(3), 200–215.

McDonald, R. P. (1979). The structural analysis of multivariate data: A sketch of a general theory. *Multivariate Behavioral Research*, *14*, 21–38. doi:10.1207/s15327906mbr1401_2

Mogilner, A., & Mogilner, T. (2006). *Children's writer's word book* (2nd ed.). Cincinnati, OH: Writer's Digest Books.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Shepard, L. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.). Review of Research in Education, Vol. 19. Washington, DC: AERA.

Sireci, S. (2013). Agreeing on validity arguments. *Journal of Educational Measurement, 50*(1), 99–104.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement, 11*(4), 263–267. doi:10.1111/j.1745-3984.1974.tb00998.x

Taylor, S. E., Frackenpohl, H., White, C. E., Nieroroda, B. W., Browning, C. L., & Birsner, E. P. (1989). *EDL core vocabularies in reading, mathematics, science, and social studies*. Austin, TX: Steck-Vaughn Company.

Thompson, S., & Thurlow, M. (2002). *Universally designed assessments: Better tests for everyone!* (Policy Directions No. 14). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

U.S. Department of Education. (2007). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind act of 2001*. Retrieved from http://www.ed.gov/policy/elsec/guid/saaprguidance.pdf

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187–213. doi:10.1111/j.1745-3984.1993.tb00423.x

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233–251. doi:10.1111/j.1745-3984.1993.tb00425.x