# The North Carolina Department of Public Instruction

# Edition 5

# Mathematics 3 – 8 End-of-Grade (EOG)

# NC Math 1 and NC Math 3 End-of-Course (EOC)

# Technical Report

# 2018–2019

**PUBLIC SCHOOLS OF NORTH CAROLINA**
**Division of Accountability Services | North Carolina Testing Program**
Department of Public Instruction | State Board of Education

Prepared by:

North Carolina Department of Public Instruction
Accountability Division
December 2020

Tammy Howard, Ph.D., Director of Accountability

Maxey-Moore, Section Chief, Test Development

Kinge Mbella, Ph.D., Lead Psychometrician

Thakur Karkee, Ph. D., Psychometrician

## TABLE OF CONTENTS

## TABLE OF TABLES

## TABLE OF FIGURES

## APPENDICES

# CHAPTER 1 INTRODUCTION

The intent of this technical report is to provide comprehensive and detailed evidence in support of the validity and reliability of the North Carolina State Testing Program (NCSTP). The first part of this report presents a brief overview of the revision and eventual adoption of new mathematics content standards which is used to justify the development of new assessments. The remaining sections describe a brief history of the NCSTP followed by documentation of item development and review, field test and analysis, and form development and review. The report concludes with summaries of standard setting workshop used to set achievement levels for reporting and interpreting, student results, and validity evidences for the *Edition 5* End-of-Grade (EOG) and End-of-Course (EOC) mathematics summative assessments.

## 1.1 Purpose and Background of the North Carolina State Testing Program

The General Assembly GCS 115C-174.10T specified the purpose of the NCSTP as:

*"(i) to assure that all high school graduates possess those minimum skills and that knowledge thought necessary to function as a member of society; (ii) to provide a means of identifying strengths and weaknesses in the education process in order to improve instructional delivery; and (iii) to establish additional means for making the education system at the State, local, and school levels accountable to the public for results."*

With the above purposes as a guide, the North Carolina State Board of Education (NCSBE) developed the School-Based Management and Accountability Program to improve student performance in the early 1990s. The current vision of the NCSBE is for *"Every public school student will be empowered to accept academic challenges, prepared to pursue their chosen path after graduating high school, and encouraged to become lifelong learners with the capacity to engage in a globally-collaborative society."* The current mission of the NCSBE is to use its constitutional authority to guard and maintain the right of sound and basic education for every child in North Carolina Public Schools. The NCSBE's three main goals are to:

- *Eliminate opportunity gaps by 2025*
- *Improve school and district performance by 2025*
- *Increase educator preparedness to meet the needs of every student by 2025.*

Starting from the early 1990s, North Carolina has continually sought innovation in the design, development, and ways to use state assessments to increase academic expectations, so students are prepared for success after high school. This is evident in the NCSBE stated goals and policy of continuous academic content standards evaluation and review. The NCSBE mandates that the North Carolina Department of Public Instruction (NCDPI) review content standards every five to

seven years after they were first adopted. This also implies that state assessments are also reviewed and redesigned to ensure they are up to date with current measurement practices and aligned to academic expectations of current North Carolina *Standard Course of Study* (NC*SCoS*).

In 1994, EOG assessments, designed to measure the NCSBE-adopted content standards, were administered for the first time to all students in grades 3–8. In 1996, the accountability system, referred to as Accountability, Basics and Local Control (ABCs), was implemented using data from the EOG assessments to inform parents, educators and the public annually on the status of achievement at the school level. In the 1997–98 school year, EOC tests were added and used to the ABCs school accountability model. The ABCs model business rules were fine-tuned to ensure schools were being held accountable for all students.

In 2013, ABCs was replaced by the READY accountability model after the NCSBE adopted new *Common Core State Standards* for mathematics and English Language Arts/Reading and the North Carolina Essential Standards for Science. The NCDPI developed and administered new EOG and EOC assessments aligned to the newly adopted common core standards. The READY model was used to measure the progress of students in grades 3 through 8 and high school. The assessment results provided summative evaluative data aimed at informing parents, teachers and students on their relative standing based on grade level expectation as specified in the adopted content standards. Student test data from the EOG and EOC were also used to determine each school's Adequate Yearly Progress (AYP) as required by the federal No Child Left Behind Act (NCLB).

In 2016, the NCSBE replaced the *Common Core State Standards* with new NC*SCoS* for high school mathematics. In 2017 under the leadership from the NCSBE, the *Common Core State* Standards for mathematics and reading were replaced by new North Carolina Standards. To maintain strong content alignment and validity evidences of uses and score interpretations, EOG and EOC assessments were redesigned and new operational forms aligned to the new NC*SCoS* were operationally administered in the 2018–19 school year. This technical report documents all steps and processes that were implemented in the development, administration, scoring and reporting of results for *Edition 5* of EOG and EOC mathematics assessment. The purpose of this report is to demonstrate the NCDPI's continuous commitment to the highest standards and technical quality of its EOG and EOC assessments.

## 1.2 North Carolina Content Standards Review, Revision, and Implementation Processes

General Assembly of North Carolina *Session Law 2014-78 Senate Bill 812* (see *Appendix 1-A*) has enacted the Academic Standards Review Commission (ASRC) composed of 11 members to conduct a comprehensive review of all English Language Arts and Mathematics standards that

were adopted by the State Board of Education under G.S. 115C-12(9c) and propose modifications to ensure that those standards meet all of the following criteria:

- Increase students' level of academic achievement
- Meet and reflect North Carolina's priorities
- Are age-level and developmentally appropriate
- Are understandable to parents and teachers
- Are among the highest standards in the nation

In accordance with these frameworks, the ASRC started the comprehensive review process of English Language Arts (ELA) and mathematics content standards in 2015. The findings and recommendations of ASRC's reviews are documented in the commission's report (*Appendix 1-B*). In early 2016, the NCDPI division of Standards, Curriculum and Instruction (SC&I) started reviewing the recommendations. A formalized review framework of the recommendations was built on four guiding principles with the aim to promote transparency and stakeholder engagement throughout every step of the standards review, revision and implementation process. The four principles are:

- Feedback-Based: The NCDPI formally collects feedback on the current standards from educators, administrators, parents, students, institutions of higher education, business/industry representatives, national organizations and other education agencies.
- Research-Informed: The NCDPI reviews contemporary research on standards and learning in the content area under review. Benchmarking with other states, third-party reviews and comparability of national and international standards and trends in order to inform the process.
- Improvement-Oriented: The NCDPI provides the State Superintendent and State Board of Education an annual report summarizing feedback received from stakeholders concerning standards and implementation.
- Process-Driven: The system process includes three phases: review, revision and implementation.

Using these four guiding principles as a framework, the SC&I division developed and implemented a plan of action and timeline in 2016 to review and revise the mathematics content standards. During the review phase, SC&I worked with the ASRC as facilitator to help with research and provided guidance on state and federal policy requirements. The SC&I division's role was also to gather and present inputs from stakeholder groups (educators, parents, business and industry leaders, community leaders and members of society at large) through survey and webinars. The division was also tasked with updating the NCSBE on the commission progress throughout the process.

Following the review, SC&I adopted a 6-step iterative process summarized below to revise and draft new mathematics content standards.

- Establish and convene content-standard writing teams.
- Share drafted standards with local districts, charter schools and other stakeholders for at least 30 days of review and input.
- Engage the data review committee to compile feedback to share with the writing teams.
- Reconvene the writing teams to review the feedback and incorporate changes.
- Share additional drafts for stakeholder reviews and inputs.
- Submit the final revised standards to the NCSBE for approval.

The final phase in the framework was the implementation of the new content standards. To ensure a smooth transition at every level of the school system in the areas of instruction and assessment SC&I also enacted a detailed 4-step implementation plan summarized below:

- Launched and disseminated a state-level standards implementation plan including samples, phase wise extension and full-fledged implementation to local districts and charter schools.
- Modified the annual statewide assessment program as necessary in accordance with the revised standards.
- Facilitated statewide professional development training and supports for educators on the revised standards.
- Collected data and evaluated the implementation of the revised standards.

*Tables 1.1* and *1.2* outline detail timelines and brief descriptions of actions that were implemented by the NCDPI during the review, revision and implementation of the new NC*SCoS* for mathematics from 2016 through 2018. *Table 1.1* shows timelines for high school mathematics content standards for NC Math 1, NC Math 2 and NC Math 3. *Table 1.2* shows the timeline for K–8 mathematics content standards. These timelines show how the four (4) principles outlined by the NCSBE were operationalized and implemented into actionable steps during the review, revision and implementation of the new mathematics standards. Additional review materials and activities highlighted in *Table 1.1* are detailed in *Appendix 1-C.* Additional details for K–8 mathematics content standards review, revision and implementation are documented in *Appendix 1-D.*

*Table 1. 1 NC Math 1–3 Standards Review, Revision and Implementation Timeline*

| Date | Actions | Descriptions |
|---|---|---|
| January 2016, | Material preparations and group invitation | Comments from teacher surveys, community surveys, teacher focus groups and Academic Standards Review Commission's final report were assembled in preparation for data review teamwork. |
| February 4 –5 and March 4 –5, 2016 | Data analysis/review group meetings | Team of 23 educators (composed of ten (10) teachers, seven (7) district mathematics leaders and six (6) professors from NC colleges) from across the state convened in Raleigh to conduct a thorough analysis of all feedback as a part of the data analysis/review component of the standards review process. Business and industry were represented through Chamber of Commerce, which sent two representatives to address the group and sit in on part of the process. The responsibilities of the data review team included:<br>1) Read through all data by standard across all three high school courses,<br>2) Identify and record common themes,<br>3) Apply professional input based on comments,<br>4) Reach small group consensus on whether to keep standard, revise, remove or move and revise,<br>5) Reach large group consensus by standard and course and<br>6) Develop guidelines for revision work. |
| March 9 –20, 2016 | Revision and rewriting process preparation | The members of the data review team were invited to begin the revision and rewriting phase. Invitees included seven (7) teachers, four (4) district/state leaders and four (4) higher-education representatives. Western, Central and Eastern regions of NC were represented.<br><br>The review teams were formed by conceptual category and then regrouped by course. Each revision team received summary documents from review meetings. |

| Date | Actions | Descriptions |
|---|---|---|
| March 21 – April 15, 2016, | Revise and rewrite | Writing teams met virtually and in-person to apply recommendations of the data review team. These recommendations included overarching guidance such as:<br><br>1) Examine and rewrite all standards for clarity by removing unnecessary language and examples.<br>2*) Move Geometry standards according to topic.<br>3) Delete/move identified standards to fourth mathematics courses.<br>4) Limit overly broad standards by rewriting with clearer direction for teachers based on the identification of functions and systems per course.<br>5*) Remove identified standards that span 2 or 3 courses that were viewed as duplicative and unnecessary.<br>6*) A first draft was produced at the conclusion of the meeting held in Greensboro and communicated with LEAs for public comments. |
| April 18 – April 26, 2016 | Public review of draft | Draft of revised high school mathematics courses made public for comments and posted to eBoard for the NCSBE. |
| April 26 – May 3, 2016 | Data feedback summary | The NCDPI's SC& I prepared feedback on drafts of revised courses for NCSBE presentation. Noted any revisions needed and applied. |
| May 4 –5, 2016<br>June 1 –2, 2016 | NCSBE meeting | The revised draft was presented to NCSBE for discussion. The NCSBE voted for adoption of revised high school mathematics courses. |
| June 6 – July 30, 2016 | District Leaders and Teacher training | Regional meetings with district leaders and high school mathematics teachers across the state were held to communicate and begin training under new standards using statewide system of support and service. |

* Recommended by or contained in final report of the Academic Standards Review Commission (ASRC)

*Table 1. 2 Grades K – 8 Mathematics Standards Review, Revision and Implementation Timeline*

| Date | Action | Description |
|---|---|---|
| September–October 2016 | K– 8 mathematics data review committee* meeting | The committee conducted a thorough analysis of all feedback from teacher surveys, community surveys, teacher focus groups, leader focus group and the ASRC. |
| December 2016–January 2017 | Mathematics data review committee findings compiled and shared | Updates were presented at the December 2016 and January 2017 NCSBE meeting. Results from September–October 2016 data review committee meetings shared including how the results and data from ASRC, parent and community survey, teacher survey, mathematics leader focus group and teacher focus groups were organized and prioritized for K–8 to identify priorities, concerns and changes. |
| January – February 2017 | Writing teams** convene | Writing teams convened to create drafts of K–8 standards. |
| March 2017 | Drafts K–8 standards released | Drafts of K–8 mathematics standards were released to public for comments and shared with the NCSBE. |
| April 2017 | Update on drafts | Updates provided to the NCSBE on data collection from public comments. |
| May 2017 | K–8 mathematics drafts presented to the NCSBE | Final drafts with comments shared with the NCSBE. |
| June 2017 | Actions taken on K–8 mathematics drafts | Presented K–8 draft mathematics standards to the NCSBE for adoption. |
| June – August 2017 | Regional professional development (PD) sessions | The NCDPI hosted regional PD and information sessions on revisions of K–8 mathematics standards by grade bands of K–2, 3–5 and 6–8 for teachers and district leadership. |
| August 2017 | Implementation of the standards | Districts implemented new standards. The NCDPI continued its support as a PD trainings and webinars. |

*K–8 Mathematics Data Review Committee: 60–75 Educators including teachers, district mathematics leaders, and Institute of Higher Education (IHE) representatives

**K–8 Mathematics Writing/Revision Committee: 35–40 Educators including teachers, district leaders, and IHE representatives

The attributes described above are a part of validity evidences to show that North Carolina mathematics standards are research based and have adequate rigor and expectation to prepare North Carolina students for college and/or challenging careers after high school. To maintain content and construct validity evidences of EOG and EOC assessment score uses and interpretation, North Carolina redesigned and administered new assessments that are aligned to the new adopted mathematics content standards. *Table 1.3* shows an overview of the timeline beginning with adoption of new content standards to development and reporting of scores aligned to these new mathematics content standards.

*Table 1. 3 NCDPI Accountability and Testing Highlights*

| Year | Action |
|---|---|
| June 2016 | The NCSBE adopted the revised standards for high school Mathematics (NC Math 1, 2 and 3). |
| August 2016 | The revised standards for high school mathematics were implemented. |
| June 2017 | The NCSBE adopted the revised standards for grade 3–8 Mathematics. |
| August 2017 | The revised standards for mathematics grades 3–8 were implemented. |
| 2017–18 | Item development and field-test |
| 2018–19 | Edition 5 EOG and EOC assessments developed, administered, and score reported on new achievement level scale. |

## 1.3 Overview of the North Carolina Statewide Assessment Program

The NCDPI designs, develops and administers high-quality statewide mathematics assessments in grades 3–8 and high school that are aligned to NC*SCoS* with Career- and College-Ready (CCR) expectations for students. EOG and EOC assessment scores provide valid and reliable information intended to serve two general purposes: measure students' performance and progress as it relates to their proficiency towards grade-level content standards and serve as a quantitative indicator for use in federal and statewide accountability models.

- Measure students' performance and progress: North Carolina EOG and EOC assessments are used to measure whether students are performing at a level that indicates they consistently demonstrate mastery of the content standards. These assessments are designed to measure student performance on the full breadth and depth of grade-level content standards. Student performance on EOG and EOC assessments is reported using scale scores grouped into one

of four achievement levels (Not Proficient, Level 3, Level 4, and Level 5). Additionally, state board policy requires that EOC scores make up a minimum of 20% of student course grades.

- Federal and State Accountability Models: EOG and EOC assessments are used, as required by federal and state law, as indicators in the school accountability models. These models are designed to identify schools in need of support. Specifically, these assessment scores are used as measures of proficiency and academic growth as defined using SAS© Education Value-Added Assessment System (EVAAS) under the current accountability systems.

The North Carolina *Testing Code of Ethics* (https://www.dpi.nc.gov/documents/testing-code-ethics) cautions educators to use EOG and EOC test scores and reports only for these intended uses as approved by the NCSBE and for which the NCDPI has provided validity evidence to support these intended uses. It also reiterates that test scores are only one of many indicators of student achievement. The use of EOG and EOC test scores for purposes other than those intended by the NCDPI must be supported by evidences of validity, reliability/precision, and fairness.

## 1.4 Overview of the Technical Report

Chapter 1 provides a brief history of testing in North Carolina; the standards review, revision and implementation process; and overview of the North Carolina statewide assessment program.

Validity is a unifying and core concept in test development and, thus, Chapter 2 documents an overview of NCSTP test design, item development process and field-test plans. The test design sections include description of test specifications meetings, test blueprints, cognitive complexity, item format and mode of test administration. An overview of item development process which includes item writer training, item writing, and reviews is also documented. Final section describes field-test plans to replenish item pool for future test development.

Chapter 3 describes the field-test item analysis plans using Classical and Item Response theory as well as differential item functioning analysis. The NCDPI has set internal criteria for filtering out items with less-than-optimal characteristics. Final sections describe summary of item analysis and separate and concurrent item parameter calibration of item responses for the purpose of building parallel forms.

Chapter 4 starts with automated form assembly process using *Edition 4* test characteristic curves and test information functions as preliminary statistical targets. In subsequent sections, descriptions of 26-step operational form assembly and review processes are documented. Summary analysis of parallel forms developed for each of the EOG and EOC grades/levels, based on the field-test statistics are documented. This chapter also documents evidences to show parallel forms are comparable and meet all content, blueprint, and statistical specifications. The chapter further documents the structure of the base forms in terms of item types and cognitive

complexity, and descriptive classical and IRT statistics based on the field-test data. Also, figures displaying test characteristic curves, test information functions, and conditional standard error of measurements are presented.

Chapter 5 documents procedures put in place by the NCDPI to assure the administration of EOG and EOC assessments are standardized, fair, and secured for all students across the state. The chapter also describes training provided to test administrators, test security, and accommodation procedures implemented to ensure all students have equal and fair access to EOG and EOC assessments. The chapter concludes with description of student participation and processes used for identifying test irregularities and misadministration.

Chapter 6 describes processes used for scoring and scale development procedure adopted to create final reportable scale scores. The chapter begins with describing IRT scoring and scale scores, documenting final IRT results based on post calibration, IRT summed score procedure and score comparability across forms and modes. Final sections describe raw to scale scores and score certification processes.

Chapter 7 presents a summary of the standard setting study that was conducted in July 2019 after the first operational administration of EOGs and EOCs. The NCDPI contracted with Data Recognition Corp (DRC) to conduct a standard setting workshop to recommend cut scores and achievement levels for the newly developed mathematics EOG and EOC assessments. The chapter is a condensed version of the final report prepared by the DRC describing the full workshop and final cut score recommendations. Final sections document validity of the standard setting in terms of participants' evaluation of standard setting processes as well as evaluation of the process by external evaluators.

Chapter 8 summarizes performance results for EOG and EOC assessments for the 2018–19 operational administrations. This chapter is organized into three main sections. The first section highlights descriptive summary results of scale scores and achievement levels for EOG and EOC forms across major demographic variables. The second section presents sample reports and descriptions and stakeholders of the various standardized reports created by the NCDPI. The final section briefly describes confidentiality of student information.

Chapter 9 presents validity evidences collected in support of the interpretation of EOG and EOC test scores. The first two sections in this chapter present validity evidences in support of internal structure of EOG and EOC assessments. Evidence presented in these sections includes reliability, standard error estimates, classification consistency summary of reported achievement levels and exploratory Principal Component Analysis in support of the unidimensional analysis and interpretation of scores. The final sections of the chapter document validity evidences based on relation to other variables summarized from the EOG/EOC Quantile® Framework linking study,

and the last section presents a summary of procedures used to ensure EOG and EOC assessments are accessible and fair to all students.

*Table 1. 4 Glossary of Abbreviations*

| Abbreviations | Full Form |
|---|---|
| 2PL | Two-Parameter Logistic |
| 3PL | Three-Parameter Logistic |
| ALD | Achievement Level Descriptor |
| ASRC | Academic Standards Review Commission |
| AYP | Annual Yearly Progress |
| CBT | Computer-Based Test |
| CCR | Career- and College-Ready |
| CMH | Cochran-Mantel-Haenszel |
| CTT | Classical Test Theory |
| DD | Drag and Drop |
| DIF | Differential Item Functioning |
| DLP | Data Leak Protection |
| DOK | Depth of Knowledge |
| DRC | Data Recognition Corporation |
| EAP | Expected a Posteriori |
| EC | Exceptional Children |
| EDS | Economically Disadvantaged Students |
| EL | English Learner |
| ELA | English Language Arts |
| EOC | End-of-Course |
| EOG | End-of-Grade |
| FERPA | Family Educational Rights and Privacy Act |
| GR | Graded Response |
| HOSS | Highest Obtainable Scale Score |
| ICC | Item Characteristic Curve |
| IEP | Individualized Education Plan |
| IRT | Item Response Theory |
| LOSS | Lowest Obtainable Scale Score |
| MC | Multiple Choice |
| MCE | Minimally Competent Examinee |
| MH | Mantel-Haenszel |
| MOU | Memorandum of Understanding |
| | North Carolina |
| | North Carolina Department of Public Instruction |

| Abbreviations | Full Form |
|---|---|
| NCLB | No Child Left Behind |
| NCSBE | North Carolina State Board of Education |
| NC*SCoS* | North Carolina Standard Course of Study |
| NCSTP | North Carolina State Testing Program |
| NCSU-TOPS | North Carolina State University-Technical Outreach for Public Schools |
| NCTAC | North Carolina Technical Advisory Committee |
| NE | Numeric Entry |
| OTISS | Online Testing Irregularity Submission System |
| PBT | Paper-Based Test |
| PCA | Principle Component Analysis |
| PII | Personally Identifiable Information |
| QSC | Quantile Skills and Concepts |
| RAC | Regional Accountability Coordinator |
| SC&I | Standards, Curriculum and Instruction |
| SE | Standard Error |
| TCC | Test Characteristic Curve |
| TD | Targeted Drop |
| TDS | Test Development System |
| TE | Technology Enhanced |
| TI | Text Identify |
| TIF | Test Information Function |
| TMS | Test Measurement Specialist |
| VI | Visually Impaired |

# CHAPTER 2 TEST DESIGN, ITEM DEVELOPMENT, AND FIELD-TEST PLAN

This chapter documents steps implemented by the NCDPI during the development of *Edition 5* mathematics EOG and EOC assessments in adherence with Standard 4.0 (AERA, APA, & NCME, 2014) which states *"…Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population*" (p. 85). Specifically, this chapter describes the test specification processes – content blueprint, test format, item development and review. The last section describes the item tryout plans used to field-test newly developed items for *Edition 5* EOG and EOC forms.

## 2.1 Test Specifications

The EOG and EOC assessments are standards based that serve summative purposes. These assessments were redesigned so they were aligned with new mathematics content standards adopted in 2016–17 to ensure adequate validity evidences in support of standard-based interpretation of test scores. The second step in the development of the new assessments is guided by the overall test specifications which outline all essential content, cognitive and psychometric specifications.

The NCDPI recruited North Carolina teachers and educators from across the state and conducted two on-site test specification workshops in April and July 2017. Participants invited to these meetings represented North Carolina educators and teachers from across all geographic regions, demographic subgroups and experiences. Participants also included Special Education and English Learners educators to ensure fairness and accessibility of EOG and EOC assessments for all North Carolina students. Full agendas, surveys, and complete demographic characteristics of workshop participants by grade span are tabulated in *Appendix 2–A*. The main purposes of these test specification workshops were to specify content, cognitive rigor, test format blueprints and psychometric specifications for *Edition 5* EOG and EOC assessments.

### 2.1.1 Content Blueprint

The two-day on-site test specification workshops were facilitated by the NCDPI Test Development staff and designed to get participants to recommend content blueprints for *Edition 5*. The workshops were held by grade spans: NC Math 1 and NC Math 3, EOG grades 6–8 and EOG grades 3–5. During these workshops, participants were tasked to recommend content domain blueprints for each grade. Workshops started with an overview presentation of the purposes of EOG and EOC assessments followed by an overview of the new mathematics content standards. Participants were then separated into smaller work groups, and each group

was assigned a group lead to facilitate discussions. The first major task for participants was to recommend content blueprint weights by domain. These recommendations were done in two rounds with large group discussions between rounds.

In Round 1, following group discussions of grade-level content standards as they relate to EOG and EOC assessments, participants were directed to individually assign 0–10 ratings on a Google form with "0" indicating a particular standard cannot be assessed based on the proposed assessment design to "10" indicating a standard can be assessed and is of the highest importance. At the conclusion of Round 1, all ratings were aggregated and summarized to generate recommended domain content distribution weights.

The Round 1 recommendations from all participants were aggregated and presented to the larger group for open discussions. Group discussions were prioritized for standards with the highest ranges of ratings among participants. During these group discussions participants were given an opportunity to justify their ratings and share their rationale with the entire room. Following large group discussions, participants returned to their smaller groups for one final round of recommendations.

In Round 2, participants were encouraged to rely on information shared from the lager group discussions to determine if they wanted to revise any ratings. At the conclusion of Round 2 reviews, the updated recommended content weights were presented as their final grade-level content blueprint recommendations.

At the end of test specification workshops, the NCDPI team members from Test Development and SC&I reviewed the recommended blueprints to ensure adequate across-grades articulation. The final recommendations shown in *Table 2.1* and *Table 2.2* were then adopted as *Edition 5* mathematics content blueprints for EOC and EOG assessments.

*Table 2.1 Mathematics EOC Test Blueprint (%), High School*

| EOC Domains | NC Math 1 | NC Math 3 |
|---|---|---|
| Functions | 32–36 | 32–36 |
| Geometry | 8–12 | 20–24 |
| Statistics and Probability | 18–20 | 8–12 |
| Number and Quantity and Algebra | 36–40 | 32–36 |

*Table 2.2 Mathematics EOG Test Blueprint (%), Elementary and Middle School Grades*

| Domains | Grades | | |
|---|---|---|---|
| **Elementary** | **3** | **4** | **5** |
| Operations and Algebraic Thinking | 32–36 | 14–18 | 9–13 |
| Number and Operations in Base Ten | 9–13 | 25–29 | 25–29 |
| Number and Operations - Fractions | 28–32 | 30–34 | 39–43 |
| Measurement and Data, Geometry | 23–27 | 23–27 | 19–23 |
| **Middle-School** | **6** | **7** | **8** |
| Ratios and Proportional Relationships | 24–28 | 24–28 | |
| The Number System | 20–24 | 8–12 | |
| Expressions and Equations | 22–26 | 20–24 | |
| Geometry | 12–16 | 16–20 | 24–28 |
| Statistics and Probability | 12–16 | 22–26 | 16–20 |
| The Number System, Expressions and Equations | | | 24–28 |
| Functions | | | 28–32 |

### 2.1.2 Content Cognitive Complexity

On Day 2 of the test specification workshop, participants were tasked to evaluate and recommend content cognitive complexity expectation ranges for all assessable standards to guide item and test development. The NCDPI adopted the Norman Webb depth of knowledge (DOK) classification (Hess, 2013) as the basis for evaluating content complexity for EOG and EOC assessment items. A general definition for each of the four DOK levels is shown in *Appendix 2–B.* The DOK levels offer a framework for content experts to differentiate learning expectations and outcomes by considering the level of thinking required by students to successfully engage with items aligned to specific content standard expectations. Prior to the test specification workshops, the NCDPI Test Development and SC&I staff received training on Webb's DOK classifications on April 2017 from Dr. Karen Hess. The Webb's DOK levels guide used during training by Dr. Hess is shown in *Appendix 2–C.*

At the test specification workshop, the NCDPI staff provided an overview training on Webb's DOK to ensure participants had the necessary working knowledge needed for this activity. They then participated in two rounds of discussions and recommendations of DOK expectations.

In Round 1, participants were separated into smaller working groups and their task was to set DOK range expectations by standards. Classification ratings from each group were recorded using Google forms and the final data from all groups were uploaded into a final table and

reviewed with the entire large group. The large group discussions were used to give participants an opportunity to review and justify their ratings and make any necessary changes.

The final recommended DOK classifications from Round 2 were then adopted as the expected content cognitive complexity recommendations for assessed mathematics content standards. At the conclusion of the meeting, the NCDPI's Test Development and Standard, Curriculum and Instruction division staffs reviewed these recommended classifications to ensure coherent alignment with grade-level content standards expectations and summarized the data into DOK range specifications for EOG and EOC assessments. The final content cognitive complexity specifications for *Edition 5* EOG and EOC mathematics tests are shown in *Table 2.3*.

*Table 2.3 Proposed DOKs Across Grades*

| Grade | Number of Items | Category (%) | | |
|---|---|---|---|---|
| | | DOK 1 | DOK 2 | DOK 3 |
| 3 | 40 | 40–50 | 50–60 | |
| 4 | 40 | 35–45 | 50–60 | 5 |
| 5 | 40 | 30–40 | 50–60 | 8–10 |
| 6 | 45 | 25–35 | 50–60 | 8–15 |
| 7 | 45 | 25–35 | 50–60 | 8–15 |
| 8 | 45 | 25–35 | 50–60 | 8–15 |
| NC Math 1 | 50 | 20–30 | 60–65 | 8–12 |
| NC Math 3 | 50 | 20–30 | 60–65 | 8–12 |

### 2.1.3 Item Format

The NCDPI has a long tradition of using selected response items in its summative assessments. Rationales for using selected response items are driven by psychometric, practical and policy considerations. From a psychometric perspective, selected response items such as "multiple-choice" have an extensive reference in educational measurement literature to be a reliable item format in largescale summative standards-based assessments. Also, the validity argument is that scoring of selected response items can be easily automated, highly reliable, and fair for all students. On the practical side, current NCDPI policies are directed towards ensuring state assessments have a minimum effect on instructional time and resources yet are still able to guarantee reliable score for valid uses.

Per G.S. §115C-174.12(a)(4), "all annual assessments of student achievement adopted by the State Board of Education pursuant to G.S. §115C-174.11(c)(1) and (3) and all final exams for courses shall be administered within the final ten (10) instructional days of the school year for

yearlong courses and within the final five (5) instructional days of the semester for semester courses." NCSBE Policy TEST-001 states "LEAs shall report scores resulting from the administration of districtwide and state-mandated tests to students and parents or guardians along with available score interpretation information within thirty (30) days from the generation of the score at the LEA level or receipt of the score and interpretive documentation from the NCDPI."

The NCDPI also has a long tradition of decentralized scoring that ensures students' test scores are readily available following testing for their respective district and student level reporting. The current turnaround time for most state assessments for reporting scores back to schools by their local district is about 2–4 days. The timeliness of test score reporting and teachers using the report for students' proficiency status are significant variables in determining item types to use for state EOG and EOC assessments. Considerations to include new item types, in addition to potential psychometric consequences, are weighed against any potential implication to scoring and reporting delays.

In developing *Edition 5* EOG and EOC assessments, the NCDPI recognized the need to diversify the item pool to incorporate innovative item types that would allow for improved authentic assessment and to include a higher frequency of cognitively complex items to better align assessments with challenging content standard expectations. In addition to traditional multiple-choice (MC) items, *Edition 5* EOG test forms also included open-ended item formats such as numeric entry (NE) or gridded response (GR) items. The current EOC forms also include technology enhanced (TE) item types such as drag-and-drop (DD), text identify (TI) and targeted drop (TD) items. The test development plan is to incorporate TE item types to EOG assessments when these assessments are required online. The final test structures of the base EOG and EOC forms are described in Chapter 4.

### 2.1.4 Calculator Use

Participants at the test specification meetings were also tasked to specify the proportion of EOG and EOC mathematics assessments that should require calculator use. The participants stated that content standard expectations were different across grade levels. For elementary school students, the expectation was to be able to understand foundational mathematics principles, while for higher grades the focus was more on broader mathematical concepts. As a result, they recommended calculator use for 50% of EOG grades 3–5 assessments items and calculator use for about 67% of grades 6–8 and NC Math 1 items. For NC Math 3, the recommendation was to allow calculator use for the entire assessment.

## 2.2 Mode of Test Administration

In 2014, the NCDPI began a steady transition from paper-based test (PBT) administrations to computer-based test (CBT) administrations. This transition has been gradual and systematic

across districts and schools in order for them to be able to provide the necessary technological capacity and comfort for reliable statewide CBT administration. Throughout the transition period, the NCDPI continues to conduct testing in both modes. In 2017–18, all EOG and EOC assessments were available in both modes, and schools had the option to choose how their students were assessed. Beginning from 2018–19 administration, the NCDPI policy requires all EOC assessments to be administered online. PBT forms are only available to students and schools with documented special accommodation needs. Therefore, *Edition 5* of EOC NC Math 1 and NC Math 3 forms were developed as CBT forms and all students were required to participate in CBT mode. The NCDPI, however, continues to provide EOC in PBT mode as an accommodation to students with accessibility issues and schools with an approved technological hardship. The final transition to all CBT mode for *Edition 5* EOG mathematics assessments is planned for the 2020–2021 school year. PBT accommodation forms will always be available for students with a documented need, to ensure accessibility and fairness for all students across the state.

## 2.3 Item Writer and Reviewer Training

The first step of item development is item writer and reviewer training. The main pool of item writers and reviewers for EOG and EOC assessments are classroom teachers from North Carolina. Teachers who want to serve as item writers or reviewers for test development are required to successfully complete in-person or online training courses available through the NC Education website: (https://center.ncsu.edu/ncpd/course/). These courses are designed for anyone interested in learning how to write and/or review assessment items for the North Carolina student testing program based on the North Carolina *Standard Course of Study*.

These courses provide an overview of the test development process and the basic rules and structures of item formats used by the North Carolina State Testing Program. Upon completion of at least one B-level course and at least one C-level course, those interested in item writing and/or reviewing should complete an application for becoming an item writer or reviewer.

The design of these courses is generally linear, requiring the online participant to step through each resource (Web page, PDF, etc.) in a structured sequence. At the end of most topic areas, participants are required to take a short quiz before moving to the next topic area to demonstrate understanding of the presented material. All online quizzes may be taken as many times as needed in order to meet the requirements for moving forward in the course. Once participants have viewed a resource, they are able to return to it for reference at any time. The online item writer training courses can be accessed using the website: https://center.ncsu.edu/ncpd/course/index.php?categoryid=5.

Item writer and reviewer training incorporates the concept of universal design and comprehensible access to the content being measured. Item writers are also required to complete

a mathematics-grade-specific course on the newly adopted content standards. For more information regarding the item writer training and how educators become an item writer or reviewer for the North Carolina student testing program, visit the website: https://center.ncsu.edu/ncpd/course/view.php?id=128.

## 2.4 Item Development Process

The item development process for *Edition 5* began after the NCSBE adopted the new NC*SCOS* for EOC NC Math 1 and NC Math 3 in June 2016 and for EOG grades 3–8 in June 2017. North Carolina test items are written and reviewed by trained North Carolina teachers who serve as item writers. Additionally, the NCDPI's SC&I staff and Test Measurement Specialist (TMS) in partnership with content specialists at North Carolina State University Technical Outreach for Public Schools (NCSU-TOPS) at participate in item development processes. Ultimately, the NCDPI's TMSs serve as final staff reviewers for all EOG and EOC assessment items. Educators with classroom and grade-level content standards experience across the state are recruited, trained, and awarded contracts to write EOG and EOC assessment items. The use of classroom teachers from across the state for item writing is evidence of instructional validity pertaining to how well the test items reflect classroom instruction. Every year a diverse group of North Carolina educators is recruited to write items to replenish EOG and EOC item pools.

Standard 3.2 (AERA, APA, & NCME, 2014) states, "*Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics*" (p. 64). Each new item undergoes a NCDPI iterative 19-step item development and review process. Full details of the 19-step processes are documented in *Appendix 2–D* (p. 1-6).

The first two steps of the item development/review are mostly content focused. Upon receipt of newly written items, Content Specialists at TOPS review the item for accuracy of content, appropriateness of vocabulary (both subject-specific and general), adherence to item writing guidelines and sensitivity and bias concerns. They also verify if items are assigned to the correct attributes:

- a primary standard,
- a secondary standard (when appropriate),
- a DOK rating,
- a targeted achievement level (more recently),
- correct answer/appropriate foil, and
- cited sources of any stimulus material for items (if applicable).

All items that successfully pass initial content evaluation are then sent through an initial production review phase where items needing revisions outside the technical scope of the Content Specialist (such as artwork and graphs) are revised by production staff. Items with stimulus materials are reviewed by copyright staff for copyright concerns and proper citation. At Step 4, each item is independently reviewed by two North Carolina teachers or educators as item reviewers. These reviewers look for any quality issues or bias/sensitivity issues and suggest improvements, when necessary. Any comments or suggested edits to an item are addressed and reconciled by the content and production teams during the next iterative Steps (4–6).

The next steps are designed to address any potential accessibility issues and to ensure items are fair to all students. Exceptional Children (EC)/English Learner (EL)/Visually Impaired (VI) specialist reviews the item for accessibility concerns for EC, EL, and VI students, such as accessibility of graphics for students with or without vision, and also consider accessibility in braille. These reviews address concerns arising from bias or sensitivity issues, such as contexts that might elicit an emotional response and inhibit students' ability to respond or contexts that students may be unfamiliar with for cultural or socioeconomic reasons. Review of reading level of the item is considered along with stem and multiple-choice option (foil) qualities. The item is reviewed to ensure stem is a clear and complete question, foils are straightforward, no repetitive words, and the grammar of the stem agrees with the foils. The review also includes modifying words and making suggestions for bold print and italics or removal, and looking for idioms and two-word verbs that may provide an accessibility issue for EL. Any items with comments that cannot be reconciled is deleted. All other items that either have no issues or had minor suggested reviews that were reconciled are forwarded to a second production edit step (Step 9) and grammar review (Step 10).

At Step 11, a security check is performed on all new items by production staff to make sure no duplicate copy of the item exists in the test development databases. If there is a duplicate copy of the item or a requested revision was not made, then the item is flagged and sent back to Step 8.

In Steps 12–18, items undergo final content and production reviews by content lead (Step 12), SC&I specialist (Step 14), final production and grammar edits (Steps 16 and 17) and a final thorough content review at Step 18 by a Test Measurement Specialist (TMS). The TMS reviews for overall item quality and also check that quality control measures have been followed by reading the comments from all previous reviews and verifying that the comments have been addressed by the content specialists. The TMS has four options at Step 18:

- Approve the item as is; the item proceeds to Step 19 (Item Approved).
- Indicate edits are needed; the item is moved back to Step 15 for review by a content specialist.
- Recommend SC&I to review the item again; the TMS moves the item back to Step 14.
- Delete the item.

The item development and review process are continuous cycles to ensure sufficiency of the item pool. The finalized approved items are then field tested and must undergo a post-field-test round of statistical reviews before they become an operational item.

## 2.5 Field-Test Plan

An embedded field-test design was adopted for the development of *Edition 5* EOC and EOG mathematics items for the North Carolina summative assessments. The main purpose of field testing prior to the development of new operational forms is to gather reliable item level data to evaluate all aspects of item statistical characteristics, accessibility and fairness and to provide baseline statistical targets to assemble pre-equated parallel forms. With the adoption of new content standards, the use of stand-alone field-test administration may have offered a flexible opportunity to gather essential item level data. However, the NCDPI moved to an embedding field-test plan for future item development. The justifications to move away from a traditional stand-alone field-test plan that had been used to develop previous edition of the EOC and EOG assessments were twofold.

First, the embedded field-test design addresses noted shortcomings of a stand-alone field-test by reducing the test burden on students. A stand-alone field-test requires an additional test administration other than operational administration where data shows students are generally less motivated and that usually leads to less reliable item level data.

Second, from a policy perspective, the NCSBE is continuously looking for innovative ways to reduce the impact of testing in public schools. The embedded field-test design offers the opportunity to reduce the testing burden to students and schools. An embedded field-test plan for *Edition 5* allows the NCDPI to get more reliable item level data in a seamless design that offers very little interruption in terms of administrative and instructional impact for students and schools.

### 2.5.1 Field-Test Design for EOG and EOC *Edition 5* Tests

The plan for *Edition 5* was to field-test about 540 items for each EOG grades 3–8 and 300 items for NC Math 1 aligned to new grade-level mathematics content standards. The goal for each item bank was to create between three and four new parallel operational pre-equated test forms. A matrix sampling design that included 10 field-test embedding slots within *Edition 4* forms was used to create sub-versions called “flavors” to embed new field-test items.

The rationale to embed new items aligned to *Edition 5* content standards with *Edition 4* operational tests was that the revised mathematics content standards for *Edition 5* are closely related to and significantly overlap with previously assessed standards. In those instances where new content standards and item types are introduced in *Edition 5*, those items were added at the

end of the operational form. To accomplish this, the number of embedding slots from the *Edition 4* base forms administered in the 2017–18 administration was modified to add five additional item slots at the end of the form. These additional item slots were primarily reserved for new content or new item types. This was done to protect the integrity and validity of students' scores from any effect of new content. Also, to further ensure fairness, this plan was communicated to the field and included as a part of test administration training materials. Table 2.4 shows the matrix embedded field-test plan used to generate item pool for the new EOG and EOC aligned to new content standards.

*Table 2.4 Grades 3–8 Mathematics and NC Math 1 Item Embedding Plan, 2017–18*

| Grades | Base Forms | Items Base Operational Form | Flavors | FT Items/ Flavor | Total FT Items |
|---|---|---|---|---|---|
| 3–5 | 3 | 44 | 12 | 15 | 540 |
| 6–8 | 3 | 50 | 12 | 15 | 540 |
| NC Math 1 | 2 | 50 | 15 | 10 | 300 |

### 2.5.2 Field-Test EOC NC Math 3

One exception to the embedded field-test plan presented above was NC Math 3 where the NCDPI had initially planned to conduct a stand-alone field test for the development of the new EOC assessment. The EOC NC Math 3 is a new summative assessment that was introduced in *Edition 5* to offer an alternate high school assessment for students who took EOC NC Math 1 in middle school. With the adoption of the new content standards, there was no current state assessment that covered the same content aligned to the new NC Math 3.

The plan was to field-test about 500 newly developed NC Math 3 items aligned to the new content standards with the goal to create up to three 50 item parallel operational test forms for 2018–19 administration. To minimize the instructional and administrative impact of stand-alone field-tests, 20 unique mini forms of 25 items each were assembled and randomly spiraled to a sample of students from 212 high schools across North Carolina public schools in 2017–18 administration. The purpose of the stand-alone item field test was to get preliminary baseline item statistics to help inform form assembly.

# CHAPTER 3 ITEM ANALYSIS

This chapter summarizes procedures and criteria the NCDPI uses to analyze and evaluate the statistical and psychometric characteristic of newly developed test items. Item analysis serves as the final quantitative process for item review and to establish grade level operational item pool for form development. Standard 4.10 (AERA, APA, & NCME, 2014) states, "*When a test developer evaluates the psychometric properties of items, the model used for that purpose should be documented. … The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented*" (p.89).

Most large-scale assessment programs rely on two broad measurement models – Classical Test Theory (CTT) and Item Response Theory (IRT) – to screen and evaluate items for calibration, form assembly and scoring. Another important procedure in traditional item analysis is the statistical evaluation of DIF used to evaluate fairness and potential item bias across major groups. The NCDPI psychometric specifications for item review use statistical criteria from both CTT and IRT measurement models in addition to Mantel-Haenszel DIF statistics. These procedures and their various criteria used for item screening and analysis are explained and described in the following sections.

## 3.1 Statistical Item Flagging Criteria

All field-test items are classified into one of three NCDPI item flagging categories (Keep, Reserve, and Weak) with the goal to rank items in the final pool based on overall statistical quality during form assembly. These specifications are routinely updated to continuously ensure that the highest quality items are selected for EOG and EOC assessments.

- Keep: These are items with good statistical properties from CTT, IRT and DIF statistical procedures used for item analysis. Items flagged as "Keep" are first choice from the item pool during form assembly. Their main statistical properties are within the established NCDPI ranges considered as optimal items.
- Reserve: These are items with at least one major statistical parameter that is barely outside the range to be considered as reserve items. These items are only included in the final form assembly pool if they are needed to meet content or statistical specifications of the operational form. When any item flagged as "Reserve" from field tests is placed on a new form it must undergo additional content review to ensure the content is accurate.
- Weak: These are items with at least one major statistical parameter being significantly outside the range to be considered as optional items based on field-test analysis. When complete field-test data are available, these items are generally not included in the item pool used for form assembly. The only exception to this rule is when exceptional

circumstances cause field-test data to be incomplete or unreliable. In such situations, thorough vetting is required from the content experts and psychometricians.

## 3.2 CTT Based Item Analysis

Item level CTT statistics like percent correct (p-value), item-to-total correlations (biserial correlation), and distractor analysis are used as a first step to screen item quality following field tests. In accordance with the NCDPI policy, whenever possible, all items must first be field-tested prior to placing them on operational form. After items are field tested, the first step involves conducting a series of CTT analysis to determine if these items meet the minimum psychometric requirements to be considered for further evaluation. The NCDPI uses a custom-developed SAS® Macro item analysis routine with a combination of procedures to process student response data from field tests and compute CTT statistics.

- Item p-value summarizes the proportion of examinees from a given sample answering the item correctly and is used as an indicator of preliminary item difficulty. Valid p-value for dichotomously scored items ranges between 0 and 1, where values close to 0 indicate extremely difficult items (few students selected the correct response) and values close to 1 indicate easier items (almost all students answered correctly).
- The biserial correlation coefficient is a special case of the Pearson correlation coefficient and describes the relationship between a dichotomous variable and a continuous variable. The biserial coefficient provides evidence of the strength of the relationship between the item and the unidimensional construct being measured. Theoretical range for biserial coefficient is –1 to 1. Negative biserial correlation generally indicates the item might be measuring a separate unintended construct. *Table 3.1* shows the CTT-based item flagging criteria.

*Table 3. 1 CTT Item Flagging Criteria*

| CTT Statistics | Flagging Criteria |
|---|---|
| $0.150 \leq$ p-value $\leq 0.850$ | Keep |
| $0.100 \leq$ p-value $\leq 0.149$ or $0.851 \leq$ p-value $\leq 0.900$ | Reserve |
| p-value $\leq 0.099$ and p-value $\geq 0.901$ | Weak |
| biserial $\geq 0.250$ | Keep |
| $0.150 \leq$ biserial $\leq 0.249$ | Reserve |
| biserial $< 0.150$ | Weak |

Grades 3-8 and NC Math 1 CTT Descriptive Summaries

The CTT descriptive summary from field-test in 2017–18 for EOG and EOC mathematics items are shown in *Table 3.2*. This table shows the combined CTT summary statistics across both paper- and computer-based test modes by grade.

*Table 3. 2 CTT Descriptive Summary of Grades 3–8 and NC Math 1 Field-Test Item Pool, Spring 2018*

| Grade | CTT Flag | Total Items | P–value | | | | Biserial Correlation | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | SD | Min | Max | Mean | SD | Min | Max |
| 3 | Keep | 376 | 0.66 | 0.15 | 0.25 | 0.85 | 0.47 | 0.08 | 0.20 | 0.65 |
| | Reserve | 99 | 0.77 | 0.16 | 0.14 | 0.90 | 0.36 | 0.07 | 0.16 | 0.49 |
| | Weak | 65 | 0.80 | 0.20 | 0.16 | 0.97 | 0.27 | 0.09 | -0.01 | 0.42 |
| 4 | Keep | 413 | 0.65 | 0.14 | 0.19 | 0.85 | 0.48 | 0.08 | 0.25 | 0.65 |
| | Reserve | 80 | 0.75 | 0.15 | 0.21 | 0.90 | 0.35 | 0.07 | 0.13 | 0.49 |
| | Weak | 47 | 0.84 | 0.14 | 0.33 | 0.98 | 0.27 | 0.09 | -0.10 | 0.41 |
| 5 | Keep | 413 | 0.64 | 0.15 | 0.19 | 0.85 | 0.47 | 0.08 | 0.21 | 0.68 |
| | Reserve | 84 | 0.70 | 0.20 | 0.14 | 0.90 | 0.36 | 0.08 | 0.13 | 0.48 |
| | Weak | 43 | 0.80 | 0.19 | 0.03 | 0.94 | 0.29 | 0.07 | 0.09 | 0.42 |
| 6 | Keep | 439 | 0.51 | 0.16 | 0.16 | 0.85 | 0.49 | 0.10 | 0.20 | 0.68 |
| | Reserve | 56 | 0.47 | 0.25 | 0.11 | 0.88 | 0.34 | 0.10 | 0.13 | 0.55 |
| | Weak | 43 | 0.34 | 0.32 | 0.00 | 0.96 | 0.26 | 0.12 | 0.05 | 0.47 |
| 7 | Keep | 443 | 0.49 | 0.16 | 0.15 | 0.85 | 0.50 | 0.10 | 0.20 | 0.72 |
| | Reserve | 49 | 0.46 | 0.24 | 0.11 | 0.89 | 0.32 | 0.11 | 0.12 | 0.54 |
| | Weak | 48 | 0.25 | 0.22 | 0.00 | 0.84 | 0.18 | 0.16 | -0.12 | 0.47 |
| 8 | Keep | 309 | 0.42 | 0.16 | 0.15 | 0.85 | 0.40 | 0.09 | 0.19 | 0.63 |
| | Reserve | 112 | 0.40 | 0.22 | 0.10 | 0.90 | 0.29 | 0.10 | 0.12 | 0.55 |
| | Weak | 119 | 0.23 | 0.21 | 0.00 | 0.92 | 0.16 | 0.12 | -0.16 | 0.42 |
| NC Math 1 | Keep | 237 | 0.46 | 0.14 | 0.16 | 0.85 | 0.47 | 0.12 | 0.20 | 0.69 |
| | Reserve | 43 | 0.35 | 0.20 | 0.10 | 0.74 | 0.31 | 0.12 | 0.12 | 0.56 |
| | Weak | 20 | 0.25 | 0.19 | 0.04 | 0.67 | 0.23 | 0.15 | 0.07 | 0.46 |

The initial CTT results indicate about 70% or more of items field-tested for EOG grades 3–7 and NC Math 1 were classified as meeting the NCDPI optimal standards of "Keep". For items field tested in EOG grade 8, only about 57% of the item bank was classified as optimal for form development. The significant reduction in the pool quality for grade 8 mathematics was attributed to the significant change in the 8$^{th}$ grade mathematics population. Due to policy changes, the top third performing students dual-enrolled in 8$^{th}$ grade mathematics and NC Math 1

were required to only participate in NC Math 1 assessment. This resulted in a significant shift of the ability distribution for 8th grade students evident by the low sample p-values for grade 8 compared to other EOG grades. Moreover, the p-value and biserial ranges show the item pool had enough range of item difficulty and biserial correlation for high quality operational form assembly.

NC Math 3 CTT Descriptive Summaries and Time Data

The NC Math 3 field test was initially designed as a stand-alone administration. Preliminary item analysis from Fall administration indicated students may not have taken the test seriously because they knew that the assessment was a field test that would not have direct testing consequences. Results for the 2017–18 stand-alone field test shown in *Table 3.3* and *Table 3.4* support this concern. First, the item pool with "Keep" and "Reserve" categories have an average p-value of 0.33 across 20 forms indicating the forms consisted of generally difficult items. Second, the average number correct score across the 20 forms with 25-item from each form was seven (7). Timing data also shown students spent on average 44 minutes to complete a 25-item test. The percentile raw scores and time indicates 95% of the students completed the 25 items test in 78 minutes or lower and scored 13-point or less. Third, Cronbach alpha, a measure of internal consistency, across the forms did not exceed 0.7 indicating that the statistics generated from the stand-alone field test were less reliable.

*Table 3. 3 CTT Descriptive Summary of NC Math 3 Field–Test Item Pool, Spring 2018*

| Grade | CTT Flag | Total Items | P–value | | | | Biserial Correlation | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | SD | Min | Max | Mean | SD | Min | Max |
| NC Math 3 | Keep | 192 | 0.33 | 0.13 | 0.15 | 0.76 | 0.39 | 0.09 | 0.2 | 0.56 |
| | Reserve | 95 | 0.33 | 0.19 | 0.1 | 0.88 | 0.35 | 0.09 | 0.12 | 0.54 |
| | Weak | 213 | 0.19 | 0.17 | 0 | 0.75 | 0.25 | 0.1 | -0.01 | 0.50 |

*Table 3. 4 NC Math 3 FT Raw Score and Timing Data (N=37,791)*

| Category | Statistics | | Percentile | | | |
|---|---|---|---|---|---|---|
| | Mean | Std Dev | 25th | 75th | 95th | 99th |
| Raw Score | 7 | 3 | 4 | 8 | 13 | 17 |
| Time (Minutes) | 44 | 39 | 30 | 54 | 78 | 93 |

These NC Math 3 results suggested that the item statistics generated from the stand-alone field test did not demonstrate enough reliability to interpret and generalize these item statistics as reflective of NC Math 3 performance. As a direct consequence, fewer items could be categorized

as “Keep” and “Reserve”, resulting in an insufficient number of items to develop new forms. The NCDPI decided to abandon the stand-alone field test and switch to an operational field test design for NC Math 3. New NC Math 3 forms were assembled primarily relying on content expertise with the plan to statistical adjust and balance the forms after operational administration in 2017-18.

## 3.3 IRT-Based Item Analysis

IRT offers a more robust approach to item analysis compared to CTT. CTT uses relatively weak assumptions based on the relationship between true score and error. A limitation of CTT is that it focuses on properties of a given test and results are often group dependent (Hambleton, 2000, Yen & Fitzpatrick, 2006). The IRT-based item parameters, on the other hand, are assumed to be sample independent, and item performance is related to the estimate of students’ latent trait called “ability” measured by the test (Anastasi & Urbina, 1997). IRT offers many features to the testing program that may be difficult to get with CTT mostly because IRT defines a scale for the underlying latent variable that is measured by test items. This aspect of IRT means comparable scores may be computed for examinees who did not take the same test questions without intermediate equating steps (Thissen & Orlando, 2001).

IRT offers a series of statistical models used to describe the probabilistic relationship between examinee responses given the item characteristics. All IRT models assume this relationship to be monotonic, meaning that as the trait level increases, the probability of a correct response also increases. According to Yen & Fitzpatrick (2006, p. 112), all IRT models can be classified by the type of item data, like number of dimensions, they use to describe examinee and item characteristics, and the number and type of item characteristics they describe relative to each dimension.

Since EOG and EOC mathematics items are binary scored (only two possible outcomes: correct or incorrect), the NCDPI uses two main IRT unidimensional models to describe items’ characteristics for item calibration, to develop item banks for form building and for scaling. The NCDPI uses the three-parameter logistic (3PL) unidimensional model for multiple-choice and technology enhanced items and the two-parameter logistic (2PL) model for numeric or gridded response items. These models make three major assumptions:

- unidimentionality – that there is one dominant latent trait being measured which in this case is mathematics and that this trait is the driving force for the responses observed for each item in the measure,
- local independence – that responses to different questions on the test are conditionally independent given the underlying ability level, and
- sample invariance – that item parameter estimates are invariant to any group of subjects who have answered the item.

The mathematical function for the 3PL IRT model is:

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-Da_i(\theta - b_i)]}$$

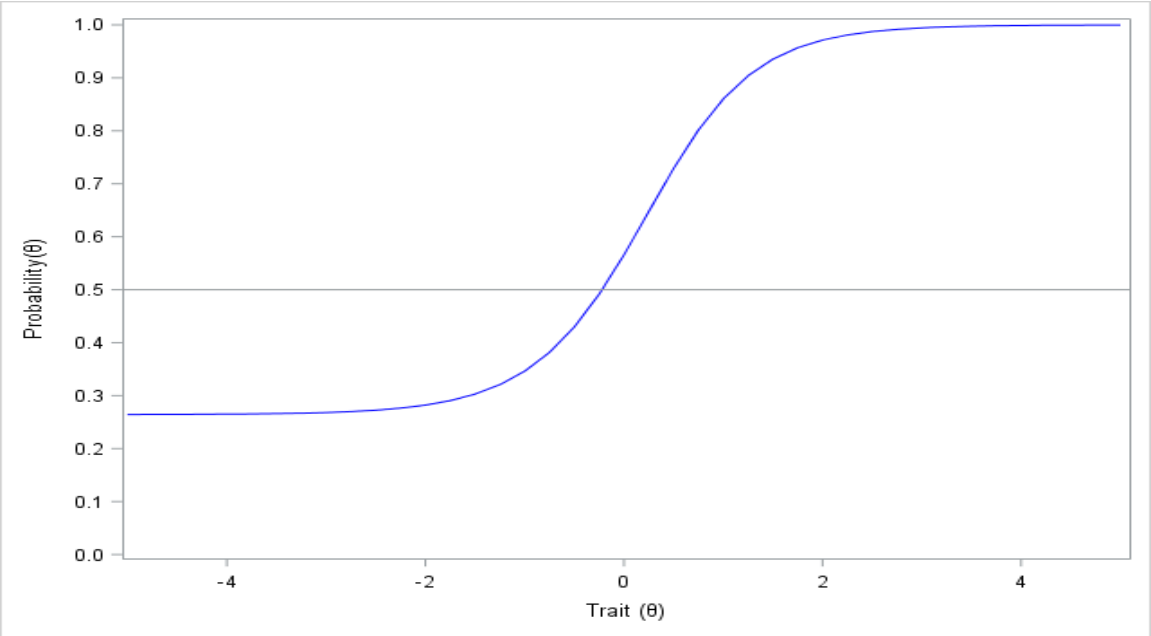
where $P_i(\theta)$ is the probability that a randomly chosen examinee of given ability answers item *i* correctly (this is an S-shaped curve with values between 0 and 1 over the ability scale), $a_i$ is the slope or the discrimination power of the item, $b_i$ is the threshold or difficulty parameter of an item, $c_i$ is the lower asymptote or pseudo-chance level parameter, and *D* is a scaling factor of 1.7. The major difference between a 3PL model and a 2PL model is that the 2PL model does not directly account for a chance-score parameter. The 2PL model can be expressed as a special case of the 3PL model with $c_i = 0$ (see Equation below). For numeric response items, students are required to provide their answers rather than to select an answer from several choices, and the chance to get an item right by guessing is almost zero. The mathematical function for the 2PL model is:

$$P_i(\theta) = \frac{1}{1 + \exp[-Da_i(\theta - b_i)]}$$

Once parameters for items are calibrated, a probabilistic relationship between each item along the ability continuum of -∞ to +∞ can be represented with a nonlinear monotonically increasing curve called an item characteristic curve (ICC) or trace line (Hambleton & Swaminathan, 1985). The ICCs represent a summary figure, which can be used to evaluate the statistical properties for each item. Inferences about difficulty, discrimination, and chance score for each item can be made for examinees at different ability levels along the ability continuum. Such inferences are critical during form assembly when items are selected to match a statistical target.

An example of the ICC is shown in *Figure 3.1*. The vertical axis represents the probability of a correct response and the horizontal axis represents the underlying latent ability scale. If the ICC is towards the left on the ability scale (less than 0), that will indicate the item is relatively easier for most examinees. The ICC in *Figure 3.1* shows an item with about medium difficulty in which an examinee with average ability will have about a 50% probability to answer the item correctly. The slope describes the discriminatory power of the item that indicates the level of measurement precision attributed to that item conditional on the ability scale. The lower asymptote of the curve is the 3PL model adjustment for what is usually referenced in IRT literature as an adjustment for guessing (c parameter). In the case of the 2PL for numeric response items the c parameter is fixed to zero.

*Figure 3. 1 Graphical Representation of Item Characteristic Curve or Trace Line*




For final item quality, the NCDPI uses IRT parameters flagging criteria displayed in *Table 3.5* to classify field-test items into one of the three categories. As stated in Section 3.1, the final item pool for form development is made of items flagged as psychometric "Keep" and "Reserve". During form assembly, priority is given to items with a "Keep" status.

*Table 3. 5 IRT Items Flagging Criteria*

| IRT Parameters | Flagging Criteria |
|---|---|
| Threshold Value (b) | |
| $-2.500 \leq b \leq 2.500$ | Keep |
| $-3.000 \leq b \leq -2.501$ or $2.501 \leq b \leq 3.000$ | Reserve |
| $b \leq -3.001$ or $b \geq 3.001$ | Weak |
| Slope Value (a) | |
| $1.190 \leq a$ | Keep |
| $0.850 \leq a \leq 1.189$ | Reserve |
| $a < 0.848$ | Weak |
| Asymptote or Guessing Value (c) | |
| $\leq 0.350$ | Keep |
| $0.351 \leq c \leq 0.450$ | Reserve |
| $> 0.451$ | Weak |

## 3.4 IRT Parameter Estimation

IRT parameters of the embedded field-test items are estimated by calibrating student responses using IRTPRO® software (Cai, Thissen, & du Toit, 2011) with the Bayesian prior for the discrimination (a) parameter set to Lognormal distribution (0, 1) and pseudo-guessing parameters (c ) set to Beta distribution (5, 15). The Bayesian prior ensures appropriate parameter estimates of pseudo-guessing; that is, scores for 4-option MC items are accounted for in the 3PL model. IRT calibration phase is designed to serves two main purposes:

- Form Development: The first main purpose of calibration is to develop an item bank with items of known statistical properties that are on the same latent IRT grade-level ability scale. These calibrated items expressed on the same IRT scale offers the NCDPI the flexibility to build multiple alternate forms without the need for traditional post equating.
- Scaling: The second purpose of calibration is to establish final IRT parameters for field-test items that are later used to create an IRT raw-to-scale table for alternate new forms before they are operationally administered. This is the essence of the NCDPI decentralized and immediate scoring for EOG and EOC assessments.

The NCDPI uses two main methods of calibration based on data collection design attributed to modes of testing: a single random group calibration for field-test items administered in predominantly in one test mode and a concurrent calibration with a mode DIF sweep step for field-test items administered in both modes.

### 3.4.1 Single Group Calibration

During each EOG and EOC test administration window, multiple parallel (alternate) forms are administered in every grade. Subsets of field-test items are embedded with operational items on base forms using a matrix sampling design shown in *Figure 3.2* to create form flavors to embed and collect student responses. All form and flavor combinations are randomly spiraled within schools at the student level across the state. This ensures base forms with field-test items are randomly administered to a representative sample of students at the grade level including students with disabilities (SWD), Rural, and economically disadvantaged student (EDS) (see *Table 3.6*). The NCDPI uses a single group design to calibrate IRT parameters for field-test items associated with each base form that were administered either majority (70% or more) in one mode or if no items flagged for mode DIF. In 2017–18, field-test items embedded in EOG mathematics grades 3 and 4 (majority paper based), and grade 8, NC Math 1 and NC Math 3 (computer based) were calibrated using a single group design. IRT field-test item parameters separately calibrated across different base forms are assumed to be on the common IRT latent ability scale. The rationale is that base forms spiraled and randomly administered to representative samples of grade level population are equivalent. *Table 3.6* and *Table 3.7* show demographic distribution of the samples in grades 3–5 and grades 6–8 and NC Math 1. It shows that the sample sizes, gender, and ethnic distribution across forms are very similar within each

grade. The grades 3–7 forms were administered in both modes, and the grade 8 and NC Math 1 forms were administered in CBT mode with paper forms for accommodation only. For the grade 8 and NC Math 1 tests, items from CBT forms only will be selected in the new operational base forms as new forms are going to be administered in CBT mode only. NC Math 3 information are not included as the test was refiled tested in 2018–19.

*Figure 3. 2 Matrix Data collection For Embedded Field–Test Design*

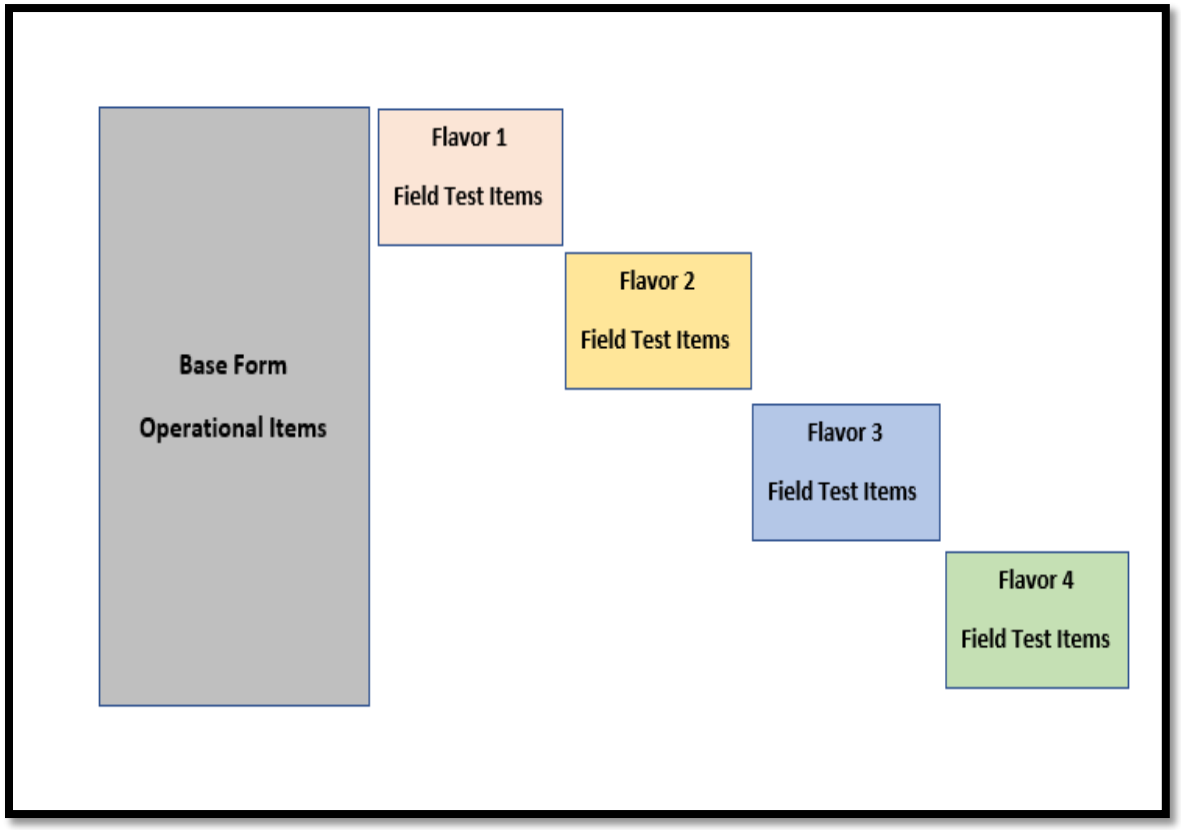


*Table 3. 6 Grades 3-5 Demographic distribution of the Field–Test Sample, 2017–18 Edition 4 Base Forms*

| Grade | Mode | Form | Total | Ethnicity | | | | Gender | | Other | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | W | B | H | Other | M | F | SWD | Rural | EDS |
| 3 | Both | A | 35,013 | 48.4 | 26.5 | 15.9 | 9.3 | 49.6 | 50.4 | 10.0 | 43.9 | 44.9 |
| | Both | B | 34,953 | 48.2 | 26.9 | 15.4 | 9.6 | 50.0 | 50.0 | 9.8 | 44.0 | 45.3 |
| | Both | C | 34,892 | 49.1 | 26.3 | 15.2 | 9.4 | 50.1 | 49.9 | 9.3 | 44.3 | 44.9 |
| | | All | 104,858 | 48.5 | 26.6 | 15.5 | 9.4 | 49.9 | 50.1 | 9.7 | 44.1 | 45.0 |
| 4 | Both | A | 31,037 | 48.9 | 26.5 | 15.2 | 9.4 | 49.7 | 50.3 | 9.3 | 40.7 | 43.5 |
| | Both | B | 38,257 | 48.5 | 25.8 | 16.2 | 9.5 | 49.7 | 50.3 | 10.9 | 45.5 | 44.7 |
| | Both | C | 37,416 | 48.9 | 25.8 | 15.9 | 9.3 | 50.0 | 50.0 | 10.4 | 46.2 | 44.7 |
| | | All | 106,710 | 48.7 | 26.0 | 15.8 | 9.4 | 49.8 | 50.2 | 10.3 | 44.3 | 44.3 |
| 5 | Both | A | 37,288 | 48.1 | 24.5 | 17.8 | 9.6 | 50.5 | 49.5 | 11.6 | 43.7 | 43.7 |
| | Both | B | 37,304 | 47.6 | 25.1 | 18.1 | 9.1 | 50.3 | 49.7 | 11.6 | 43.6 | 43.9 |
| | Both | C | 36,652 | 48.1 | 24.8 | 17.9 | 9.3 | 50.0 | 50.0 | 11.0 | 44.2 | 44.1 |
| | | All | 111,244 | 47.9 | 24.8 | 18.0 | 9.3 | 50.3 | 49.7 | 11.4 | 43.9 | 43.9 |

Note: W=White, B=Black, H=Hispanic, M=Male, F=Female

*Table 3. 7 Grades 6–8 and NC Math 1 Demographic distribution of the Field–Test Sample, 2017–18 Edition 4 Base Forms*

| Grade /Course | Mode | Form | Total | Ethnicity | | | | Gender | | Other | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | W | B | H | Other | M | F | SWD | Rural | EDS |
| 6 | Both | A | 37,260 | 48.1 | 25.0 | 17.5 | 9.4 | 50.2 | 49.8 | 13.6 | 44.6 | 45.2 |
| | Both | B | 37,977 | 48.8 | 24.1 | 18.1 | 8.9 | 50.3 | 49.7 | 12.6 | 44.5 | 44.9 |
| | Both | C | 36,420 | 47.5 | 25.1 | 18.4 | 9.1 | 51.0 | 49.0 | 13.3 | 44.0 | 46.0 |
| | | All | 111,657 | 48.2 | 24.7 | 18.0 | 9.1 | 50.5 | 49.5 | 13.2 | 44.4 | 45.4 |
| 7 | Both | A | 36,580 | 49.3 | 24.6 | 17.1 | 9.0 | 50.6 | 49.4 | 12.5 | 45.0 | 43.1 |
| | Both | B | 36,852 | 49.2 | 24.7 | 17.5 | 8.6 | 50.5 | 49.5 | 13.0 | 45.1 | 43.7 |
| | Both | C | 36,743 | 49.5 | 24.3 | 17.9 | 8.3 | 50.5 | 49.5 | 12.3 | 45.1 | 43.2 |
| | | All | 110,175 | 49.3 | 24.5 | 17.5 | 8.7 | 50.5 | 49.5 | 12.6 | 45.0 | 43.3 |
| 8 | CBT | A | 26,139 | 45.1 | 28.4 | 18.7 | 7.9 | 51.9 | 48.1 | 16.0 | 46.8 | 49.5 |
| | CBT | B | 27,096 | 45.1 | 28.1 | 18.8 | 8.0 | 52.0 | 48.0 | 16.5 | 47.2 | 49.4 |
| | CBT | C | 24,385 | 46.8 | 27.6 | 18.1 | 7.6 | 51.7 | 48.3 | 15.1 | 50.5 | 49.1 |
| | | All | 77,620 | 45.6 | 28.1 | 18.5 | 7.8 | 51.9 | 48.1 | 15.9 | 48.1 | 49.3 |
| NC Math 1 | CBT | All | 65,648 | 50.1 | 23.7 | 16.8 | 9.5 | 51.2 | 48.8 | 12.1 | 46.5 | 42.2 |

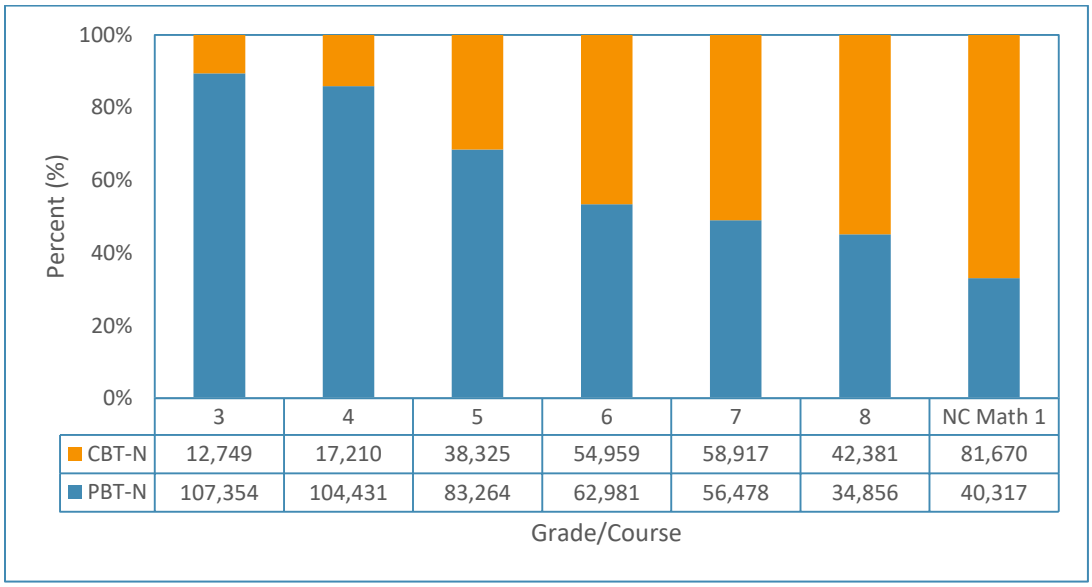
Note: W=White, B=Black, H=Hispanic, M=Male, F=Female

### 3.4.2 Concurrent Calibration with Mode DIF Sweep

Beginning with 2017–18, the NCDPI offers all assessments in both PBT and CBT modes. The plan is to gradually transition to a CBT mode for all EOG and EOC tests. *Figure 3.3* shows the proportion of students by mode with student count in the inset table for EOG and EOC mathematics in 2017–18. Notice that the proportion of students who took the test on CBT mode increased as grade level increased.

As a part of the NCDPI's effort to ensure score comparability across test administration mode, a concurrent groups calibration with a DIF sweep step was used to calibrate field-test items on forms with a significant participation in both modes. The use of concurrent groups calibration with DIF sweep allows the nonequivalent ability distribution of students across modes to be accounted for in the final IRT parameters. District and schools self-select when they are ready to transition to CBT mode. This calibration method is designed to disentangle the ability differences between the two groups while properly modeling any DIF due to mode effect of groups taking the same items between PBT and CBT modes.

*Figure 3. 3 Proportion of Students by Mode, 2017–18*




| | 3 | 4 | 5 | 6 | 7 | 8 | NC Math 1 |
|---|---|---|---|---|---|---|---|
| CBT-N | 12,749 | 17,210 | 38,325 | 54,959 | 58,917 | 42,381 | 81,670 |
| PBT-N | 107,354 | 104,431 | 83,264 | 62,981 | 56,478 | 34,856 | 40,317 |

Concurrent calibration with mode DIF sweep is done in a two-step calibration process in IRTPRO® software with an evaluation phase between calibration steps to identify possible candidates DIF items. The first step is an exploratory calibration phase with the goal to identify candidate DIF items. During the initial phase, the latent ability of the reference group is centered to have a normal (0, 1) distribution and the focal group's ability distribution is freely estimated.

Using estimate of group ability distribution, separate IRT parameters are estimated for all items in both groups. Once IRT parameters are estimated for each group, they are then evaluated to identify candidate DIF items. In 2015, after a series of empirical analyses and discussions, the NCDPI recommended to its Technical Advisory Committee to update their DIF analysis procedure from the chi-square hypothesis test performed in IRTPRO® to a more robust residual-based effect size methodology. The rationale was that the chi-square procedure was flagging too many false positives due to the larger sample sizes common in North Carolina statewide grade level assessment items (usually the sample size exceeds 2,000 observations per item.)

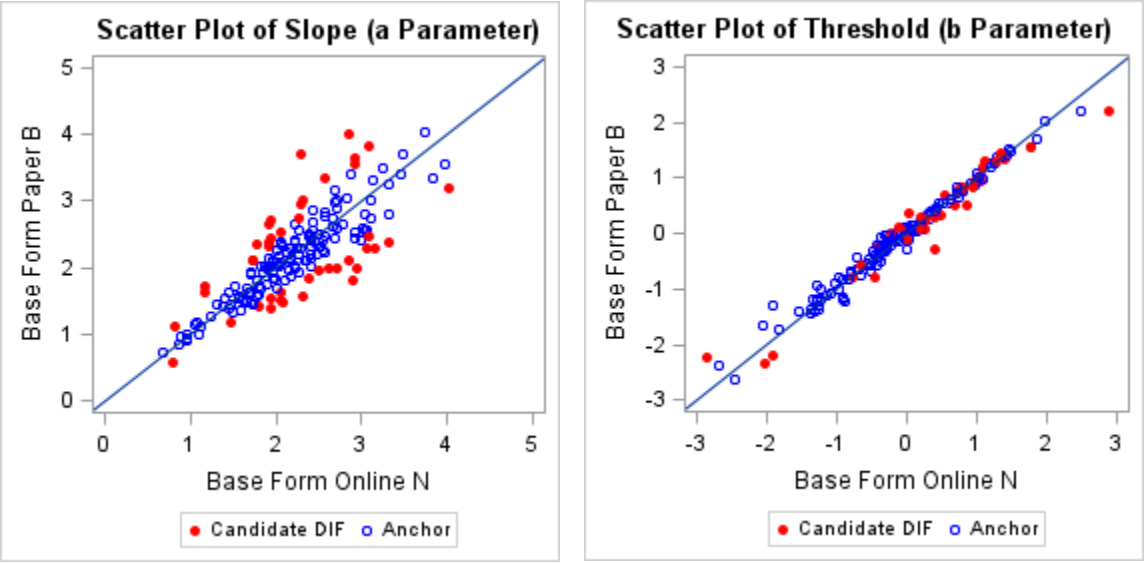
A flagging criterion of 0.20 was set to identify candidate DIF items is either the following conditions were met:

$$\left|\mathbf{1} - \frac{\boldsymbol{a_{1j}}}{\boldsymbol{a_{2j}}}\right| \text{ or } \left|\boldsymbol{b_{1j}} - \boldsymbol{b_{2j}}\right| > 0.20 \qquad (3\text{-}1)$$

where $a_{1j}$ and $a_{2j}$ are the slope parameters and $b_{1j}$ and $b_{2j}$ are the threshold parameters for the focal and reference groups respectively. The rationale for choosing a very stringent criterion was that a type 1 error was a bigger threat to score comparability and the NCDPI wanted to make sure all potential DIF items were flagged.

*Figure 3.4* shows a visual example of residual DIF analysis performed for EOG grade 6 mathematics items in spring 2018. The slope and threshold parameters for items administered in paper and computer mode are plotted using scatter diagrams. Items along the diagonal are shown to display no DIF and these items will be used as anchor in step 2, whereas those items shown in the off diagonal are hypothesized as candidate DIF items. Separate parameters will be estimated for those items in each group during the next step.

*Figure 3. 4 Example of Residual DIF Analysis*



The second step is set up as a confirmatory calibration step where all items with residual effect size less than 0.20 are considered as anchor items and used to estimate examinees' abilities between groups, and items above the criterion are placed as candidate DIF items. During calibration, IRTPRO® will estimate joint item parameters for the anchor items, and separate item parameters are estimated for candidate DIF items conditioned on group ability estimated using anchor items. This then places both sets of item parameters on a common IRT scale. If no items are flagged as candidate DIF the final item parameter values are concurrently calibrated without taking the mode of delivery into account.

## 3.5 IRT Calibration Summary From 2017–18

*Table 3.8* and *Table 3.9* show descriptive statistics of IRT parameters for items from the spring 2018 embedded field-test plan. The items flagged as "Keep" and "Reserve" are considered as acceptable and made up the final item pool for form assembly. Since almost even proportion of students took grades 6 and 7 forms in two modes, the NCDPI performed concurrent calibration with DIF sweep for paper and computer modes. Single group calibration was performed for grades 3, 4, 5, 8, and NC Math 1 as majority of students took paper forms in grades 3 and 4, no items were flagged for mode DIF in grade 5, and grade 8 and NC Math 1 were CBT tests with paper option for accommodations only.

*Table 3.8 Grades 3–5 Descriptive Statistics of IRT Parameters for the EOG and EOC Math Field-Test, Spring 2018*

| Grade | Flags | N* | % | Slope(a) | | | | Threshold(b) | | | | Asymptote(g) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Mean | SD | Min | Max | Mean | SD | Min | Max | N* | Mean | SD | Min | Max |
| 3 | Keep | 376 | 70 | 1.99 | 0.50 | 1.19 | 3.77 | –0.28 | 0.64 | –1.42 | 1.52 | 376 | 0.17 | 0.06 | 0.04 | 0.35 |
| | Reserve | 99 | 18 | 1.52 | 0.46 | 0.86 | 2.74 | –0.87 | 0.99 | –2.39 | 2.51 | 99 | 0.22 | 0.10 | 0.08 | 0.45 |
| | Weak | 64 | 12 | 1.43 | 0.59 | 0.51 | 3.64 | –1.34 | 1.41 | –3.22 | 2.69 | 64 | 0.23 | 0.09 | 0.13 | 0.57 |
| 4 | Keep | 413 | 76 | 2.05 | 0.48 | 1.19 | 4.27 | –0.21 | 0.62 | –1.38 | 1.63 | 413 | 0.18 | 0.07 | 0.02 | 0.35 |
| | Reserve | 80 | 15 | 1.58 | 0.55 | 0.86 | 3.18 | –0.66 | 0.94 | –1.87 | 1.65 | 80 | 0.25 | 0.09 | 0.08 | 0.41 |
| | Weak | 46 | 9 | 1.54 | 0.66 | 0.21 | 3.73 | –1.36 | 1.14 | –3.16 | 1.18 | 46 | 0.29 | 0.12 | 0.15 | 0.59 |
| 5 | Keep | 413 | 76 | 1.95 | 0.46 | 1.20 | 3.80 | –0.22 | 0.64 | –1.39 | 1.54 | 289 | 0.21 | 0.07 | 0.05 | 0.35 |
| | Reserve | 84 | 16 | 1.59 | 0.57 | 0.87 | 3.14 | –0.47 | 1.07 | –1.86 | 2.03 | 66 | 0.25 | 0.10 | 0.06 | 0.43 |
| | Weak | 43 | 8 | 1.51 | 0.68 | 0.43 | 3.00 | –0.98 | 1.29 | –2.55 | 3.14 | 41 | 0.26 | 0.13 | 0.12 | 0.67 |

*Note: 1) No Asymptote value for TE items, 2) Items with negative biserial flagged from CTT were excluded from IRT calibration.

*Table 3.9 Grades 6–8 and NC Math 1 Descriptive Statistics of IRT Parameters for the EOG and EOC Math Field-Test, Spring 2018*

| Grade | Flags | N* | % | Slope(a) | | | | Threshold(b) | | | | Asymptote(g) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Mean | SD | Min | Max | Mean | SD | Min | Max | N* | Mean | SD | Min | Max |
| 6 (PBT) | Keep | 446 | 83 | 2.28 | 0.60 | 1.25 | 5.51 | 0.25 | 0.60 | –1.27 | 1.52 | 336 | 0.21 | 0.06 | 0.05 | 0.35 |
| | Reserve | 54 | 10 | 1.93 | 0.91 | 0.91 | 4.35 | 0.48 | 1.08 | –1.80 | 2.14 | 35 | 0.27 | 0.09 | 0.11 | 0.41 |
| | Weak | 37 | 7 | 2.09 | 1.44 | 0.35 | 8.38 | 0.93 | 1.76 | –2.49 | 4.44 | 22 | 0.27 | 0.09 | 0.13 | 0.52 |
| 6 (CBT) | Keep | 439 | 82 | 2.27 | 0.58 | 1.21 | 4.80 | 0.27 | 0.62 | –1.78 | 1.76 | 329 | 0.20 | 0.07 | 0.01 | 0.35 |
| | Reserve | 56 | 10 | 1.84 | 0.84 | 0.91 | 4.35 | 0.63 | 1.03 | –1.86 | 2.16 | 39 | 0.27 | 0.11 | 0.06 | 0.41 |
| | Weak | 38 | 8 | 1.98 | 0.93 | 0.35 | 4.98 | 1.00 | 1.68 | –2.49 | 4.78 | 21 | 0.25 | 0.13 | 0.01 | 0.50 |
| 7 (PBT) | Keep | 447 | 83 | 2.44 | 0.69 | 1.20 | 6.05 | 0.38 | 0.54 | –1.12 | 1.74 | 328 | 0.21 | 0.06 | 0.06 | 0.35 |
| | Reserve | 45 | 8 | 2.08 | 1.05 | 0.95 | 4.33 | 0.69 | 1.11 | –1.79 | 2.20 | 38 | 0.28 | 0.08 | 0.10 | 0.44 |
| | Weak | 36 | 9 | 2.51 | 1.79 | 0.13 | 8.68 | 1.47 | 1.22 | –2.17 | 4.04 | 18 | 0.25 | 0.09 | 0.10 | 0.49 |
| 7 (CBT) | Keep | 443 | 82 | 2.43 | 0.70 | 1.20 | 5.28 | 0.38 | 0.56 | –1.12 | 2.21 | 326 | 0.20 | 0.07 | 0.02 | 0.35 |
| | Reserve | 49 | 9 | 2.01 | 1.16 | 0.85 | 5.63 | 0.67 | 1.06 | –1.82 | 1.86 | 38 | 0.26 | 0.09 | 0.11 | 0.44 |
| | Weak | 32 | 9 | 2.45 | 1.23 | 0.52 | 4.92 | 1.61 | 1.20 | –2.01 | 4.77 | 17 | 0.28 | 0.12 | 0.10 | 0.52 |
| 8 | Keep | 309 | 57 | 1.89 | 0.51 | 1.19 | 3.83 | 0.70 | 0.66 | –1.31 | 2.02 | 244 | 0.20 | 0.06 | 0.06 | 0.35 |
| | Reserve | 112 | 21 | 1.44 | 0.52 | 0.86 | 3.22 | 1.00 | 1.07 | –1.79 | 2.59 | 86 | 0.22 | 0.07 | 0.06 | 0.44 |
| | Weak | 91 | 22 | 1.41 | 0.82 | 0.45 | 4.02 | 2.12 | 1.32 | –2.05 | 5.66 | 53 | 0.23 | 0.06 | 0.14 | 0.42 |
| NC Math 1 | Keep | 237 | 79 | 2.12 | 0.53 | 1.24 | 4.45 | 0.50 | 0.56 | –1.16 | 1.72 | 174 | 0.21 | 0.08 | 0.02 | 0.35 |
| | Reserve | 43 | 14 | 1.83 | 0.82 | 0.86 | 3.76 | 1.23 | 0.75 | –0.63 | 2.83 | 29 | 0.25 | 0.12 | 0.03 | 0.44 |
| | Weak | 20 | 7 | 2.25 | 1.08 | 0.39 | 3.89 | 1.87 | 0.58 | 0.10 | 2.78 | 12 | 0.27 | 0.14 | 0.12 | 0.61 |

*Note: 1) No Asymptote value for TE items, 2) Items with negative biserial flagged from CTT were excluded from IRT calibration.

## 3.6 Bias and Sensitivity DIF Analysis

As the developers of the NC assessments, it is the responsibility of the NCDPI to examine all assessment items for possible sources of bias. The Standard 3.3 (AERA, APA, & NCME, 2014) states "*Those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test"* (p. 64). Statistical DIF procedure sometimes referred to as bias analysis examine the degree to which students of various groups (e.g., males versus females) perform differently on an item. It is expected that students with the same ability should have similar probability for answering items correctly, regardless of background characteristics. An item is considered as exhibiting DIF when students from different socioeconomic or demographic background with similar estimated knowledge and skill on the overall construct being tested perform substantially different on the same item (AERA, APA, & NCME, 2014). It is important to remember that the

presence or absence of true bias is a qualitative decision based on the content of the item and the curriculum context within which it appears.

The NCDPI utilizes Mantel-Haenszel (MH) DIF statistics with ETS Delta classification codes for flagging candidate DIF for multiple-choice items (Camilli & Sheppard, 1994) to quantitatively identify suspect items for further qualitative bias and sensitivity scrutiny by expert panels. The MH chi-square statistic tests the alternative hypothesis that a linear association exists between the row variable (score on the item) and the column variable (group membership). The MH odds ratio (*Table 3.10*) is computed using the Cochran-Mantel-Haenszel (CMH) option in PROC FREQ Procedure in SAS® for j matched groups.

$$\boldsymbol{\alpha_{MH}} = \frac{\sum_j \boldsymbol{A_j D_j / T_j}}{\sum_j \boldsymbol{B_j C_j / T_j}} \qquad (3\text{-}2)$$

Where, in j 2X2 tables, $A_j$ and $C_j$ are the numbers of examinees in the reference and focal groups, respectively, who answer the item correctly; and $B_j$ and $D_j$ are the numbers of examinees in the reference and focal groups, respectively, who answered the item incorrectly.

*Table 3.10 MH Odds Ratio Calculation*

| Group | Score on Studied Item | | Total |
|---|---|---|---|
| | 1 | 0 | |
| Reference (R) | Aj | Bj | nRj |
| Focal (F) | Cj | Dj | nFj |
| Total | m1j | m0j | Tj |

Transforming the odds ratio by the natural logarithm provides the DIF measure, such that:

$$\beta_{MH} = \log_e(\alpha_{MH}) \qquad (3\text{-}3)$$

The ETS classification scheme first requires rescaling the MH value by a factor of -2.35 providing the Delta (D) statistic as follows:

$$|D| = -2.35\beta_{MH} \qquad (3\text{-}4)$$

Items are then classified based on their Delta statistic into three categories:

- ‘A’ items are not significantly different from 0 using |D|<1.0. No substantial difference on item performance between the two groups is found for items with A+ or A– classifications.
- ‘B’ items significantly different from 0 and D is not significantly greater than 1.0 or |D|<1.5. An item with a B+ rating marginally favors the focal group (Females, African

Americans, Hispanics, or Rural students). An item with a B– rating on the other hand marginally favors the reference group (favors Males, Whites, or Non-rural students).

- 'C' items have D significantly greater than 1.0 and |D|≥1.5. An item with a C+ rating favors the focal group (Females, African Americans, or Hispanics, Rural, Economically Disadvantaged Students or EDS). Item with a C– rating disfavors the focal group (favors Males, Whites, Rural, EDS).

All field-test items are quantitatively evaluated for DIF based on five main demographic and socioeconomic groupings:

- Demographic:
  - Males (reference) and Females (focal)
  - Whites (reference) and Blacks (focal)
  - Whites (reference) and Hispanics (focal)
- Socioeconomic:
  - Urban schools (reference) and Rural schools (focal)
  - Not Economic Disadvantaged (reference) and Economic Disadvantaged (focal)

*Table 3.11* shows field-test EOG and EOC item pool DIF summary by flagging classification from 2017–18 administration. The NCDPI rule is to exclude all items from the final pool that are flagged as DIF "C" suggesting significant DIF. These items are either retired or sent back to Step 1 of the item writing process to undergo significant revisions and a new round of field tests and analysis. Items flagged as DIF "B" are kept in the pool but will need to undergo further bias review by a panel if selected to be placed on a form. The panel decides whether the items are free of implied bias.

*Table 3.11 Mantel-Haenszel Delta DIF Summary for the EOG and EOC Mathematics Field-Test, Spring 2018*

| Grade | Gender | | | Ethnic | | | | | | Urban/Rural | | | Economically Disadvantage | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | White/Black | | | White/Hispanic | | | | | | | | |
| | A | B | C | A | B | C | A | B | C | A | B | C | A | B | C |
| 3 | 518 | 20 | 2 | 492 | 43 | 5 | 512 | 27 | 1 | 537 | 3 | 0 | 536 | 4 | 0 |
| 4 | 506 | 31 | 3 | 506 | 29 | 5 | 511 | 27 | 2 | 539 | 1 | 0 | 538 | 2 | 0 |
| 5 | 517 | 22 | 1 | 494 | 34 | 12 | 511 | 26 | 3 | 538 | 2 | 0 | 538 | 2 | 0 |
| 6 | 486 | 43 | 8 | 474 | 49 | 14 | 479 | 47 | 11 | 520 | 11 | 6 | 526 | 11 | 0 |
| 7 | 478 | 47 | 13 | 463 | 48 | 27 | 494 | 36 | 7 | 527 | 11 | 0 | 518 | 19 | 1 |
| 8 | 469 | 48 | 18 | 460 | 50 | 22 | 484 | 28 | 20 | 518 | 11 | 6 | 521 | 9 | 5 |
| NC Math 1 | 276 | 22 | 2 | 275 | 15 | 10 | 285 | 13 | 2 | 296 | 3 | 1 | 289 | 9 | 2 |

A=A+/A-, B=B+/B-, C=C+/C-

At the conclusion of item analysis based on field-test data, the final item pool for form assembly is made up of items with a psychometric classification of "Keep" or "Reserve" and a DIF flag of "A" or "B". All items with field-test psychometric flag of "Weak" or DIF classification of "C" are excluded from consideration during form assembly.

# CHAPTER 4 OPERATIONAL FORM ASSEMBLY, ANALYSIS, AND REVIEW

AERA, APA, & NCME (2014) states, "*The test developer is responsible for documenting that the items selected for the test meet the requirements of the test specifications. In particular, the set of items selected for a new test form or an item pool for an adaptive test must meet both content and psychometric specifications*" (p. 82). To adhere to the standard, Chapter 4 documents the iterative IRT-based automated form assembly processes used to create parallel forms. This chapter also summaries all the quality and content review steps the NCDPI uses to finalize new operational base forms from the field-test pool. In all, the NCDPI has instituted a 26-step iterative form building and review process documented in *Appendix 2–D* (p.12–18).

## 4.1 IRT Automated Form Assembly

The first step in form assembly requires the initial selection of items to match the test blueprint discussed in Chapter 2 and a statistical target for new forms. The NCDPI uses a two-phase form assembly process to select and review forms. In Phase 1, an automated form assembly custom SAS® macro uses sampling procedures to optimally select items from the pool to match test blueprint and statistical specifications to recommend the most appropriate form. The automated form assembly macro relies on two main IRT based statistics: test characteristic curve (TCC) and test information function (TIF).

### Test Characteristics Curves

In IRT, TCCs are essential for form assembly and scaling. A TCC is generally 'S-shaped' figure with flatter ends that show the expected summed score as a function of theta (θj) (Thissen, Nelson, Rosa, & Mcleod, 2001). Mathematically, the TCC function is the sum of ICCs for all items on the test (see equation 4-1). During form assembly, items with known parameters were selected from the item bank based on a predetermined blueprint to match a target or base TCC. According to Thissen, Nelson, Rosa, & Mcleod (2001, p.158), TCCs for parallel forms plotted on the same graph is an easy way to examine the relation of summed score with theta.

$$TCC = \sum_{i=1}^{n} p_i(\theta) \text{ , i=1……..n} \qquad (4\text{-}1)$$

Where $p_i(\theta)$ is the probability of answering item(s) correctly and provides ICCs across ability (θ) ranges.

## Test Information Function (TIF) and Conditional Standard Error (CSE)

The concept of reliability ($\rho$) is central in CTT when evaluating the overall consistency of scores over replications and it is generally reported in terms of standard error, which is defined as $s_x\sqrt{1-\rho}$ . Under the CTT framework, reliability and standard error are sample based and, regardless of where examinees are on the score scale, the amount of measurement error is uniform. Thissen and Orlando (2001, p. 117) highlighted that, in IRT, standard errors usually vary for different response patterns for the same test. Examinees with different response patterns or at different points on the theta scale will show variations in the amount of measurement precision. No single number characterizes the amount of precision of an entire test on an IRT base scale. Instead, the pattern of precision over the range conditional on ability may be inferred using the Test Information Function (TIF) (see equating 4-2) and the inverse of TIF is interpreted as conditional standard error. The concept of measurement precision as reported by TIF or CSE has been well documented in IRT literature.

$$I(\theta) = \sum_{i=1}^{n} \frac{[P\prime_i(\theta)]^2}{P_i(\theta)Q_i(\theta)} \qquad (4\text{-}2)$$

For more information see Hambleton & Swaminathan (1985), and Thissen & Orlando (2001). Some features of TIF as noted in Hambleton & Swaminathan (1985, p. 104) are:

- TIF is defined for a set of test items at each point on the ability scale.
- The amount of information is influenced by the quality and number of test items.
- $I(\theta)$ is the test information function, $P_i(\theta)$ is obtained by evaluating the item characteristic curve model at $\theta$, $P_i'(\theta) = \delta P/\delta\theta$, and $Q_i(\theta) = (1- P_i(\theta))$
- The steeper the slope, the greater the information.
- The smaller the item variance, the greater the information.
- $I(\theta)$ does not depend upon the particular combination of test items. The contribution of each test item is independent of the other items in the test.
- The amount of information provided by a set of test items at an ability level is inversely related to the error associated with ability estimates at the ability level.

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \qquad (4\text{-}3)$$

In Phase 2 of assembly, IRT parameters and the recommended form from the macro are output into an interactive excel worksheets where any further review to the form base on content and or production feedback are manually handled. All revisions made to the form are done with respect to the blueprint and statistical targets.

## 4.2 Statistical Targets of New Forms

*Edition 5* EOG and EOC mathematics assessments were developed to realign the new assessments to the revised *Edition 4* content standards. As documented in chapters 1 and 2 of this report, changes from *Edition 4* to *Edition 5* were substantial for some parts of the assessed standards. The statistical properties of the old base forms were used as a baseline in specifying the targets for new forms. The TCCs of the old base forms were used as starting targets for the new base forms and these were adjusted to enhance measurement precision along the critical areas of the scale. If the existing base form indicated the test was more precise for examinees with below-average estimated ability, the new reference was adjusted to make sure there was enough measurement precision at the middle of the distribution. The goal was to maximize measurement precision around the achievement level cuts at Not Proficient/Level 3, Level 3/Level 4, and Level 4/Level 5. These points are the most critical reporting decisions made on the EOG and EOC scales.

Since the NCDPI no longer plans to report EOG and EOC mathematic on a developmental scale, the statistical targets are determined independently for each grade based on the content complexity of grade level content standards. The final statistical targets for base forms across grade are not intended to imply a vertical scale.

The ideal TCCs for the parallel forms would be perfectly overlay to each other. The TCCs of the newly developed parallel forms across grades 3–8 and NC Math 1, based on the IRT item parameters estimated from spring 2018 embedded field tests, are shown in *Figure 4.1* through *Figure 4.7*. Since NC Math 3 forms were constructed based on 2018-19 operational parameters, the TCCs based on field-test are not shown. The TIFs and conditional standard error of measurements (CSEMs) are shown in *Appendix 4–A*. The *Figure 4.1* through *Figure 4.7* show that the TCCs for parallel forms are closely overlap, with small variations in some grades, along the ability scale. These small variations in TCCs from parallel forms are acceptable and could be accounted for during scaling using summed score methodology, where separate raw-to-scale tables will ensure all examinees with the same expected ability have the same expected outcome regardless of the test form. All this is because item parameters used to generate these forms are on the same IRT scale, which makes it possible to compare performance of students taking completely different forms without the need to conduct additional traditional equating.

*Figure 4. 1 TCCs Based on Field-Test Item Parameters, Grade 3*

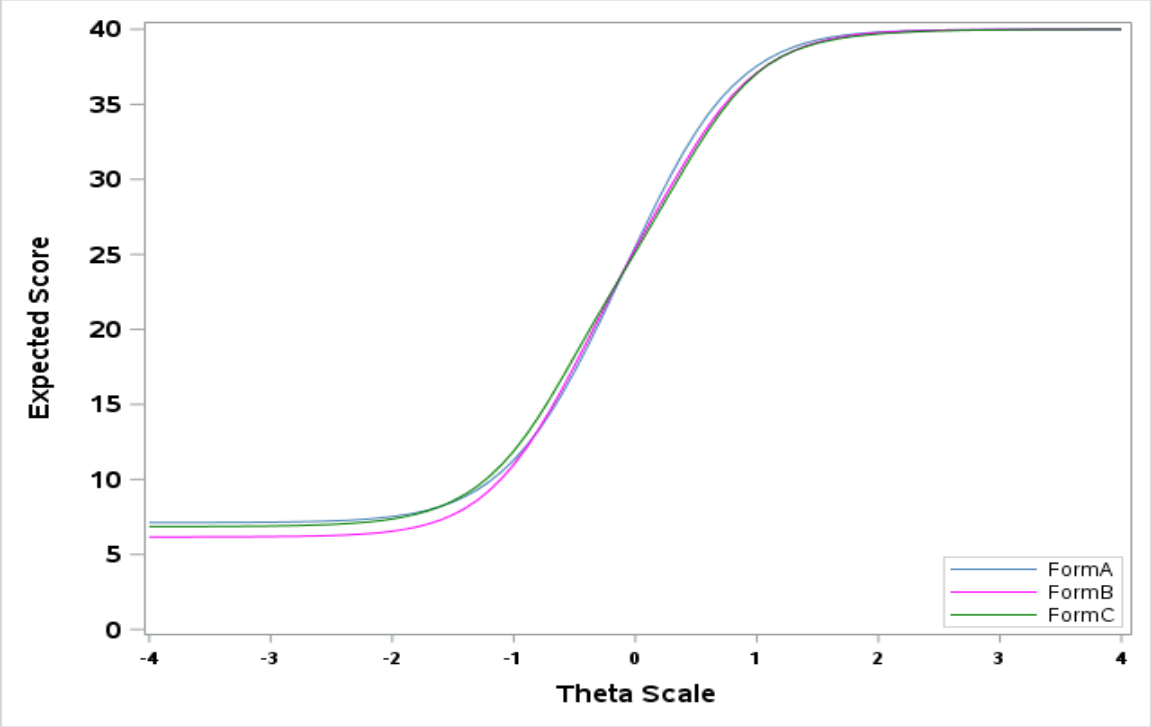



*Figure 4. 2 TCCs Based on Field-Test Item Parameters, Grade 4*

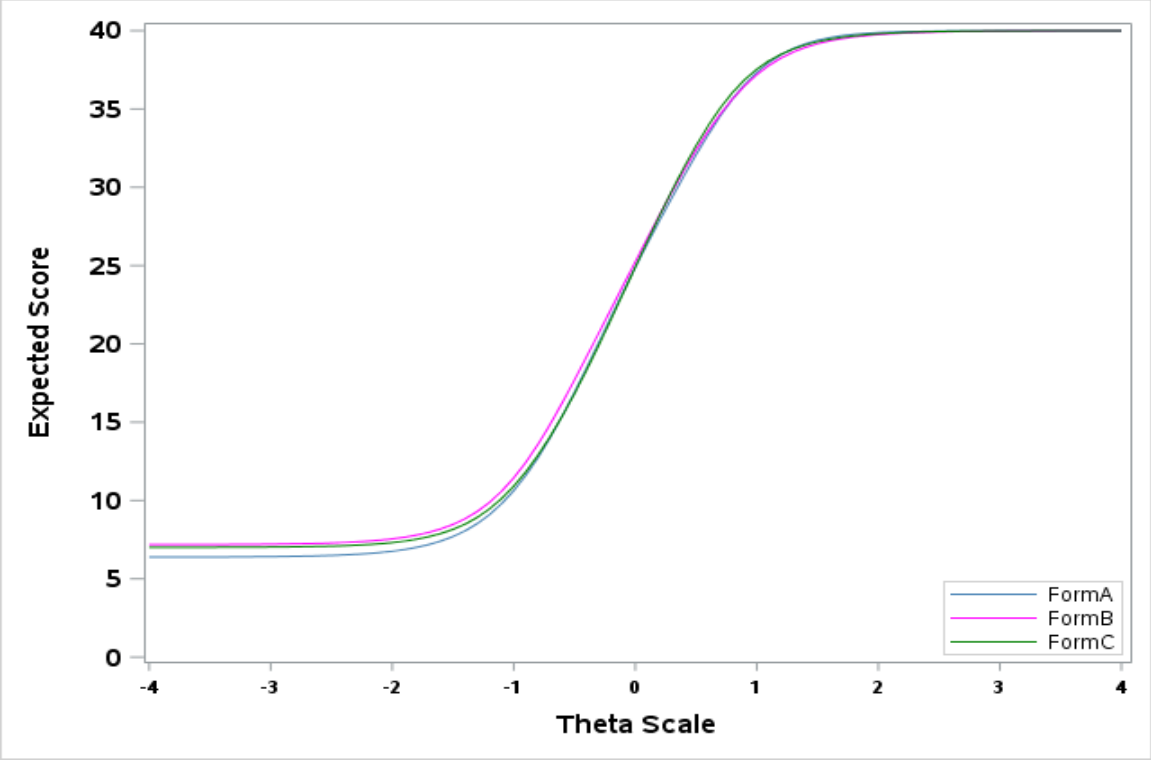


*Figure 4. 3 TCCs Based on Field-Test Item Parameters, Grade 5*

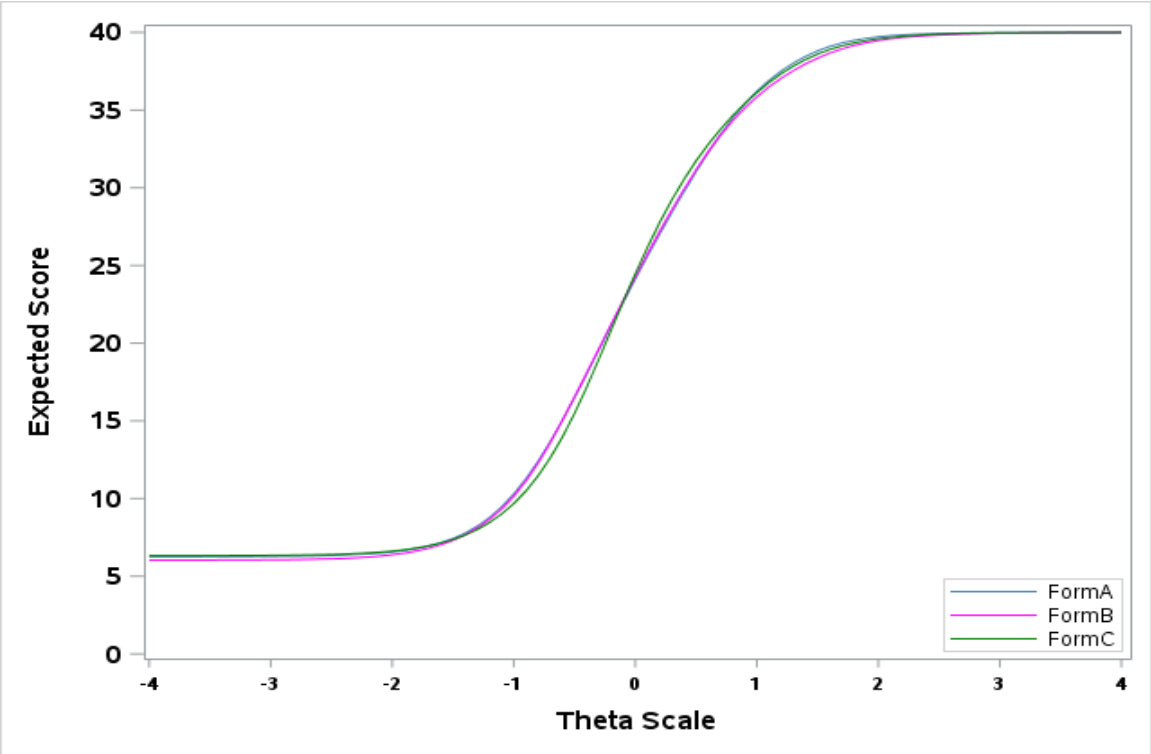



*Figure 4. 4 TCCs Based on Field-Test Item Parameters, Grade 6*

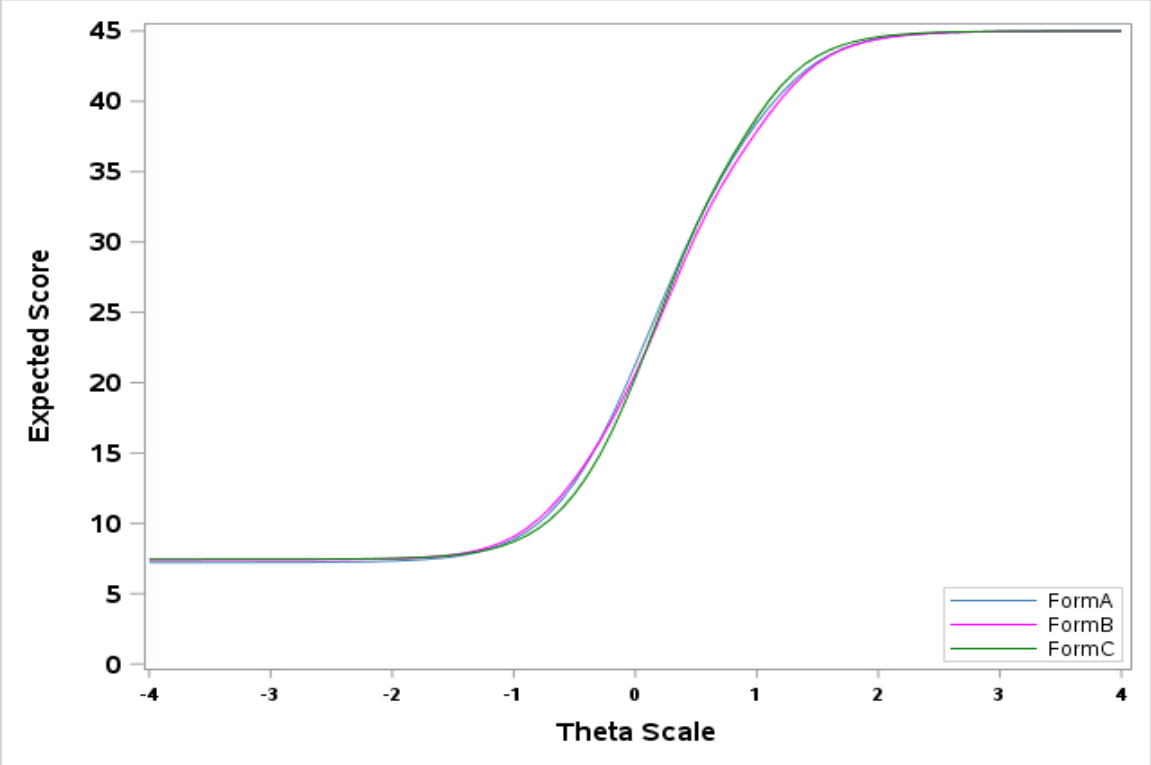


*Figure 4. 5 TCCs Based on Field-Test Item Parameters, Grade 7*

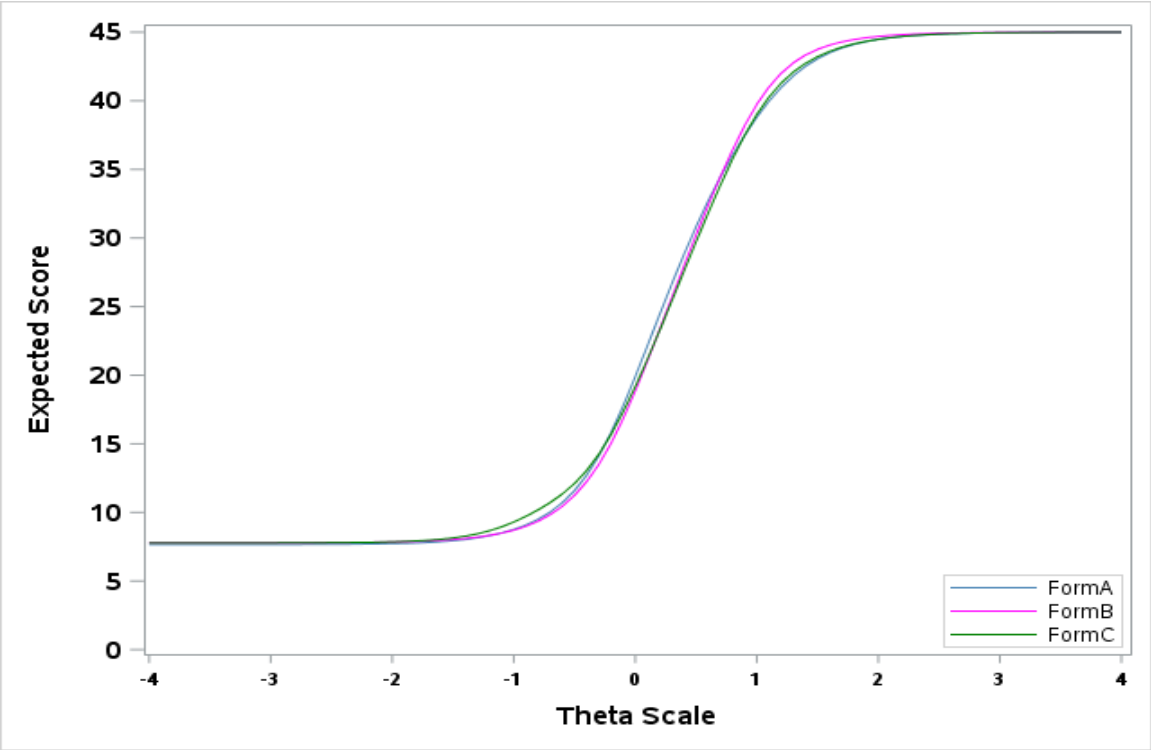



*Figure 4. 6 TCCs Based on Field-Test Item Parameters, Grade 8*

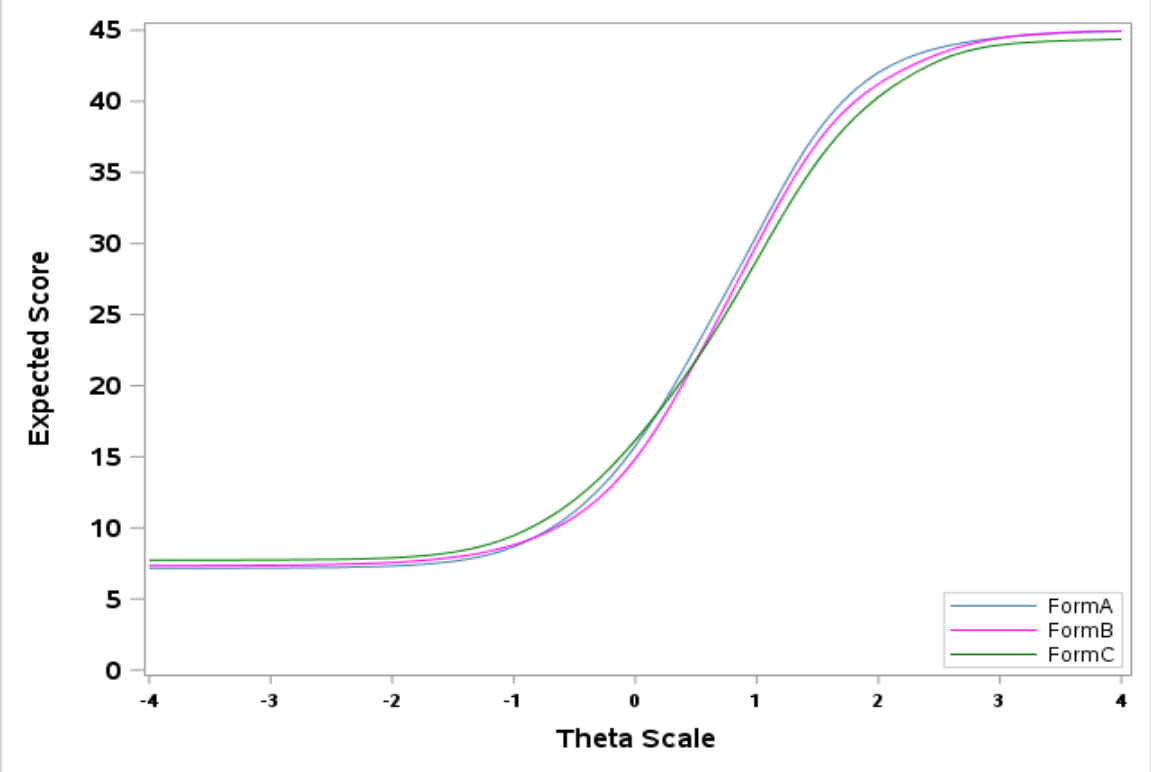
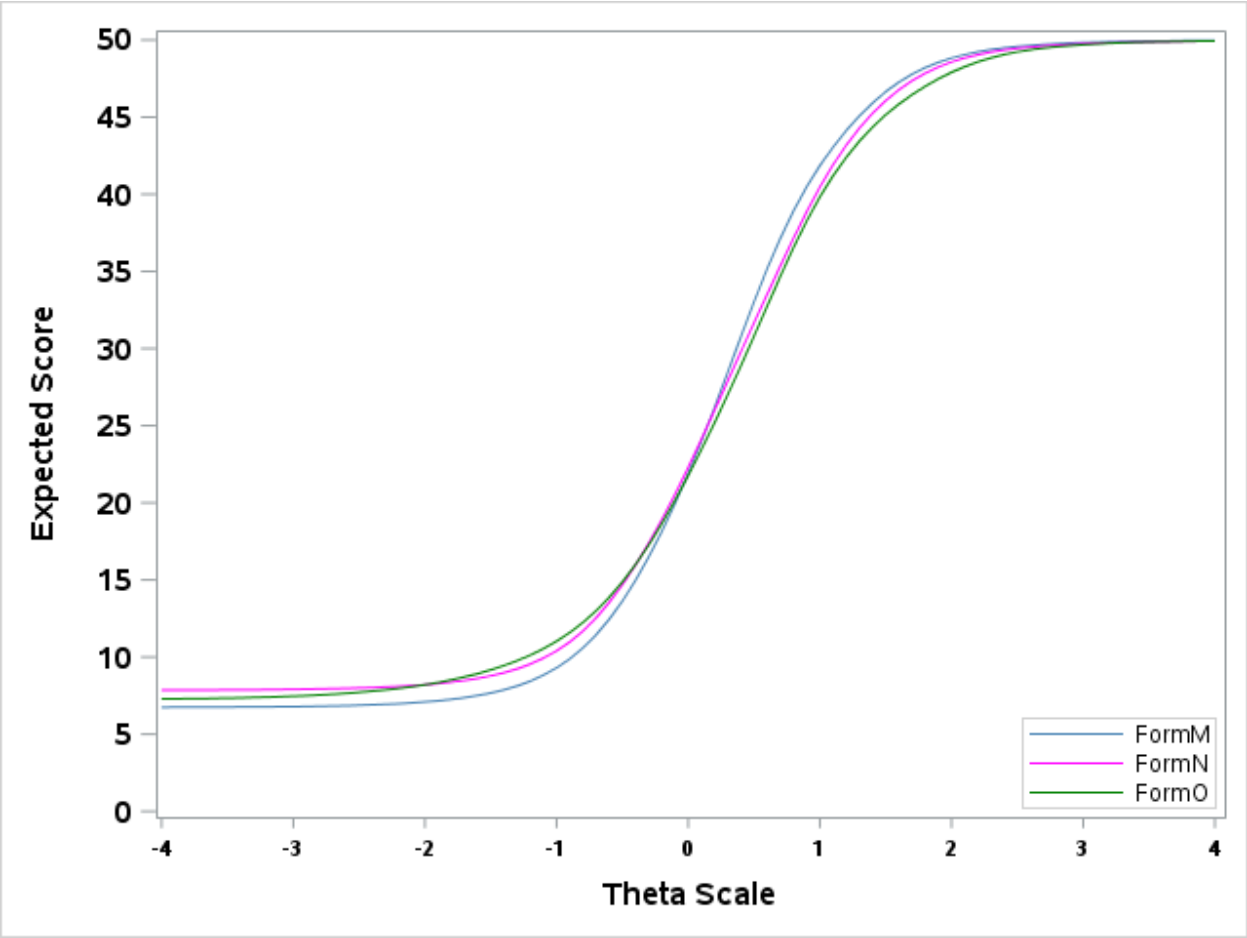


*Figure 4. 7 TCCs Based on Field-Test Item Parameters, NC Math 1*




## 4.3 Form Review

After the initial assembly and statistical review (Step 1) of the form development process is complete, the form then undergoes a series of iterative review steps which can be summarized into content and production reviews (*Appendix 2–D*). At each critical review step, if there is a recommendation to replace an item the form is sent back to Step 1 for final consideration. If there is a replacement item from the bank that maintains the blueprint and statistical properties of the form, then a quick swap is made, and the form sent back through the review process.

### 4.3.1. Content Reviews

The main content review steps of the 26-step processes (*Appendix 2–D*) are Steps 3–7, Steps 11–14, Steps 16–18 and Step 21. These content review steps are done at various stages by a NCSU-TOPS content specialist, an NCDPI TMS, and an external outside content reviewer. The ultimate

objective of content reviewers is to make sure all items selected on forms are appropriate and aligned to grade-level content. They also check to make sure items on forms do not cue and are not repetitive (like overemphasis on a subtopic, e.g. if all area problems in one form were isosceles triangles). Criteria for evaluating each test form included the following:

- The content of the test forms reflects the goals and objectives of the North Carolina *Standard Course of Study* for the subject (content validity).
- The content of test forms reflects the goals and objectives as taught in North Carolina schools (instructional validity).
- Items are clearly and concisely written, and the vocabulary is appropriate to the target age level (universal design).
- Content standards of the test forms are balanced and. items do not cue other items on a form.
- All selected response items have one and only one best correct response choice. The distractors should appear plausible for someone who has not achieved mastery of the representative objective (one best answer).

The outside content reviewers are instructed to complete a mock administration of a test form and to provide written comments and feedback next to each item. Each reviewer independently documents his or her opinion as to how well the tests met the five criteria listed above. These comments are further reviewed by the NCSU-TOPS and the NCDPI content with the goal to address concerns ranging from a simple grammatical fix to replacing the item in the form.

At Step 21, a content manager reviews comments/suggestions and makes any necessary revisions to embedded items. The manager checks the form for overall quality and reviews the form comment history to ensure all comments have been addressed. After reviewing the form, the Content Manager may choose one of the following options:

- Approve the form and send it to Step 23 (Audio Approval) if the form will be recorded online.
- Approve the form and send it to Step 24 (Compare) if the form will be unrecorded or on paper only.
- Send the form to Step 20 (Psychometrician) if there are suggested revisions to operational items for the Psychometrician to consider.
- Send the form to Step 22 (Production Edits) for revisions to artwork, graphs, or reading selections.

### 4.3.2. Production Reviews

Production and grammar reviews of text, artwork or graphs, and copyright are continuously monitored and checked in several steps (Steps 2, 9, 10, 17, 19, 20, 23 and 24). Most of the

production steps are used for item revisions such as minor grammatical edits, formatting and revision of artwork or figures on items. All proposed revisions to base form items must be approved by the psychometrician who will determine if proposed edits are significant to the point that it might affect the interpretation of field-test statistics. If it is ruled the proposed revision will invalidate the item field-test statistics, then a recommendation is made to replace the item.

At Step 23, a content specialist reviews the audio for each item and either approves the audio or indicates it needs correction. After all items' audio has been approved, the form is sent to Step 24 for PDF/Online Check for forms that will be administered in both computer and paper modes.

At Step 24, PDF/Online Check, production staff exports the form as a document and formats the document per formatting guidelines. The form is placed in a folder with a signoff sheet where:

- First, two editors review the form for formatting concerns as well as any grammatical issues, and
- Second, a content specialist reviews the form for content and evaluates any comments and or suggestions from Editing reviews.

If there are any edits to execute in the online test development system, the Content Specialist indicates with each item what edits are approved and sends the form back to Step 21. Any suggestions that are rejected should be noted in the form comments. Any suggested edits to operational items that Content Staff feel warrant consideration are directed to the TMS and Psychometrician for consideration.

After final review of the online version, the computer-based forms are exported from the TDS application into the NCTest platform. In this stage, a series of quality checks are performed by NCSU-TOPS staff to ensure all the specified interactions between items and the NCTest platform are fully functional across the different end users' approved devices. The NSCU–TOPS and the NCDPI test development have instituted a four-phase quality check protocol. This protocol focuses on issues ranging from technical and network comparability aspects to accessibility aspects such as verifying that high contrast, large font and read aloud files are working properly. Summary description of the four-phase quality checks performed on all computer-based forms are:

- Phase 1 – forms are assigned to demo students. Each form is assigned to a demo student and forms are chosen to display the accessibility/accommodation features of large font and high contrast along with test read aloud.
- Phase 2 – NCSU–TOPS employees conduct quality checks using the demo students to ensure the correctness of the forms and the items themselves. The Editing/Production groups are notified if issues arose with respect to the content, whereas the NCTest group is notified if there are any issues with the apps or supporting resources.

- Phase 3 – operations staff and TMSs at the NCDPI listen to all audio recordings, review all test features (highlighting, strike out answers, reset, etc.) and view all items. The accommodated forms are viewed with presentation settings of large font or high contrast. The check of all forms is performed on the secure browser, Chrome app for Chromebooks and/or the iPad app for iPads to ensure items functioned and displayed appropriately. Findings are then reported to NCSU–TOPS for corrections and all corrections are monitored and verified as complete by the NCDPI.
- Phase 4 – forms are checked to ensure the data is being recorded accurately and the scoring keys for the items on each form are accurate. The NCDPI accountability division IT group validates the data collected at this stage.

All forms that are also offered online are sent to Step 25 and the form is operationally locked to prevent any further revisions. This is to ensure that the published versions of the form, items and selections are preserved electronically.

## 4.4 Bias and Sensitivity DIF Reviews

When constructing test forms, it is important to know the extent to which items perform differentially for various groups of students. The first step was flagging items for DIF. The second step was convening a fairness review panel to examine potential DIF flagged items selected on operational test forms. Standard 3.6 (AERA, APA, & NCME, 2014) states, "*Where credible evidence indicates that test scores may differ in meaning for relevant subgroups in the intended examinee population, test developers and/or users are responsible for examining the evidence for validity of score interpretations for intended uses for individuals from those subgroups. What constitutes a significant difference in subgroup scores and what actions are taken in response to such differences may be defined by applicable laws*" (p. 65).

This specific standard place responsibility on test publishers to examine all sources of possible construct-irrelevant variance. In order to satisfy this standard, the TOPS convened the Fairness Review panel to review all items flagged as DIF "B" that were placed on a test form. In 2017–18, the Fairness Review panel for EOG and EOC mathematics was made up of 12 participants representing teachers and educators. These members were selectively recruited based on their expert knowledge of mathematics content. Their demographic information is summarized in *Table 4.1*.

*Table 4.1 Demographic Information of Fairness Review Panels, Spring 2018*

| Category | Subcategory | N | % |
|---|---|---|---|
| Gender | Female | 4 | 33% |
| | Male | 8 | 67% |
| Ethnicity | Black | 1 | 8% |
| | White | 11 | 92% |
| Highest Degrees Earned | J.D./Ed.D./Ph.D. | 8 | 67% |
| | MA/MS/M.Ed. | 4 | 33% |
| Year of Experience | >20 | 3 | 25% |
| | 10–20 | 7 | 58% |
| | 1–10 | 2 | 17% |

Prior to reviewing items, panelists had to complete an online fairness review training process through the NC Review System. See *Appendix 4–B* for an overview of fairness review training process. The current operational goal is to minimize the use of DIF B items on operational forms. *Table 4.2* shows the distribution of items in operational forms by DIF category for EOG and NC Math 1 tests from spring 2018 administration and NC Math 3 from 2018-19 operational administration. Notice that DIF flags for the forms across grades and courses were mostly category "A" and few "B". All category "B" flagged items were reviewed and approved by the Fairness Review panel.

During form review, all DIF B items shown in *Table 4.2* were reviewed and approved by the DIF review panel. Panelists were asked to evaluate the item based on the following criteria:

- Does the item contain language that is not commonly used statewide or has different connotations in different parts of the state or in different cultural or gender groups?
- Does the item contain any local references that are not a part of the statewide curriculum?
- Does the item portray anyone in a stereotypical manner? (This could include activities, occupations, or emotions.)
- Does the item contain any demeaning or offensive materials?
- Does the item have offensive, stereotyping, derogatory, or proselytizing religious references?
- Does the item assume that all students come from the same socioeconomic background? (e.g., a suburban home with two-car garage)
- Does the artwork adequately reflect the diversity of the student population?
- Are there other bias or sensitivity concerns?

*Table 4.2 Mathematics Edition 5 Test Structure by DIF Types*

| Grade | Mode | Form | DIF Classification | | Total Items |
|---|---|---|---|---|---|
| | | | DIF A | DIF B | |
| 3 | Both | A | 34 | 6 | 40 |
| | Both | B | 35 | 5 | 40 |
| | Both | C | 32 | 8 | 40 |
| 4 | Both | A | 36 | 4 | 40 |
| | Both | B | 38 | 2 | 40 |
| | Both | C | 35 | 5 | 40 |
| 5 | Both | A | 33 | 7 | 40 |
| | Both | B | 39 | 1 | 40 |
| | Both | C | 36 | 4 | 40 |
| 6 | Both | A | 45 | 0 | 45 |
| | Both | B | 40 | 5 | 45 |
| | Both | C | 41 | 4 | 45 |
| 7 | Both | A | 39 | 6 | 45 |
| | Both | B | 43 | 2 | 45 |
| | Both | C | 39 | 6 | 45 |
| 8 | CBT | M | 33 | 12 | 45 |
| | CBT | N | 35 | 10 | 45 |
| | CBT | O | 36 | 9 | 45 |
| NC Math 1 | CBT | M | 41 | 9 | 50 |
| | CBT | N | 40 | 10 | 50 |
| | CBT | O | 44 | 6 | 50 |
| NC Math 3 | CBT | M | 43 | 7 | 50 |
| | CBT | N | 42 | 8 | 50 |
| | CBT | O | 46 | 4 | 50 |

The review panelists used an online review platform in which they are able to provide additional content for any category they responded “Yes” indicating they suspect an item is associated with a bias, sensitivity or accessibility issue.

Based on the reviews from all panelist, a final determination is made whether to retain or delete any of these items from the operational form. Any item that receives an affirmative response to any of these questions asked during fairness review are further reviewed by content test specialists at the NCDPI and the NCSU-TOPS to make a final recommendation of whether to replace these items from the form. Furthermore, all experts must agree these flagged items

measure the content that is expected of students with no obvious indication of specific construct-irrelevant variance.

## 4.5 Summary of Final Operational Forms

This section details test format and statistical properties of new *Edition 5* mathematics EOG and EOC test forms that were built in 2018 using embedded field-test items. All forms were built based on test specification criteria outlined in Chapter 2. EOC NC Math 3 and, to some extent, EOG grade 8 forms were the only exceptions in which the NCDPI had to deviate from its standard test development protocol. The details and special circumstances surrounding these forms will be addressed in this section.

### 4.5.1 *Edition 5* EOG and EOC Mathematics Test Format

*Table 4.3* and *Table 4.4* display test format of the final assembled operational base forms in terms of items counts, item types and calculator use. Each item carried a maximum of 1 score point. The grades 3–8 forms were the same in both modes (paper-based, or PBT, and computer-based, or CBT).

The NC Math 1 and NC Math 3 forms included technology item types – Drag-and-Drop, Text Identify and Targeted Drop – and were primarily designed for CBT. For the PBT forms, the technology items in the EOC forms were replaced with MC items. Examples of the TE item types can be accessed from the NCDPI website: https://www.dpi.nc.gov/districts-schools/testing-and-school-accountability/state-tests/end-course-eoc#released-forms.

*Table 4.3 Test Format of EOG Mathematics Grades 3–8*

| Grade | Number of Items | Mode | Form | Item Types | | Calculator Use | |
|---|---|---|---|---|---|---|---|
| | | | | MC | TE | Yes | No |
| 3 | 40 | Both | A | 40 | | 20 | 20 |
| | 40 | Both | B | 40 | | 20 | 20 |
| | 40 | Both | C | 40 | | 20 | 20 |
| 4 | 40 | Both | A | 40 | | 20 | 20 |
| | 40 | Both | B | 40 | | 20 | 20 |
| | 40 | Both | C | 40 | | 20 | 20 |
| 5 | 40 | Both | A | 30 | 10 | 20 | 20 |
| | 40 | Both | B | 30 | 10 | 20 | 20 |
| | 40 | Both | C | 30 | 10 | 20 | 20 |
| 6 | 45 | Both | A | 35 | 10 | 30 | 15 |
| | 45 | Both | B | 35 | 10 | 30 | 15 |
| | 45 | Both | C | 35 | 10 | 30 | 15 |
| 7 | 45 | Both | A | 35 | 10 | 30 | 15 |
| | 45 | Both | B | 35 | 10 | 30 | 15 |
| | 45 | Both | C | 35 | 10 | 30 | 15 |
| 8 | 45 | CBT | A | 35 | 10 | 30 | 15 |
| | 45 | CBT | B | 35 | 10 | 30 | 15 |
| | 45 | CBT | C | 35 | 10 | 30 | 15 |

*Table 4.4 Test Format of NC Math 1*

| Course | Number of Items | Mode | Form | Score Per Item | Item Types | | | | | Calculator Use | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | MC | GR | DD | TI | TD | Yes | No |
| NC Math 1 | 50 | PBT | A | 1 | 39 | 11 | | | | 35 | 15 |
| | 50 | CBT | M | 1 | 29 | 16 | 3 | 1 | 1 | 35 | 15 |
| | 50 | CBT | N | 1 | 29 | 16 | 3 | | 2 | 35 | 15 |
| | 50 | CBT | O | 1 | 29 | 16 | 2 | 2 | 1 | 35 | 15 |

### 4.5.2 DOK Distributions

Test specification guidelines for cognitive complexity using DOK are shown in *Table 4.5* for grades 3–8 and *Table 4.6* for NC Math 1 and NC Math 3. This table confirms the DOK distribution by forms across grades is consistent with recommendations made during test specification meetings.

*Table 4.5 Mathematics DOK Distributions, EOG Grades 3–8*

| Grade | Category | Blueprint (%) | Form A | | Form B | | Form C | |
|---|---|---|---|---|---|---|---|---|
| | | | N | % | N | % | N | % |
| 3 | DOK1 | 40–50 | 19 | 48 | 17 | 43 | 19 | 48 |
| | DOK2 | 50–60 | 21 | 53 | 23 | 58 | 21 | 53 |
| 4 | DOK1 | 35–45 | 16 | 40 | 14 | 35 | 14 | 35 |
| | DOK2 | 50–60 | 22 | 55 | 24 | 60 | 24 | 60 |
| | DOK3 | 5 | 2 | 5 | 2 | 5 | 2 | 5 |
| 5 | DOK1 | 30–40 | 15 | 38 | 16 | 40 | 13 | 33 |
| | DOK2 | 50–60 | 22 | 55 | 21 | 53 | 23 | 58 |
| | DOK3 | 8–10 | 3 | 8 | 3 | 8 | 4 | 10 |
| 6 | DOK1 | 25–35 | 14 | 31 | 15 | 33 | 14 | 31 |
| | DOK2 | 50–60 | 27 | 60 | 26 | 58 | 27 | 60 |
| | DOK3 | 8–15 | 4 | 9 | 4 | 9 | 4 | 9 |
| 7 | DOK1 | 25–35 | 14 | 31 | 14 | 31 | 14 | 31 |
| | DOK2 | 50–60 | 27 | 60 | 27 | 60 | 27 | 60 |
| | DOK3 | 8–15 | 4 | 9 | 4 | 9 | 4 | 9 |
| 8 | DOK1 | 25–35 | 14 | 31 | 14 | 31 | 15 | 33 |
| | DOK2 | 50–60 | 27 | 60 | 27 | 60 | 26 | 58 |
| | DOK3 | 8–15 | 4 | 9 | 4 | 9 | 4 | 9 |

*Table 4.6 DOK Distributions, EOC NC Math 1 and NC Math 3*

| Course | Category | Blueprint (%) | M | | N | | O | |
|---|---|---|---|---|---|---|---|---|
| | | | N | % | N | % | N | % |
| NC Math 1 | DOK1 | 20–30 | 15 | 30 | 14 | 28 | 15 | 30 |
| | DOK2 | 60–70 | 31 | 62 | 32 | 64 | 31 | 62 |
| | DOK3 | 8–12 | 4 | 8 | 4 | 8 | 4 | 8 |
| NC Math 3 | DOK1 | 20–30 | 13 | 26 | | | 13 | 26 |
| | DOK2 | 60–70 | 33 | 66 | | | 32 | 64 |
| | DOK3 | 8–12 | 4 | 8 | | | 5 | 10 |

### 4.5.3 Summary Statistics of Base Forms

*Table 4.7* and *Table 4.8* present form-level summary CTT statistics (p-values and biserial) and IRT statistics [slope (a), threshold (b), pseudo-guessing (g)] for new *Edition 5* EOG and NC Math 1 forms. Form level statistics were based on embedded spring 2018 field-test data. Both CTT and IRT statistics confirmed forms within grade were built to the same statistical target. This evidence suggests forms within grades are statistically equivalent or parallel. The NCSBE removed double testing requirements from 2018-19 administration. As a result, those grade 8 students who took NC Math 1 did not have to take the grade level test resulting in a relatively difficult grade 8 forms with smaller p-values and higher threshold parameters.

*Table 4.7 Average CTT and IRT Statistics for Grades 3–5, Spring 2018 Field–Test*

| Grade | Mode | Form | Number of Items | CTT | | IRT | | |
|---|---|---|---|---|---|---|---|---|
| | | | | P-value | Biserial-Corr. | Slope (a) | Threshold (b) | Asymptote (g) |
| 3 | Both | A | 40 | 0.62 | 0.47 | 1.96 | –0.12 | 0.18 |
| | Both | B | 40 | 0.62 | 0.48 | 2.00 | –0.12 | 0.15 |
| | Both | C | 40 | 0.62 | 0.47 | 2.05 | –0.12 | 0.17 |
| 4 | Both | A | 40 | 0.62 | 0.48 | 2.00 | –0.09 | 0.16 |
| | Both | B | 40 | 0.63 | 0.48 | 2.10 | –0.08 | 0.18 |
| | Both | C | 40 | 0.62 | 0.48 | 1.99 | –0.07 | 0.18 |
| 5 | Both | A | 40 | 0.60 | 0.46 | 1.94 | –0.02 | 0.16 |
| | Both | B | 40 | 0.60 | 0.46 | 1.97 | –0.02 | 0.15 |
| | Both | C | 40 | 0.60 | 0.47 | 1.87 | –0.02 | 0.16 |

*Table 4.8 Average CTT and IRT Statistics for Grades 6–8 and NC Math 1 FT, Spring 2018*

| Grade | Mode | Form | Number of Items | CTT | | IRT | | |
|---|---|---|---|---|---|---|---|---|
| | | | | P-value | Biserial-Corr. | Slope (a) | Threshold (b) | Asymptote (g) |
| 6 | Both | A | 45 | 0.52 | 0.48 | 2.15 | 0.27 | 0.16 |
| | Both | B | 45 | 0.52 | 0.49 | 2.33 | 0.29 | 0.16 |
| | Both | C | 45 | 0.52 | 0.49 | 2.33 | 0.30 | 0.17 |
| 7 | Both | A | 45 | 0.52 | 0.50 | 2.43 | 0.33 | 0.17 |
| | Both | B | 45 | 0.51 | 0.51 | 2.38 | 0.32 | 0.17 |
| | Both | C | 45 | 0.52 | 0.50 | 2.36 | 0.32 | 0.17 |
| 8 | Both | A | 45 | 0.43 | 0.39 | 1.76 | 0.68 | 0.16 |
| | Both | B | 45 | 0.43 | 0.38 | 1.79 | 0.75 | 0.16 |
| | Both | C | 45 | 0.43 | 0.37 | 1.81 | 0.84 | 0.17 |
| NC Math 1 | CBT | M | 50 | 0.48 | 0.49 | 1.99 | 0.36 | 0.13 |
| | CBT | N | 50 | 0.49 | 0.47 | 2.10 | 0.39 | 0.16 |
| | CBT | O | 50 | 0.47 | 0.46 | 1.99 | 0.43 | 0.14 |

## 4.6 EOC NC Math 3 Form Development

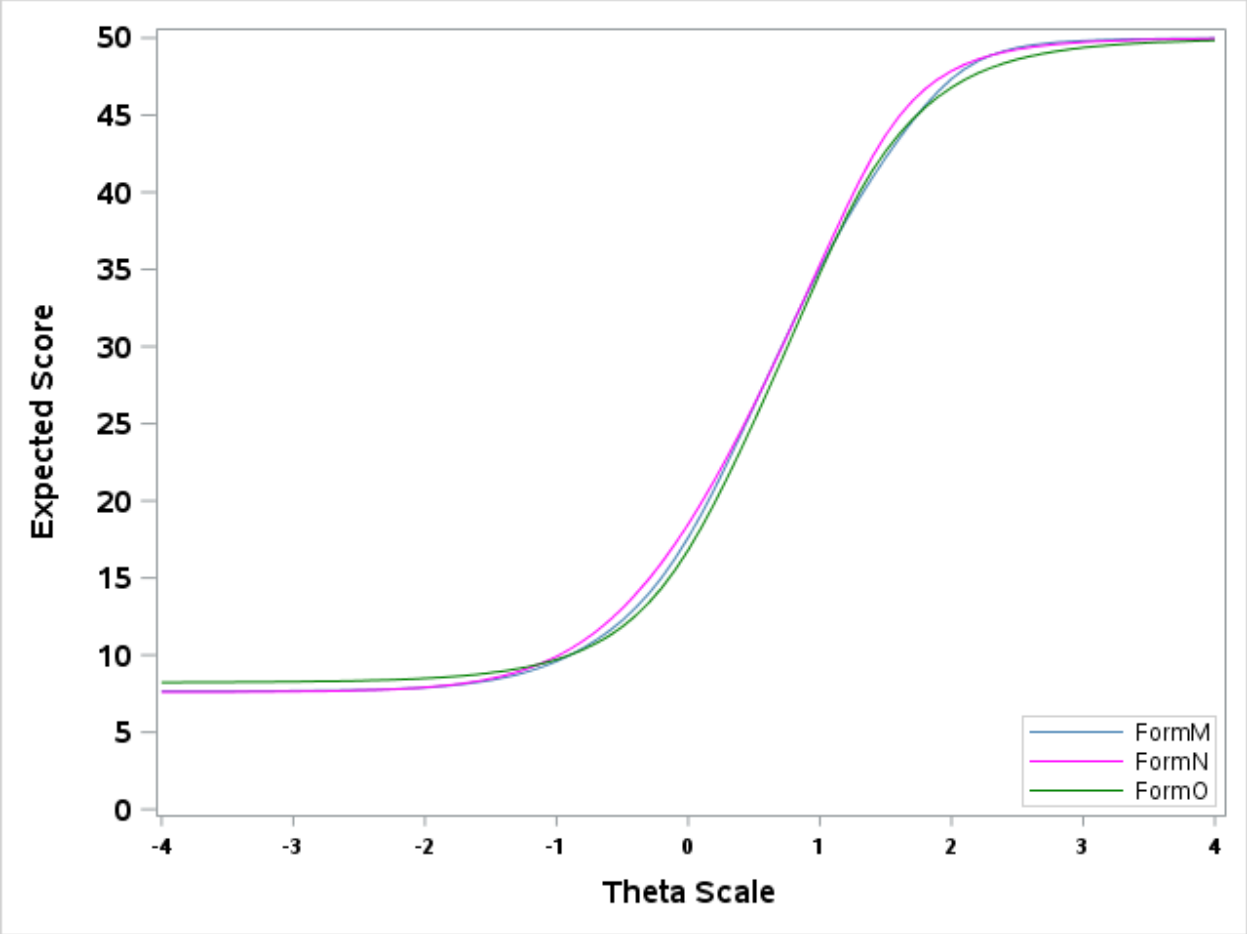
As 2017–18 NC Math 3 stand-alone field data was determined not reliable and could not be used to set statistical targets for new forms, the NCDPI used a hybrid form development plan for the 2018–19 administration with the expectation to conduct post administration analysis and revise the forms for operational use to ensure they match content and statistical specifications.

During the form development for 2018–19, three 60-item forms were selected from the 2017–18 item pool with heavy reliance on content judgement. The normal statistical rules used to evaluate and rank field-test items were relaxed and content preferences on item quality was given priority. All forms were built based on the same content blueprint and followed the usual established content and production review steps described in the previous form review section. The plan with the 60-item field test was to allow flexibility during post administration analysis to select 50 optimal items for each operational form as specified by the blueprint while maintaining statistical and content balance across all forms. *Table 4.9* shows content structure of NC Math 3 field-test and operational forms. The NC Math 3 forms are administered in CBT mode except for accommodations and approved technical hardships. The TCCs of the NC Math 3 operational forms are shown in *Figure 4.8.*

*Table 4.9 Test Format of NC Math 3*

| NC Math 3 | Number of Items | Mode | Form | Item Types | | | | | Calculator Use |
|---|---|---|---|---|---|---|---|---|---|
| | | | | MC | GR | DD | TI | TD | Yes |
| Field–Test | 60 | CBT | M | 38 | 15 | 3 | 2 | 2 | 60 |
| | 60 | CBT | N | 38 | 15 | 3 | 3 | 1 | 60 |
| | 60 | CBT | O | 38 | 15 | 2 | 3 | 2 | 60 |
| Operational | 50 | CBT | M | 33 | 13 | 1 | 2 | 1 | 50 |
| | 50 | CBT | N | 32 | 13 | 3 | 1 | 1 | 50 |
| | 50 | CBT | O | 34 | 12 | 1 | 2 | 1 | 50 |

*Figure 4. 8 TCCs Based on 2018-19 Operational Item Parameters, NC Math 3*



# CHAPTER 5 TEST ADMINISTRATION

Standard 6.0 (AERA, APA, & NCME, 2014) states, "*To support useful interpretations of score results, assessment instruments should have established procedures for test administration, scoring, reporting, and interpretation. Those responsible for administering, scoring, reporting, and interpreting should have sufficient training and supports to help them follow the established procedures*…" (p.114). In adherence to this standard, this chapter briefly describes the NCDPI's established policies and procedures used to train test coordinators and test administrators in order to ensure standardized test administrations across the state. This chapter also provides information about test administration guides, testing windows, mode of administrations, timing guidelines, testing accommodations and mechanism for reporting test irregularities and misadministration.

## 5.1 Test Administration Guides and the Test Coordinators' Handbook

The NCDPI produces comprehensive test administration guides for each state mandated test with the exclusion of tests that are provided by a vendor. When a vendor assessment is used the school must follow the vendor's policies and procedures, which are provided in the vendor guides. The administration guides available for test coordinators and test administrators to ensure standardized administration of all tests given across the state are briefly described below with website links for more detailed descriptions.

The Proctor's Guide: The guide serves as a resource document with detailed guidelines on selecting proctors and how they should be trained. This guide also includes information about how to maintain test security, ensure appropriate testing conditions, maintain students' confidentiality, assist test administrators, monitor students, report test irregularities and follow appropriate procedures for accommodations. The document can be accessed from the NCDPI website: https://files.nc.gov/dpi/documents/files/1920proctors.pdf.

Guidelines for Testing Students Identified as English Learners (ELs) and for Testing Students with Disabilities*:* The NCDPI produces the guidelines for training test administrators and test coordinators. The document for the English Learners can be accessed from the NCDPI website at https://files.nc.gov/dpi/documents/files/new-format-19-20-tsiels_final-3_0.pdf, and the document for the students with disabilities can be accessed from the NCDPI website at https://files.nc.gov/dpi/documents/files/tswd_final-pdf-7.31.19_0.pdf. These publications include information on testing requirements, responsibilities for test coordinators and test administrators, procedures for participation (with or without accommodations) and accommodations monitoring. *Standard 4.15* (AERA, APA, & NCME, 2014), regarding the direction for test administration, states, "*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. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented*" (p. 90).

Testing Security Protocols and Procedures for School Personnel: The NCDPI publishes this document in order to maintain the integrity of the North Carolina Testing Program. It is essential for school personnel to develop awareness of proper testing protocol and procedures. Knowledge of testing policies and procedures helps ensure the North Carolina Testing Program is conducted in a manner that is fair, consistent and equitable for all students. The purpose of this publication is to provide principals, teachers and other school personnel with a reference for implementing secure, uniform test administrations for the North Carolina Testing Program. This testing security guide may be kept in the schools. Additional copies may be downloaded from the NCDPI website: https://files.nc.gov/dpi/documents/files/2019-20-testing-security-working_final.pdf.

North Carolina Test Coordinators' Policies and Procedures Handbook: The purpose of the handbook is to provide public school unit test coordinators with a reference for implementing proper test administrations for the North Carolina Testing Program. The handbook can be accessed from the NCDPI website: https://files.nc.gov/dpi/documents/files/19-20-tc-handbook.pdf. The handbook provides information to ensure the integrity of the testing program is maintained, results generated from the program are valid and any subsequent reporting is accurate and appropriate. It is essential for school personnel to develop awareness of proper testing procedures in order to provide accurate test data for decision-making. The North Carolina Testing Program must be conducted in a manner that is fair, consistent and equitable for all students. The Handbook also details the design of each assessment in order for preparations necessary before test day, on test day and after the test is complete and the purpose of the assessments, student eligibility, testing windows and procedures for makeup testing.

## 5.2 Test Administrators Training

The test administrators' training utilizes the *North Carolina Test Coordinators' Policies and Procedures Handbook* as well as all other NCDPI publications discussed in Section 5.1. These documents contain comprehensive information on test administration including test security, roles and responsibilities of test administrators, test administration preparation, monitoring, testing accommodations, online testing, testing irregularities and available resources. The North Carolina Testing Program uses a train-the-trainer model to prepare test administrators to administer all North Carolina tests. Regional Accountability Coordinators (RACs) receive training from the NCDPI Testing Policy and Operations staff during scheduled monthly training sessions. Subsequently, the RACs provide training to public school unit test coordinators on the processes for proper test administration. Public school unit test coordinators provide this training to school test coordinators. The training includes information on the test administrators'

responsibilities, proctors' responsibilities, preparing students for testing, eligibility for testing, policies for testing students with special needs (students with disabilities and students with limited English proficiency), accommodated test administrations, test security (storing, inventorying and returning test materials) and the *Testing Code of Ethics* that may be downloaded from the NCDPI website: https://files.nc.gov/dpi/documents/files/testing_code_of_ethics_0.pdf.

## 5.3 Test Security and Administration Policies

Test security is an ongoing concern for the North Carolina testing program. When test security is compromised, it can undermine the validity of test scores. For this reason, the NCDPI has taken extensive steps to ensure the security of the assessments by establishing protocols for school employees administering tests.

### 5.3.1 Protocols for Test Administrators

Only school system employees are permitted to administer secure state tests. Those employees must participate in the training for test administrators as described in Section 5.2. Test administrators may not modify, change, alter, or tamper with student responses on answer sheets or in test books. Test administrators must thoroughly read and be trained on the appropriate *Test Administration Guide* and the codified North Carolina *Testing Code of Ethics* prior to the test administration. Test administrators must follow the instructions to ensure a standardized administration and read aloud all directions and information to students as indicated in the manual. The school test coordinator is responsible for monitoring test administrations within the building and responding to situations that may arise during test administrations.

### 5.3.2 Protocol for Handling of Paper-Based Tests

When administering paper-based test, school systems are mandated to provide a secure area for storing tests. The Administrative Procedures Act 16 NCAC 6D.0302 states, in part, that LEAs shall (1) account to the NCDPI for all tests received; (2) provide a secure, locked storage area for all tests received; (3) prohibit the reproduction of all or any part of the tests; and (4) prohibit their employees from disclosing the content of, or specific items contained in, the test to persons other than authorize employees of the LEA.

At the individual school, the principal is responsible for all test materials received. As established by SBE policy GCS-A-010, the *Testing Code of Ethics*, the principal must ensure test security within the school building and store the test materials in a secure, locked facility except when in use. The principal must establish a procedure to have test materials distributed immediately before each test administration. Every LEA and school must have a clearly defined system of check-out and check-in of test materials to ensure at each level of distribution and

collection (district, school and classroom) all secure materials are tracked and accounted for. Public school unit test coordinators must inventory test materials upon arrival from NCSU-TOPS and must inform NCSU-TOPS of any discrepancies in the shipment.

Before each test administration, the school test coordinator shall collect, count and return all test materials to the secure, locked storage area. Any discrepancies are to be reported to the public school unit test coordinator immediately and a report must be filed with the Regional Accountability Coordinator (RAC). At the end of each test administration cycle, all testing materials must be returned to the school test coordinator according to directions specified in the test administration guide. Immediately after each test administration, the school test coordinator shall collect, count and return all test materials to the secure, locked facility. Any discrepancies must be reported immediately to the public school unit test coordinator. Upon notification, the public school unit test coordinator must report the discrepancies to the RAC and ensure all procedures in the Online Testing Irregularity Submission System (OTISS) are followed to document and report the testing irregularity. The procedures established by the school for tracking and accounting for test materials must be provided upon request to the public school unit test coordinator and/or the NCDPI Division of Accountability Services/North Carolina Testing Program.

At the end of the testing window, the NCDPI mandates that all test administration guides, used test booklets that do not contain valid student responses, unused test booklets and unused answer sheets be immediately securely destroyed by the district at the LEA. Secure test materials are to be retained by the LEA district/school in a secure, locked facility with access controlled and limited to one or two authorized school personnel only. After the required storage time has elapsed, the LEA should securely destroy these materials. The test materials and required storage time are listed in *Table 5.1*.

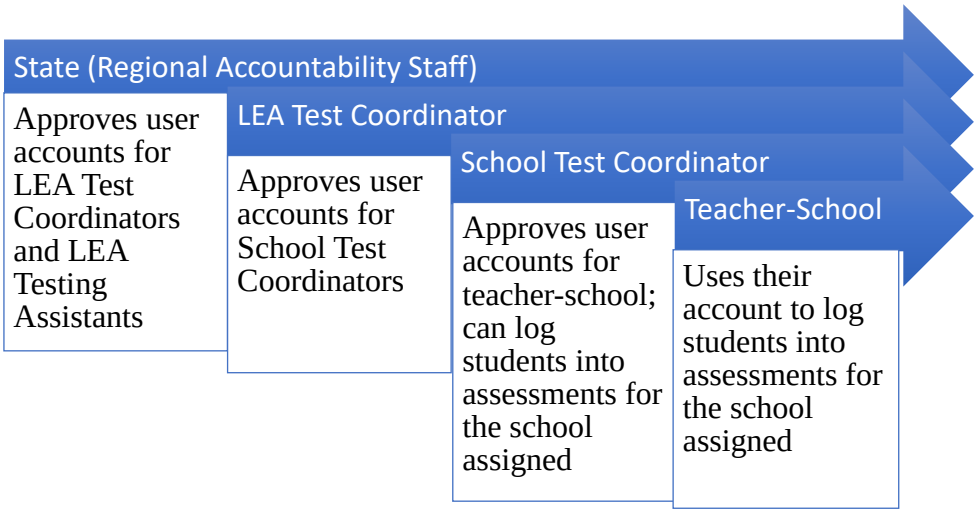
*Table 5. 1 Test Materials Designated to be Stored by the District/School in a Secure Location*

| Test Material | Required Storage Time |
|---|---|
| All used answer sheets for operational tests (including scoring sheets for W-APT) | Six months after the return of students' test scores |
| Original responses recorded in a test book, including special print version test books (i.e., large print edition, one test item per page edition, Braille edition) | Six months after the return of students' test scores |
| Original Braille writer/slate and stylus responses | Six months after the return of students' test scores |
| Original responses to a scribe | Six months after the return of students' test scores |
| Original responses using a typewriter or word processor | Six months after the return of students' test scores |
| Answer sheets with misaligned answers (keep testing irregularities in a separate file) | Six months after the return of students' test scores |
| NC General Purpose Header Sheets | Store indefinitely |
| EOC or EOG Graph Paper | Store indefinitely |
| EOC: Math 1, Biology and English II | Retain unused test materials from fall for use in spring |
| W-APT test materials (reusable except for scoring sheets) | Store indefinitely (all forms) |

### 5.3.3 Protocol for Handling of Computer–Based Tests

The NCTest platform (1024X768) is used to administer computer-based fixed-form tests. The NC Education system manages student enrollments, monitors assessment start and stop times and collects accommodation information. The NCDPI limits all public school units' access to the CBT to specific testing days. The public school unit test coordinator must enter test dates in NC Education for each assessment to be administered by computer. Assessments can only be accessed through NCTest on those specific dates. In addition, access is limited to users with a valid and verified NC Education username and password. *Figure 5.1* shows the tiers of NCTest users along with the information about who assigns access. The NCTest platform is via a safe exam browser, NCTest app for chrombooks, or the NCTest app for iPads.

*Figure 5. 1 NCTest User Access Security Protocol*




The connection is encrypted using Transport Layer Security (TLS 1.2) and authenticated using AES_128_GCM with DHE_RSA as the exchange mechanism. At the time of login, the tests are sent securely from the NCTest server at North Carolina State University (NCSU) to the local computer. Not all assessment content is sent at the time of login, only the text for all the test items are sent at that time. Graphics and audio files (for computer read-aloud accommodation) are sent as students move from item to item within the assessment. Student responses are securely sent after each item is answered to the NCTest server at NCSU using the same full HTTPS encryption process. At the conclusion of the assessment, local users are instructed to clear all cache and cookies from local machines.

After online student assessments are finalized, they are transferred nightly to the NCDPI and/or to the scoring vendors. These transfers are done following the NCDPI Secure File Transfer Protocol (SFTP) encryption rules and logic. More information on these processes can be found in the NCDPI's Test Coordinators' Policies and Procedures Handbook under "Maintaining the Confidentiality and Security of Testing and Accountability Data" section. The handbook can be accessed from the NCDPI website: (https://files.nc.gov/dpi/documents/files/19-20-tc-handbook.pdf).

## 5.4 Test Administration

*Standard 6.1* (AERA, APA, & NCME, 2014) states, "*Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and*

*any instructions from the test user*" (p. 114). The standardized procedures reduce construct-irrelevant variance and enhance the reliability and validity of the resulting test scores.

### 5.4.1 Testing Windows

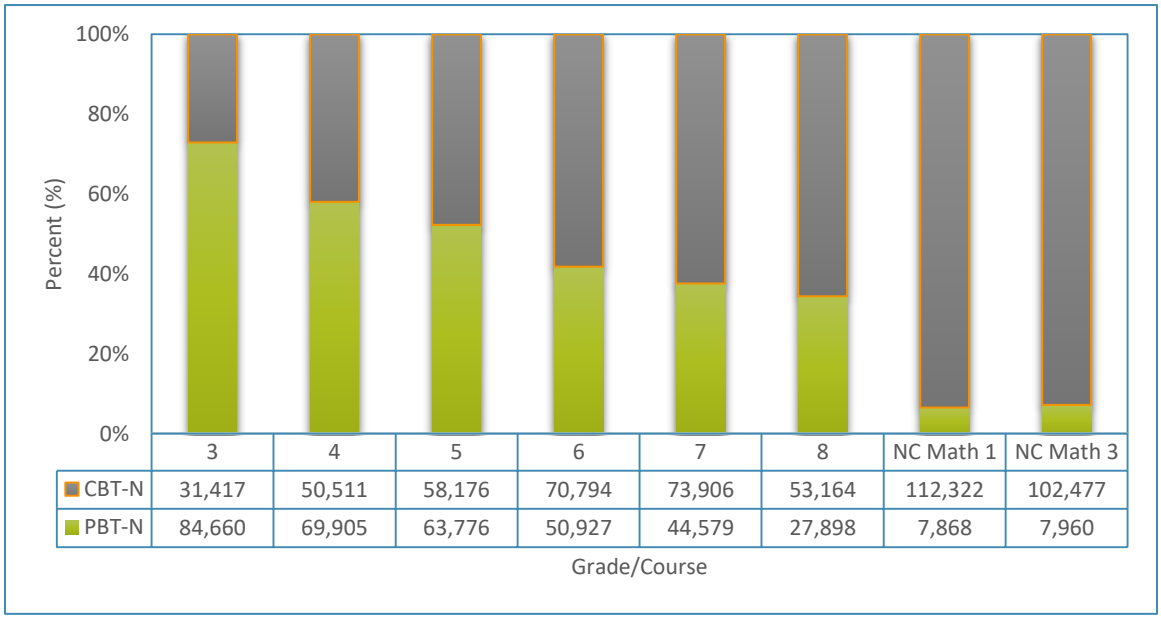
Per G.S. §115C-174.12(a)(4), "*all annual assessments of student achievement adopted by the State Board of Education pursuant to G.S. §115C-174.11(c)(1) and (3) and all final exams for courses shall be administered within the final ten (10) instructional days of the school year for yearlong courses and within the final five (5) instructional days of the semester for semester courses*." Exceptions are permitted to allow testing of a student outside the designated testing window to accommodate a student's IEP or Section 504 Plan. In rare circumstances (e.g., family emergency, family relocation, scheduled surgery during the test window) may exist and preclude an individual student from being tested during a state testing window, including makeup dates where students are permitted to test before or after the testing window. All EOG assessments are administered in spring. The mathematics EOC tests are semester courses and have two administration windows: one in fall and another in spring.

### 5.4.2 Modes of Test Administration

From the 2018–19 administration, grades 3–8 EOG mathematics assessments are available in both paper-based and computer-based modes. The state's goal is to gradually transition test administrations for EOG to the computer mode as districts can build their resources and technology capacity. The EOC NC Math 1 and NC Math 3 are both computer-based tests unless students require paper-based for accessibility purposes. *Figure 5.2* shows the proportion and total number of students who took the mathematics EOG and EOC tests by mode in the 2018–19 administration. Notice that the proportion of students who took CBT forms increased gradually as the grade level increased. The proportion of students taking CBT forms is expected to increase gradually in subsequent administration.

Three parallel forms (A, B, and C) were administered in 2018–19 administration for each grade 3–8 EOG test. These forms contained the same items and were designed to be administered in both paper-based and computer-based modes. In the NC Math 1 and NC Math 3 tests, the computer-based forms contained multiple-choice and technology-enhanced items and the paper-based forms were developed by replacing the technology items in computer-based forms with multiple-choice items. Despite different item types, the forms were developed with the same content and statistical specifications.

*Figure 5. 2 Number (N) and Percent (%) of Students by Mode, 2018–19*




| | 3 | 4 | 5 | 6 | 7 | 8 | NC Math 1 | NC Math 3 |
|---|---|---|---|---|---|---|---|---|
| CBT-N | 31,417 | 50,511 | 58,176 | 70,794 | 73,906 | 53,164 | 112,322 | 102,477 |
| PBT-N | 84,660 | 69,905 | 63,776 | 50,927 | 44,579 | 27,898 | 7,868 | 7,960 |

At every grade level, the three forms were spiraled within the classroom of a given school. Spiraling forms within the classroom ensures that the samples taking the three forms are equivalent. Therefore, it can be assumed that the item parameters estimated from the calibration of the random samples of students who were administered different test forms are on a common IRT scale and are comparable directly without need for equating. In the succeeding administrations, a combination of two forms will be administered. An embedding plan for field testing items will be proposed for developing new forms as needed.

### 5.4.3 Testing Time Guidelines

When taking the tests, all examinees are given ample time to demonstrate their knowledge of the construct being assessed. The AERA, APA, & NCME (2014) states, "*Although standardization has been a fundamental principle for assuring that all examinees have the same opportunity to demonstrate their standing on the construct that a test is intended to measure, sometimes flexibility is needed to provide essentially equivalent opportunities for some test takers*" (p.51). In adherence with the Standards, the NCDPI requires all general students be allowed ample opportunity to complete the assessments as long as they are engaged, and the maximum time allowed has not elapsed. Based on the timing data from field-test, the NCDPI's recommended time allotted for EOG tests is two hours with additional one hour if needed to complete the test. For the EOC NC Math 1 and NC Math 3, the recommended time is three hours with additional

one hour if needed to complete the test. Students with approved accommodations may take longer, as specified in their IEP or Section 504 Plan.

Summary timing data for the 2018–19 operational mathematics assessments are shown in *Table 5.2*. The table includes data for mathematics EOG and EOC computer-based forms administered under regular conditions—that is, without accommodations of *Scheduled Extended Time* and *Multiple Testing Sessions*. For grades 3–8, 95% of students completed the assessments close to within the three (3) hours (186 minutes or less) window and 99% of students in the sample completed in about four (4) hours (231 minutes or less). Note that grades 5–8 mathematics tests consisted of gridded response items. Similarly, at least 95% of high school students took about four hours or less (243 minutes for NC Math 1 and 204 minutes for NC Math 3) to complete the test. For the NC Math 3, at least 1% took more than four hours. Moreover, students took more time in NC Math 1 than in NC Math 3.

*Table 5. 2 Recorded Test Duration for Mathematics EOG and EOC Operational Forms, 2018–19*

| Grade | N | No. of OP+FT Items | Summary | | Percentile | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | SD | 25th | Median | 75th | 95th | 99th |
| 3 | 31,416 | 46 | 89.5 | 42.3 | 60 | 81 | 108 | 165 | 213 |
| 4 | 50,511 | 46 | 104.8 | 46.3 | 75 | 99 | 126 | 180 | 231 |
| 5 | 58,176 | 48 | 113.9 | 43.1 | 87 | 108 | 132 | 186 | 228 |
| 6 | 70,794 | 53 | 114.8 | 40.8 | 90 | 111 | 132 | 180 | 216 |
| 7 | 73,906 | 53 | 119.4 | 41.2 | 96 | 117 | 141 | 183 | 219 |
| 8 | 53,163 | 53 | 114.7 | 40.7 | 90 | 111 | 135 | 180 | 213 |
| NC Math 1 | 112,322 | 60 | 152.8 | 55.7 | 117 | 153 | 183 | 243 | 279 |
| NC Math 3 | 102,477 | 60 | 122.4 | 52.0 | 87 | 120 | 156 | 204 | 243 |

## 5.5 Testing Accommodations

State and federal law requires that all students, including SWD and students identified as English Learners (ELs), participate in the statewide testing program. Students may participate in the state assessments on grade level (i.e., general or alternate) with or without testing accommodations. AERA, APA, & NCME (2014) states that the eligible students participating in the EOG and EOC are provided with "*test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs"* (p. 67). Shyyan et al. (2016) define testing

## 5.5 Testing Accommodations

State and federal law requires that all students, including SWD and students identified as English Learners (ELs), participate in the statewide testing program. Students may participate in the state assessments on grade level (i.e., general or alternate) with or without testing accommodations. AERA, APA, & NCME (2014) states that the eligible students participating in the EOG and EOC are provided with "*test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs"* (p. 67). Shyyan et al. (2016) define testing accommodations as "*changes in assessment materials or procedures that address aspects of students' disabilities that may interfere with the demonstration of their knowledge and skills on standardized tests*". Accommodations are provided to eligible students with appropriate administrative procedures to assure that individual student needs are met while maintaining sufficient uniformity of the test administration.

For any state-mandated test, the accommodation(s) for an eligible student must (1) be documented in the student's current IEP, Section 504 Plan, EL Plan, or transitory impairment documentation and (2) the documentation must reflect routine use during instruction and similar classroom assessments that measure the same construct. When accommodations are provided in accordance with proper procedures as outlined by the state, results from these tests are deemed valid and fulfill the requirements for accountability.

According to Standard 6.2 (AERA, APA, & NCME, 2014), *"When formal procedures have been established for requesting and receiving accommodations, test takers should be informed of these procedures in advance of testing"* (p. 115). In compliance with the standard, the NCDPI specifies the following accommodations in North Carolina EOG and EOC assessments in the Test Administration Guide from 2018–19 and beyond:

- Special Print Versions
    - Braille Edition
    - Large Print Edition
    - One Test Item Per Page Edition
- Large Print One Test Item Per Page Edition
- Assistive Technology (AT) Devices and Special Arrangements
    - Assistive Technology Devices
    - Dictation to a Scribe
    - Interpreter/Transliterator Signs/Cues Test
    - Student Marks Answers in Test Book (not for online assessments)
    - Student Reads Test Aloud to Self
    - Test Read Aloud (in English)
    - Braille Writer/Slate and Stylus (Braille Paper)

    - Cranmer Abacus
    - Magnification Devices
    - Word-to-Word Bilingual (English/Native Language) Dictionary/Electronic Translator (EL only)
- Special Test Environments
    - Multiple Testing Sessions
    - Scheduled Extended Time
    - Testing in a Separate Room

### 5.5.1 Accommodations for Students with Disabilities

For information regarding appropriate testing procedures, test administrators who provide accommodations for students with disabilities must refer to the most recent publication of *Testing Students with Disabilities* and any published supplements or updates. The publication is available through the local school system or the NCDPI website at https://files.nc.gov/dpi/documents/files/tswd_final-pdf-7.31.19_0.pdf. In addition, test administrators must be trained in the use of the specified accommodations by the school system test coordinator or designee prior to the test administration.

According to the AERA, APA, & NCME (2014), an appropriate accommodation addresses a student's specific characteristics, but does not change the construct the test is measuring or the meaning of the score. The NCDPI's test administration guide recommends that students should only be allowed the same accommodations for assessments as those routinely used during classroom instruction and other classroom assessments that measure the same construct.

### 5.5.2 Accommodations for English Learners

North Carolina State Board of Education policy TEST-011 states that "*students identified as ELs shall participate in the statewide testing program using either the standard test administration or the standard test administration with accommodations. Consistent with State Board policies TEST-003 and TEST-016, EL students in their first year in a U.S. school shall take required EOC and North Carolina Final Exams (NCFEs), but the test scores shall not be included as at least 20% of the student's final grade for the course. This applies to English/Language Arts/Reading, Mathematics, Science, and Social Studies EOC and NCFE assessments*."

Per SBE policy TEST-011, to be identified as English Learners (ELs), students indicating a language other than English on the Home Language Survey must be assessed using the state EL identification test at initial enrollment. The NCDPI uses WIDA™ Screener Online as the state-designated EL proficiency identification test given to students in second semester grades 1 through 12 and the ACCESS for ELLs® as the state-designated EL proficiency assessment administered annually to kindergarten through twelfth grade to students who have been identified

as ELs. Students who score below Level 5.0 Bridging on the reading domain of the WIDA Screener/ACCESS for ELLs are eligible to receive state approved EL testing accommodations on all state tests. Students who score Level 5.0 Bridging or above on the reading domain of the WIDA Screener/ACCESS for ELLs or exit EL status must participate in all state tests without accommodations (SBE policy TEST-011) (see Figure 5.3). The state approved EL testing accommodations include Word-to-Word Bilingual (English/Native Language) Dictionary/Electronic Translator, Multiple Testing Sessions, Scheduled Extended Time, Testing in a Separate Room, Student Reads Test Aloud to Self and Test Read Aloud (in English).

*Figure 5. 3 Students Eligible to Receive EL Testing Accommodations*

| Subtest | 1 Entering | 2 Emerging | 3 Developing | 4 Expanding | 5 Bridging | 6 Reaching |
|---|---|---|---|---|---|---|
| Reading | **Eligible to Receive State-Approved EL Testing Accommodations for All State Tests** | | | | Must Participate in General State Test Administration without EL Testing Accommodations | |

## 5.6 Student Participation

The administrative procedures described in North Carolina Register 16 NCAC 06D. 0301 requires that all public school students enrolled in grades for which the North Carolina State Board of Education adopts an assessment, including every child with disabilities, participate in the testing program with the exception of a medical emergency. All students in grades 3 through 8 are required to participate in the EOG tests or the corresponding alternate assessment, as indicated by the student's IEP, Section 504, EL Plan/documentation, or Transitory Impairment documentation. All students enrolled in NC Math 1 and NC Math 3 as a course for credit, must be administered the EOC tests. Students who are repeating the course for credit must also be administered the EOC tests.

According to the State Board policy GCS-A-001, school systems shall, at the beginning of the school year, provide information to students and parents or guardians advising them of the district-wide and state-mandated assessments that students are required to take during the school year. In addition, school systems must provide information to students and parents or guardians to advise them of the dates the tests will be administered and how the results from each assessment will be used. Information provided to parents about the tests must include whether the NCSBE or local board of education requires the test. School systems must report test scores

and interpretative guidance from district-wide and/or state-mandated tests to students and parents or guardians within thirty (30) days of the generation of the score at the school system level or receipt of the score and interpretive documentation from the NCDPI.

### 5.6.1 Medical Exception

There may be rare circumstances in which a student with a significant medical emergency and/or condition may be excused from the required state tests. The medical emergencies may include, but not limited to, circumstances involving students who are i) in the final stage of a terminal or degenerative illness, ii) comatose, or iii) receiving extensive short-term terminal treatment due to a medical emergency. For requests that involve significant medical emergencies and/or conditions, a school may request from the Division of Accountability Services/North Carolina Testing Program a testing exception for the student. There is a process in place for requesting the medical exception. The request must be submitted on the superintendent's or school director's letterhead and include the original signature of the superintendent or school director. The request must include detailed justification explaining why the student's medical emergency and/or condition prevent participation in the respective test administration during the testing window and the subsequent makeup period. Most of what is submitted for the medical exception is housed at the school level (IEP, dates of the scheduled test administration(s) and makeup dates, number of days of instruction missed due to the emergency/condition, expected duration/recovery period, explanation of the condition and how it affects the student on a daily basis, etc.). The student's records remain confidential and any written material containing identifiable student information is not disseminated or otherwise made available to the public. For more information on the process for requesting medical exceptions based on significant medical emergencies and/or conditions, please access to "*Memo-Testing-Exceptions-Medical-Emergencies-0916.pdf*" by visiting the following the NCDPI website: *https://ec.ncpublicschools.gov/policies/nc-policies-governing-services-for-children-with-disabilities/ncdpi-communication/2016-2017/ec-division-memos/testing-exceptions-for-medical-emergencies/view*.

## 5.7 Test Irregularity and Misadministration

Standard 6.7 (AERA, APA, & NCME, 2014) states, "*Test users have the responsibility of protecting the security of test materials at all times*" p.117. Any action that compromises test security or score validity is prohibited. These may be classified as testing irregularities or misadministration. The NCDPI has a process in place to report testing irregularities and misadministration. A sample test security reporting plan is shown in the *North Carolina Test Coordinator Policies and Procedures Handbook* (https://files.nc.gov/dpi/documents/files/19-20-tc-handbook.pdf, p.91). Test administrators and proctors (if utilized) must report any alleged testing violation or testing irregularity to the school test coordinator on the day of the occurrence. The school test coordinator must contact the school system test coordinator immediately with

any allegation of a testing violation. The school test coordinator must then conduct a thorough investigation and complete the Report of Testing Irregularity provided through the Online Testing Irregularity Submission System (OTISS). Note that persons reporting irregularities in OTISS must first receive training and have a NC Education user account. The OTISS irregularity report must be submitted to the school system test coordinator within five (5) days of the occurrence. Different incidents must be documented on separate reports of testing irregularities even when the incidents occur during the same test administration in the same room. For example, if one student is disruptive during testing and another student becomes ill during the administration of the same test, two separate reports of testing irregularity must be filed in OTISS. If the superintendent or school system test coordinator declares a misadministration, the misadministration must be documented and reported using appropriate procedures outlined in OTISS. Examples of testing irregularities include, but are not limited to:

i) Eligibility Issues:

- Eligible students were not tested.
- Ineligible students were tested.

ii) Accommodation Issues:

- Approved accommodation not provided
- Approved accommodation not provided appropriately
- Accommodation provided but not approved/documented
- Accommodation Test Read Aloud (in English) or Interpreter/Transliterator Signs/Cues Test provided during the English II test administration

iii) Security Issues:

- Allowing others access to the tests, including school or district personnel who do not have a legitimate need

- Allowing students to review secure test materials before the test administration
- Missing test materials
- Secure test materials not properly returned
- For online testing, failing to maintain security of NC Education username and password
- Failing to store secure test materials in a secure, locked facility
- Failing to cover or remove bulletin board materials, classroom displays, or reference materials (printed or attached) on students' desks that provide information regarding test-taking strategies or the content being measured by the test
- Reproducing items from secure test(s) in any manner or form

- Using items from secure test(s) for instruction
- Failing to return the originally distributed number of test materials to designated school personnel
- Discussing with others any of the test items or information contained in the tests or writing about or posting them on the Internet or on social media sites.

iv) Monitoring Issues:

- Failing to prevent students from cheating by copying, using a cheat sheet, or asking for information
- Failing to prevent students from gaining an unfair advantage through the use of cell phones, text messages, or other means
- Allowing students to remove secure materials from the testing site
- Failing to monitor students and secure test materials during breaks
- For online testing, leaving computers/tablets unsupervised when secure online tests are open and visible
- Leaving the testing room unmonitored when students and secure materials are present

v) Procedural Issues:

- Paraphrasing, omitting, revising, interpreting, explaining, or rewriting the script, directions, or test questions, including answer choices
- Reading or tampering with (e.g., altering, changing, modifying, erasing, deleting, or scoring) student responses to the test questions
- Failing to administer the secure tests on the test date or during the testing window designated by the NCDPI Division of Accountability Services/North Carolina Testing Program
- Failing to follow the test schedule procedures or makeup test schedule designated by the NCDPI Division of Accountability Services/North Carolina Testing Program
- Providing students with additional time beyond the designated maximum time specified in the Administration Guide (except for students with documented special needs requiring accommodations, such as Scheduled Extended Time)
- Test administrator/proctor giving improper assistance or providing instruction related to the concepts measured by the test before the test administration or during the test administration session

vi) Technical Issues:

- Online test connectivity/technical problems
- Online test questions not displaying properly

Note that schools must report online test connectivity and technical problems that occur during the administration of online assessments when a student is not able to successfully complete the assessment. Reports do not need to be entered for students who successfully complete the assessment despite a technical issue. If the same technical problem is being reported for multiple students for the same test administration on the same day, only one OTISS report needs to be submitted. A list of all students affected should be attached to the OTISS report.

School systems must also monitor test administration procedures. According to SBE policy *TEST-001*, if school officials discover any instance of improper administration and determine that the validity of the test results has been compromised, they must (1) "notify" the local board of education, (2) declare a misadministration and (3) order the affected students to be retested. Only the superintendent and the school test coordinator have the authority to declare misadministration at the local level.

## 5.8 Data Forensics Analysis

Maintaining the validity of test scores is essential in any high-stakes assessment program and misconduct represents a serious threat to test score validity. When used appropriately, data forensic analyses can serve as an integral component of a wider test security protocol. The results of these data forensic analyses may be instrumental in identifying potential cases of misconduct for further follow-up and investigation. The possible data forensics analyses on the NCDPI's operational assessments included:

Longitudinal Performance Comparison. The NCDPI psychometricians compare longitudinal performance in terms of mean scale scores and proportion of students in different achievement levels on EOG/EOC assessments across test administrations. Any unusual performance gains may be indicative of possible irregularity issues and may suggest of further exploration.

Testing Outside of the Window Monitoring. Schools are monitored to ensure that all state testing is completed within the state-mandated testing window. The NCDPI has established set dates/windows for all state required testing. If testing occurs outside of the mandated testing window, the school must submit an irregularity report in OTISS.

# CHAPTER 6 SCORING AND SCALE DEVELOPMENT

This chapter describes procedures used by the NCDPI to collect, certify, and score EOG and EOC student responses to create final reportable scale scores. The NCDPI uses a pre-equating model based on an IRT framework for summed score and report them on a common scale. The following procedures and steps are used to ensure student response data are securely and reliably scored so uses and interpretation of EOG and EOC scale scores are valid and fair for all students across the state.

## 6.1 IRT Scoring and Scale Scores

The NCDPI uses IRT summed score procedure for form level scoring and transforming student number correct responses into reportable scale scores. The scoring tables for converting number correct responses into scale scores are generally established after form development and review is complete and before test forms are operationally administered to students. This process of establishing scoring tables for multiple parallel test forms before the forms are administered operationally to students is referred to as a pre-equated scoring model. A pre-equated scoring model has been traditionally used in North Carolina beginning in early 1990s and remained an important feature in the NCDPI grades 3–8 and high school state assessment program. The use of this model allows the NCDPI to take full advantage of test design properties offered through IRT while also allowing for decentralized scoring system based on number correct. Another practical consequence is that the NCDPI can use a short administration window for EOG and EOC that is usually the last 5–10 days of the school year and is still able to provide and use scores for end of year reporting.

## 6.2 Post IRT Calibration

An exception to using pre-equated scoring model based on embedded field-test data for scoring generally occurs during the implementation year of a new assessment edition following revision and adoption of new content standards. During this first year, scoring is done using IRT parameters calibrated from a second round following operational administration of new forms.

There were two main rationales for updating IRT parameters based on operational administration before performing scoring procedures. First, in 2017–18 when newly developed items for *Edition 5* were embedded and administered with *Edition 4* operational forms these new standards may not have been fully implemented in class. Therefore, students may not have the opportunity to learn all new grade-level content standard associated with the new items. As a result, item statistics for some new items, particularly those with revised or new content, were expected to be less reliable between field-test and operational administration. Relying on these embedded field-test IRT parameters under these circumstances for scoring would have resulted in larger-than-

expected measurement error in the final raw-to-scale tables due to instability in IRT parameters and ultimately a violation of the assumption of parameter invariance.

The second rationale was that a post-operational calibration allowed the NCDPI to set a new IRT scale for *Edition 5* forms that will be used to anchor all future forms. This offered a clean separation between the old and new EOG and EOC scales. New parameters from operational administration after students and schools offered opportunity to learn ensured a high degree confidence of parameter invariance for subsequent years.

In summer of 2018 after 100% of student response data was available from *Edition 5* operational forms, the NCDPI perform a second round of calibrations to established final form level IRT parameters for scoring. Classical and IRT form level summary statistics based on post hoc analysis from 2018–19 administration are shown in *Table 6.1 and Table 6.2*. The TCCs and TIFs/CSEMs for these forms associated with updated IRT parameters are shown in *Appendix 6–A.*

*Table 6. 1 Average CTT and IRT Statistics Grades 3–4, 2018–19*

| Grade/Course | Mode | Form | Number of Items | CTT | | IRT | | |
|---|---|---|---|---|---|---|---|---|
| | | | | P-value | Biserial-Corr. | Slope (a) | Threshold (b) | Asymptote (g) |
| 3 | Both | A | 40 | 0.66 | 0.47 | 1.84 | –0.39 | 0.14 |
| | Both | B | 40 | 0.66 | 0.48 | 1.88 | –0.34 | 0.13 |
| | Both | C | 40 | 0.66 | 0.48 | 1.9 | –0.38 | 0.13 |
| 4 | Both | A | 40 | 0.65 | 0.47 | 1.94 | –0.25 | 0.16 |
| | Both | B | 40 | 0.66 | 0.47 | 2.00 | –0.28 | 0.17 |
| | Both | C | 40 | 0.65 | 0.48 | 1.95 | –0.26 | 0.15 |
| 5 | Both | A | 40 | 0.62 | 0.44 | 1.78 | –0.18 | 0.18 |
| | Both | B | 40 | 0.63 | 0.46 | 1.87 | –0.2 | 0.2 |
| | Both | C | 40 | 0.64 | 0.45 | 1.74 | –0.23 | 0.2 |

*Table 6. 2 Average CTT and IRT Statistics Grades 6–8, NC Math 1, and NC Math 3, 2018–19*

| Grade/Course | Mode | Form | Number of Items | CTT | | IRT | | |
|---|---|---|---|---|---|---|---|---|
| | | | | p–value | Biserial–Corr. | Slope (a) | Threshold (b) | Asymptote (g) |
| 6 | Both | A | 45 | 0.58 | 0.47 | 1.97 | 0.06 | 0.21 |
| | PBT | B | 45 | 0.57 | 0.48 | 2.01 | 0.04 | 0.20 |
| | CBT | B | 45 | 0.57 | 0.48 | 2.04 | 0.06 | 0.20 |
| | PBT | C | 45 | 0.58 | 0.47 | 2.06 | 0.05 | 0.21 |
| | CBT | C | 45 | 0.58 | 0.47 | 2.07 | 0.06 | 0.21 |
| 7 | PBT | A | 45 | 0.56 | 0.49 | 2.27 | 0.13 | 0.21 |
| | CBT | A | 45 | 0.56 | 0.49 | 2.27 | 0.14 | 0.21 |
| | PBT | B | 45 | 0.55 | 0.49 | 2.20 | 0.13 | 0.21 |
| | CBT | B | 45 | 0.55 | 0.49 | 2.21 | 0.14 | 0.22 |
| | PBT | C | 45 | 0.56 | 0.49 | 2.12 | 0.09 | 0.22 |
| | CBT | C | 45 | 0.56 | 0.49 | 2.12 | 0.11 | 0.22 |
| 8 | CBT | A | 45 | 0.47 | 0.40 | 1.82 | 0.58 | 0.21 |
| | CBT | B | 45 | 0.46 | 0.38 | 1.71 | 0.57 | 0.18 |
| | CBT | C | 45 | 0.46 | 0.37 | 1.79 | 0.68 | 0.21 |
| NC Math 1 | CBT | M | 50 | 0.50 | 0.47 | 1.90 | 0.32 | 0.19 |
| | CBT | N | 50 | 0.52 | 0.46 | 1.96 | 0.31 | 0.21 |
| NC Math 3 | CBT | M | 50 | 0.40 | 0.45 | 2.14 | 0.79 | 0.15 |
| | CBT | N | 50 | 0.39 | 0.45 | 2.07 | 0.70 | 0.15 |
| | CBT | O | 50 | 0.41 | 0.46 | 2.08 | 0.79 | 0.17 |

## 6.3 IRT Summed Score Procedure

IRT parameters and students' raw responses are used with IRT summed score procedure to create final raw-to-scale conversion tables. During the implementation year, students' scores are delayed until after the standard setting workshop is complete and new performance achievement levels are adopted by the SBE before scores are reported. Two main advantages of using IRT-based scale scores over raw scale for reporting EOG and EOC scores are that:

- They provide a standard metric to report scores from multiple parallel test forms. IRT enables the continuous development and calibration and scoring of new forms on the same existing IRT scale. This allows for the NCDPI to maintain test security by administering new forms without jeopardizing any score comparability.

- Scale scores can be used to minimize differences among various forms and modes of administration of the test. By creating separate raw-to-scale tables for each form, any minor statistical form differences are accounted for and equated. Thus, it makes no difference which form was administered to students.

Estimates of students' proficiency from EOG and EOC assessments are derived from number correct scores using IRT summed score procedure based on expected a posteriori (EAP) theta estimates. These EAP theta estimates are then transformed and reported using an NCDPI custom scale metric. As affirmed in Standard 5.2 (AERA, APA, & NCME, 2014), "*the procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly*" (p.102). This section presents a summary of the IRT summed score procedures used to derive student proficiency estimates from number correct scores. For reference of full description of the IRT summed score procedure see Thissen and Orlando (2001, p.119). For any IRT model with item scores indexed ($u_i$ = 0,1), the likelihood for any summed scores $x = \sum u_i$ is:

$$L_x(\theta) = \sum\nolimits_{\sum u_i = x} L(u/\theta)$$

Where $L(u/\theta) = \prod_i T(u_i/\theta)$ and $T(u_i/\theta)$ is the traceline for response *u* to item *i*. The summation is over all such response patterns that the summed score equals *x*. The probability of each score is:

$$P_x = \int L_x(\theta) d(\theta)$$

And the expected $\theta$ associated with each summed score or expected a posteriori (EAP) scaled score associated with each score is:

$$E(\theta/x) = \frac{\int \theta L_x(\theta) d(\theta)}{P_x}$$

With posterior standard deviation given by

$$SD(\theta/x) = \sum u_i) = \left\{ \frac{\int [\theta - E(\theta/x)]^2 L_x(\theta) d(\theta)}{P_x} \right\}^{1/2}$$

The values computed using *E (θ/x)* may be tabulated and used as the IRT raw-to-scaled score transformation of the summed scores, and the values of *SD (θ/x)* may be used as a standard description of the uncertainty associated with those scaled scores commonly called standard error.

Scoring is done in IRTPRO® using calibrated item parameters to estimate EAP theta scores. To ensure students ability estimates from new parallel forms are placed on a common IRT scale, the population density distribution (mean and standard deviation) of the field-test year is used for scale transformation. For base year forms, the population density is based on estimates from post calibration.

## 6.4 Score Comparability Across Forms and Modes

As presented in Chapters 4 and 5 of this report, the NCDPI administers multiple forms of EOG and EOC during each administration window. For example, during the first administration of *Edition 5* EOG in mathematics the NCDPI administered three base forms in grade 6 across two administration modes: paper and computer. To ensure these alternate forms are statistically equivalent, the DIF sweep procedure described in Chapter 3 is used during concurrent calibration across mode to ensure items flagged as candidate mode DIF have separate item parameters.

Using the IRT summed score procedure, raw-to-scale scores tables are generated separately for each base form. For forms administered across modes, the NCDPI has a minimum participation requirement of about 1,200-1,500 students per field-test item for those items to be calibrated using the mode DIF sweep concurrent calibration procedure. For forms like grade 6 that satisfied the mode DIF sweep calibration during scoring, separate raw-to-scale tables are generated for the different modes. This is one additional measure the NCDPI uses during scoring to statistically adjust for any differences in form difficulty that might be caused by mode of administration.

Using IRT for form development and scoring takes care of the need for any further post equating to adjust for any perceived differences in form difficulty. Because the distribution of students across mode may not necessarily be random, the use of concurrent calibration with the DIF sweep step results in IRT parameters that have been adjusted for differences in mode and are all on the same IRT scale. In the case of separate forms within the same mode, the NCDPI uses separate single group calibration with the same joint population density to ensure all IRT parameters are on a common metric. This allows the NCDPI to have separate raw-to-scale tables for each grade-specific form by mode that are statistically comparable and to provide reliable estimates of student's proficiency on EOG or EOC assessments. This ensures students' scores are fair and can be used to make valid interpretation and uses.

## 6.5 Raw to Scale Scores

The NCDPI administers multiple forms of EOG and EOC within each grade every test cycle. In 2018–19 administration, the NCDPI randomly spiraled three base forms at each grade. The use of multiple pre-equated forms that are randomly spiraled to students within schools across the state offers the following advantages:

- Use of multiple forms and spiraling allows test developers to include items with broader depth of grade-level content standards sampled across forms.
- The availability of multiple forms offers an additional layer of test security. In the event of misadministration students are given an alternate form that has not been previously exposed to them.

The main implication of administering multiple forms within a single administration window is the interpretation of number correct scores commonly referred to as raw scores. Each EOG and EOC form is designed to match the same grade level blueprint but items across forms might have slightly different statistical properties. Separate raw-to-scale tables are created for each form to adjust for minor statistical differences that might exist across forms or between forms across modes. The use of IRT parameters that have been calibrated on a common IRT scale allows the NCDPI to report student performance on a common scale score metric. This common scale score is used to fairly compare student's performance across forms and between years even though students were administered completely different forms.

The raw score metric by itself cannot be used to make any valid interpretation of students' performance. This is because no adjustment is made to the raw score for students taking different forms. This is also true for students taking the same form across mode. Raw scores across forms offer no inherent interpretative meaning of students' performance because the different set of raw scores are not on the same reportable scale. A difference of 1 raw score point between group of students who took different forms or the same form across mode does not imply the student with a higher raw score performed better compared to those with the lower raw score.

The NCDPI only uses raw score in the context to IRT summed score described in section 6.2 above to create raw to scale tables that allows for decentralized scorings. The NCDPI strongly advise against reporting and interpreting raw scores from EOG or EOC assessments. *Table 8.1* through *Table 8.3* in Chapter 8 show summary raw-to-scale ranges for EOG and EOC *Edition 5* forms. These tables should only be used as a reference and part of validity evidence to ensure fairness and transparency in the scoring procedure.

## 6.6 Automated Decentralized Scoring

All items on EOG and EOC mathematics tests are designed to be machine scorable. The NCDPI's reporting group receives answer keys from the TDS verified by Psychometric team and incorporates into WinScan as individual answer key file. At the beginning of each testing window, a new release of WinScan is updated and distributed to all Public School Units with updated raw-to-scale tables. Each version is programmed using the score keys and raw-to-scale score conversion tables for all approved operational test forms.

For paper-based test forms, the school system's test coordinator establishes the schedule for receiving, scanning and scoring EOG/EOC tests at the district level. The school system's test coordinator upon receipt of student response sheets first scans the answer documents and then stores all answer sheets in a secure (locked) facility for six months following the release of test scores. After six months, all student answer sheets are recycled or destroyed in a secure manner in accordance with the NCDPI procedures. The regional accountability coordinator (RAC) and the NCSU-TOPS have the responsibility of scanning and scoring tests for charter schools and for providing long-term storage for specific test materials such as used answer sheets and used test books (e.g., *Student Marks Answers in Test Book* accommodation).

Computer-based forms are administered electronically via a centrally hosted NCSU-TOPS server and scored using the NCDPI managed server. The CBT results are posted by NCSU-TOPS nightly on the NCDPI's secure shell server which the NCSU-TOPS's scripts detect and create files for each Public School Unit with new test results which can be downloaded and imported into WinScan. Prior to the release of final results to schools, test coordinators perform quality control checks. They then provide results (reports) from the test administrations to their respective schools if no error was reported and after the NCDPI confirms its final score certification check was completed. Once the data are available, school system test coordinators can generate school rosters, class rosters and individual reports. Initial district/school-level reporting occurs at the district level. North Carolina Administrative Code (i.e., 16 NCAC 06D .0302) requires districts to report scores resulting from the administration of district-wide and state-mandated tests to students and parents or guardians along with available score interpretation information within 30 days from generation of the score at the district level or from the receipt of the score and interpretive documentation from the department.

## 6.7 Score Certification

Standard 6.9 (AERA, APA, & NCME, 2014) states, "*Those responsible for test scoring should establish and document quality control processes and criteria*" (p. 118). Prior to the release of test scores for official reporting and use for further analyses, the NCDPI performs a final certification to ensure the correct answer key was used in all phases of the scoring to record students' number correct scores. The NCDPI rule of thumb is to perform score certification analyses when 10% of the expected population has tested during the current cycle. The certification process requires the completion of two main quality control steps: In the first step, the psychometric team using the recorded student response data independently tabulates the number correct score at the student level and compares that to the recorded number correct score reported by the scoring software. The goal is to have a 100% agreement rate between scores from the official scoring software and the independent check.

The second step to complete the score certification process involves a sample review of CTT item statistics from operational forms. The goal is to check if current item level CTT statistics

are consistent with theoretical estimates based on field-test data. Forms that have previously been administered are checked against item level data from previous administration. During this step, if the form level statistics differed significantly it is further investigated at item level to make sure the scoring is correct. If any issues are found because of either a wrong scoring key or an improper rendering of any sort, the item is dropped from the form as an operational item and a new raw-to-scale table is generated for that form and the entire scoring procedure is updated with the new data.

Most recently, the NCDPI also used this opportunity to review the scoring keys for open ended "numeric entry" item types. Students' response choices for these items are re-evaluated again to make sure all possible options for example ½, 5/10, or 0.5 are each recorded as correct. If any additional response options are considered, the scoring software is updated, and students' responses are rescored.

Upon completion of score certification analyses, the generated test data are certified as accurate provided that all NCDPI-directed test administration guidelines, rules, procedures and policies have been followed at the district and school levels in conducting proper test administrations and in the generation of the student response data. Finally, the NCDPI issues an official communiqué affirming EOG and EOC scores have been certified and scale scores are approved for official reporting.

# CHAPTER 7 STANDARD SETTING

Standard setting is a process to define levels of achievement or proficiency and the cut scores corresponding to those levels. Standard 5.21 (AERA, APA, & NCME, 2014) states that "*when proposed score interpretation involves one or more cut scores, the rational and procedures used for establishing cut score should be documented*" (p. 107). For the first operational administration of the 2018–19 *Edition 5* mathematics EOG and EOC forms, the NCDPI contracted with the Data Recognition Corp (DRC[1]) to conduct a full standard setting workshop with the main goal of recommending achievement levels and cut scores for the newly developed assessments.

Since achievement levels or cut scores involve high-stakes decision-making including student, teacher and school level accountability, validity of the standard setting process and resulting cut scores is very important. Kane (2001) identified three elements of validity for standard setting: procedural, internal and external. Procedural validity evidence for these studies can be documented through the careful selection of representative, qualified panelists, use of a published standard setting method, completing the study in a systematic fashion and collecting evaluation data that indicates the panelists' confidence in the cut score recommendations they made. Internal validity evidence suggests that panelists had similar expectations for the performance of the target students. This type of evidence is provided by the reasonable standard errors in the recommended cut scores for the second round of the standard setting process. The final type of validity evidence, external, can be provided by triangulation with results from some other estimation of appropriate cut scores from outside the current standard setting process and consideration of other factors that can influence the final policy. The processes and evidences in summarized version of the *Edition 5* mathematics final standard setting are presented in the ensuing sections. A full standard setting technical report produced by DRC[1] can be found in *Appendix 7–A.*

## 7.1 Standard Setting Activities

On July 8–11, 2019, the NCDPI and DRC conducted a standard setting for the North Carolina tests of *Edition 5* general-education mathematics in grades 3–8, NC Math 1 and NC Math 3. The purpose of the standard setting was to develop achievement standards, achievement level descriptors (ALDs), and cut scores associated with four achievement levels: Not Proficient, Level 3, Level 4 and Level 5.

All together there were 60 participants for the standard setting of the general assessment in the first day of the workshop. After the pre-session training, one participant left the workshop with a remaining total of 59. Three panels (grades 3–5, grades 6–8 and NC Math 1 and NC Math 3)

[1]Copyright © 2019 Data Recognition Corp.

with a total of 59 (21 for grade 3–5, 21 for grades 6–8 and 17 for NC Math 1 and NC Math 3) North Carolina mathematics educators convened in Raleigh, North Carolina to make cut score recommendations for the assessments. The item mapping procedure (Lewis, Green, Mitzel, Baum, & Patz, 1998; Mitzel, Lewis, Patz, & Green, 2001) based on ordered item booklets prepared by DRC was used by panelists in a series of rounds to recommend cut scores. All training during the standard setting workshop was facilitated by the DRC staff.

### 7.1.1 Panelists' Backgrounds

*Table 7.1* shows the gender and ethnicity distributions of the workshop participants. As demonstrated by the information provided in these tables, panelists making up the standard setting panels showed representation of diverse gender and ethnic background across North Carolina with majority of the participants as female (85%) and white (68%).

*Table 7. 1 Panelist Gender and Ethnicity*

| N | Gender | | | Ethnicity | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | F | M | NR | AA | AI | HI | NA | WH | MI | NR |
| 60 | 85% | 15% | 0 | 21% | 3% | 5% | 0 | 68% | 0 | 1 |

*F=Female, M=Male, NR=No Response, AA=Black, AI=American Indian, HI=Hispanic, NA=Not Applicable, WH=White and MI=Mixed*

All panelists were asked to provide voluntary experience information. A brief summary of panelists' experiences in terms of years in current position and professional background are presented in *Table 7.2* through *Table 7.4*. *Table 7.2* illustrates the educational experience of the panelists in terms of the years in current position. It shows the experience ranged from less than five (5) years to more than twenty-five (25) years, indicating a very diverse group of highly experienced educators participated in the standard setting workshop.

*Table 7. 2 Panelist Experience as Educators*

| N | Panelists' Year of Experience (%) in Current Position | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1–5 | 6–10 | 11–15 | 16–20 | 21–25 | Over 25 | No Response |
| 60 | 6% | 18% | 23% | 15% | 23% | 11% | 2% |

The panelists' professional backgrounds in terms of teaching diverse group of students are summarized in *Table 7.3* and *Table 7.4*. These tables show that the teachers came from diverse teaching background including teaching general education, special education, ELs, gifted and talented as well as higher education involved in the standard setting.

*Table 7. 3 Panelist Professional Background: Three–Grade Panels*

| N | General Education Teacher | Special Education Teacher | ELL Teachers | Curriculum Staff | Higher Education | Teachers on Special Assignment | Administrator | No Response |
|---|---|---|---|---|---|---|---|---|
| 60 | 48% | 5% | 10% | 26% | 3% | 2% | 3% | 2% |

*Table 7. 4 Panelist Professional Background: Single–Grade Panels*

| N | Special ed. In a self–contained classroom | special ed. In a mainstream classroom | English learners | Gifted and talented ed. | Vocational ed. | Alternate ed. | Adult ed. | No Response |
|---|---|---|---|---|---|---|---|---|
| 60 | 6% | 63% | 56% | 61% | 3% | 2% | 11% | 13% |

### 7.1.2 Opening Session and Introductions

All participants began the workshop with a single opening session for the general and alternate assessments led by the NCDPI. During this session, the director of the NCDPI Accountability Division welcomed the participants to the workshop and described the purpose of the workshop. Subsequently, the chief of Test Development described the recent changes to the North Carolina standards and tests, and how valuable the participating educators' recommendations would be in identifying new cut scores for the tests. Following committee introductions, the three-grade level panels (grades 3–5, grades 6–8 and NC Math 1 and NC Math 3) spent the remainder of Monday, July 8, discussing achievement level descriptors (ALDs) drafted by the NCDPI in consultations with state educators. The ALDs serve as content-oriented statements describing expectations of student performance at each achievement level. Breakout-session facilitators provided panelist with ALD training that covered the purpose of ALDs, and facilitators shared several real-world examples demonstrating characteristics of effective ALDs. Panelists were trained on strategies to link ALDs to the test blueprint and curriculum standards, both of which were made available to panelists. The NCDPI provided policy ALDs for the general mathematics tests in advance of the standard setting workshop, which included general and policy-oriented statements about student achievement across levels. Panelists were tasked with adding content-oriented statements to the draft ALDs to further define student achievement in the context of the assessment. The panels' final drafted ALDs were turned over to the NCDPI for review and future revisions, as deemed necessary.

### 7.1.3 Achievement Level Descriptors

Achievement level descriptors summarize the knowledge, skills and abilities expected of students in each achievement level. Three ALDs generally considered during the standard setting process included policy ALDs, range ALDs and threshold ALDs. The North Carolina ALD development process included drafting the initial ALDs, rounds of webinars, and revisions with the North Carolina educators to finalize it. The descriptions of Not Proficient (Inconsistent Understanding), Level 3 (Sufficient Understanding), Level 4 (Thorough Understanding), and Level 5 (Comprehensive Understanding) are the policy ALDs (*Table 7.5*) for public statements about what and how much North Carolina educators want students to know and be able to do for each grade level in Mathematics.

*Table 7. 5 Policy Achievement Level Descriptors (ALDs) for General Mathematics*

| Not Proficient | Level 3 | Level 4 | Level 5 |
|---|---|---|---|
| Students at the Not Proficient level demonstrate **inconsistent understanding** of grade level content standards and will need support at the next grade/course. | Students at Level 3 demonstrate **sufficient understanding** of grade level content standards though some support may be needed to engage with content at the next grade/course. | Students at Level 4 demonstrate a **thorough understanding** of grade level content standards and are on track for career and college. | Students at Level 5 demonstrate **comprehensive understanding** of grade level content standards, are on track for career and college and are prepared for advanced content at the next grade/course. |

Range ALDs summarize the knowledge, skills and abilities expected of students for a given achievement level on a specific test. The range ALDs show the types of content, as informed by the state content standards, that should be mastered by students in each achievement level on the test at hand. Threshold ALDs are based on the range ALDs and summarize the knowledge, skills and abilities expected of students who are at the point-of-entry (the threshold) of each achievement level. For any given test, these descriptors show the types of skills needed just to be classified (lower bound) in a given achievement level (e.g., just to be classified in Level 3). At the standard setting, participants worked to develop formal range ALDs (on Day 1) and informal threshold ALDs (on Days 2–4). The range ALDs are shown in Section E of the Standard Setting Technical Report (*Appendix 7–A*).

### 7.1.4 Method and Procedure

The Bookmark Standard Setting Procedure (BSSP; Lewis, Mitzel, & Green, 1996; Lewis, Mitzel, Mercado, & Schulz, 2012) was implemented to recommend cut scores for the North Carolina general mathematics tests. This method has been used on assessments in North Carolina and across the nation (Karantonis & Sireci, 2006).

In the Bookmark method, panelists are asked to envision a response probability (RP) criterion and move through a booklet of ordered items based on a RP criterion. The selection of the RP criterion represented a policy decision and the NCDPI chose to apply RP50 to the Ordered Item Booklets for the general mathematics tests, as this criterion (a) could be applied well to grade 8 and NC Math 3; (b) could also be applied to the remaining grades; and (c) allowed for OIBs to be constructed that included a selection of easy, medium and difficult items. The RP50 implies that a minimally competent examinee (MCE) should have at least a 50 percent chance of getting the items correct if the items are before the bookmark and a less than 50 percent chance of getting the items correct if the items are after the bookmark.

A total of 59 North Carolina educators and stakeholders worked individually and collectively to recommend achievement standards for the North Carolina mathematics tests. The achievement standards were approved by the North Carolina State Board of Education on August 8, 2019.

### 7.1.5 Across-Grade Articulation and Final ALD Cuts

Throughout the standard setting process, participants were informed they would have an opportunity at the end of the workshop to consider the across-grade articulation of the achievement standards. Participants were told that achievement standards were well articulated when the impact data associated with a set of cut scores formed a reasonable, explainable pattern across grades. The table leaders were reminded about the caveats regarding the impact data: (a) that 30% of high achieving 8th graders took the NC Math 1 test instead of the Grade 8 test and (b) that the NC Math 3 test was new and students likely knew it did not count for or against their course grades during field-test administration.

During the across-grade articulation, table leaders were assembled in a room and DRC examined the ranges of cut score recommendations made by participants during the standard setting. As described to the table leaders, cut scores adopted within these ranges can be considered as reflecting the voice of the standard setting committee. DRC presented the adjusted cut scores and associated impact data to the table leaders for their inspection. The group saw how the adjustments reflected their opinions about the articulation of the students in Not Proficient and in Level 4 and above. DRC asked the group whether it felt comfortable making this set of adjusted cut scores its recommendation and the table leaders assented. DRC reminded the table leaders that the NCDPI and its advisors would be reviewing their cut score

recommendations and that adjustments may be made to the cut scores by the NCDPI for policy-related reasons.

After the revision, the final ALD Cuts (*Table 7.6*) were presented to the North Carolina State Board of Education on August 7, 2019 for consideration. After deliberation, the SBE approved the cut scores on August 8, 2019.

*Table 7. 6 Final Recommended Cuts and Proficiency Distributions*

| Grade/Course | Recommended Cuts | | | Percent of Students in Each Achievement Level Based on Recommended Cut Scores | | | |
|---|---|---|---|---|---|---|---|
| | Level 3 | Level 4 | Level 5 | Not Proficient | Level 3 | Level 4 | Level 5 |
| 3 | 545 | 551 | 560 | 36% | 20% | 31% | 14% |
| 4 | 547 | 552 | 560 | 43% | 18% | 25% | 15% |
| 5 | 546 | 551 | 561 | 40% | 18% | 31% | 11% |
| 6 | 546 | 551 | 561 | 41% | 17% | 30% | 12% |
| 7 | 546 | 550 | 560 | 42% | 14% | 31% | 13% |
| 8* | 543 | 548 | 555 | 47% | 12% | 10% | 4% |
| NC Math 1 | 548 | 555 | 563 | 45% | 25% | 22% | 8% |
| NC Math 3 | 550 | 556 | 563 | 54% | 20% | 17% | 9% |

**For general mathematics, approximately 27% of students took the NC Math 1 assessment instead of the Grade 8 assessment. These students, typically high achieving, are not included in the Grade 8 population. To help the reader see the trends in the data more easily, the impact data for Grade 8 sum to 73%.*

The raw score ranges for the proficiency levels from 2018-19 population are shown in *Table 7.7*. Notice that, for grade 3 the upper range of Level 4 and Lower range of Level 5 are overlapped. The same observation can be made for other grades/levels. The overlap is due to the fact that the raw scores corresponding to the scale score cuts across forms for a given grade/level vary slightly, mostly one scale score point. This overlapping feature of the raw scores that could potentially mislead to end users is a primary reason for reporting scale score only at the student level.

*Table 7. 7 Raw Score Ranges Across Proficiency Levels, 2018-19*

| Grade/Level | Not Proficient | | Level 3 | | Level 4 | | Level 5 | |
|---|---|---|---|---|---|---|---|---|
| | Min | Max | Min | Max | Min | Max | Min | Max |
| 3 | 0 | 20 | 21 | 27 | 28 | 36 | 36 | 40 |
| 4 | 0 | 22 | 22 | 28 | 28 | 36 | 36 | 40 |
| 5 | 0 | 20 | 20 | 26 | 26 | 35 | 35 | 40 |
| 6 | 0 | 19 | 20 | 26 | 27 | 38 | 39 | 45 |
| 7 | 0 | 18 | 18 | 24 | 24 | 38 | 38 | 45 |
| 8 | 0 | 22 | 22 | 28 | 28 | 36 | 35 | 45 |
| NC Math 1 | 0 | 20 | 20 | 31 | 31 | 42 | 42 | 50 |
| NC Math 3 | 0 | 17 | 18 | 25 | 26 | 36 | 37 | 50 |

## 7.2 Evaluation of the Standard Setting Workshop

Since standard setting process incorporates subjective judgement, it is important to document procedural validation including selection of the experts, experts' clarity of the standard setting method and their judgement, i.e., the extent to which they understand the standard setting procedure and their confidence in the cut scores. Sections below summarize the participants' evaluation of the process as well as evaluation of the processes by the external evaluator.

### 7.2.1 Participants' Evaluation

At the end of the workshop, a participant survey was conducted for their perceived validity of the workshop and their recommendations as a part of the post-session workshop evaluation. Such evaluations are important evidence for establishing the validity of performance levels (Hambleton, 2001). The survey results are presented in *Table 7.8*. Generally, 95% or higher proportion of participants were satisfied (Agree + Strongly Agree) with their recommendations and with the workshop. The results further indicated that 100% of the participants considered the threshold students when making benchmarks. They agreed that the final recommended cut scores reflected the work of their group.

*Table 7. 8 Standard Setting Workshop Evaluation Results*

| Statement | Strongly Disagree | Disagree | Agree | Strongly Agree | Agree + Strongly Agree |
|---|---|---|---|---|---|
| The training provided a clear description of the workshop goals. | 1% | 1% | 45% | 50% | 95% |
| I understood how to make my bookmarks. | 0% | 1% | 35% | 62% | 97% |
| I considered the threshold students when making my bookmarks. | 0% | 0% | 31% | 69% | 100% |
| Discussing the threshold students helped me make my bookmarks. | 0% | 0% | 46% | 54% | 100% |
| My group's work was reflected in the presentation of recommendations across grades. | 2% | 0% | 47% | 51% | 98% |
| Overall, I valued the workshop as a professional development experience. | 0% | 2% | 13% | 85% | 98% |

### 7.2.2 External Evaluation

In order to implement and evaluate any deviations from the standard setting processes by the vendor, the NCDPI contracted Dr. Gregory J. Cizek as an external independent evaluator of the mathematics standard setting workshop. Dr. Cizek is an expert in the field and is also a member of the North Carolina Technical Advisory Committee (NCTAC). His report regarding the standard setting workshop is summarized below. The detail report is available in *Appendix 7–B.*

Dr. Cizek reported that qualified educators from North Carolina were trained in the methods and led through the standard setting procedures by content and process specialists. The participants' judgments were solicited in two ways: they first generated exclusively content-based judgments and cut scores across three rounds of judgments in Phase I of the standard setting workshop; they next adjusted the system of recommended cut scores in cross-grade articulation sessions in Phase II of the workshop.

Dr. Cizek concluded that “the workshop recommended cut score can be considered to be valid and reliable estimates of appropriate performance standards for the relevant assessments. Unless the panelists’ evaluations indicate otherwise, policy makers should have confidence that the recommendations from the standard setting activity are based on sound procedures, producing credible, defensible, and educationally useful results.”

# CHAPTER 8 TEST RESULTS AND REPORTS

This chapter presents test level summary results for the EOG and EOC mathematics tests based on reported scale scores and achievement levels from 2018–19 operational administration. The chapter is divided into three main sections. Section 8.1 highlights descriptive summary results of scale scores overall and by major demographic subgroups including accommodations, gender, ethnicity, and mode as well as overall achievement level distributions for EOG and EOC forms. Section 8.2 briefly describes types of reports the NCDPI produces including those at class, school, district, and state level to share and interpret assessments results with stakeholders. Section 8.3 elaborates confidentiality requirements for sharing or reporting students' personal information as well as student data.

## 8.1 EOG and EOC Scale Score Distribution

Scale score distributions from the first operational administration of *Edition 5* EOG and EOC mathematics assessments from 2018–19 are summarized in *Figure 8.1* through *Figure 8.8*. These scores are based on results from all eligible students enrolled at the respective grade level for EOG or course for EOC. Results include both general administration and students with approved NCDPI's accommodations such as braille, large print, read-aloud and extended time.

One significant change from previous editions of EOG and EOC to *Edition 5* is the definition of the grade 8 EOG population. Beginning from 2017–18, the NCDPI stopped double testing grade 8 students who were enrolled in the NC Math 1 course. These grade 8 students are now required to take only the EOC NC Math 1 assessment. The subset of grade 8 students who take NC Math 1 generally make up about 30% of the total grade 8 population and constitute above 90% of top performing grade 7 students from the previous year's EOG. The ability distribution of current EOG grade 8 is significantly different from EOG in grades 3 through 7.

For EOG grades 3 through 7, NC Math 1 and NC Math 3, the population scale score mean was set to 550 with a standard deviation of 10. The scale score mean for the adjusted grade 8 population was set to 540 with a standard deviation of 10. This was done to make a distinction in reporting and score interpretation between grade 8 and other EOG grades. *Edition 5* mathematics scale scores are not reported using a vertical scale. Any across-grade scale score interpretations and comparisons are highly discouraged as each EOG assessment is aligned to grade level specific content standards.

*Figure 8. 1 Grade 3 Mathematics Scale Score Distribution, Spring 2019*

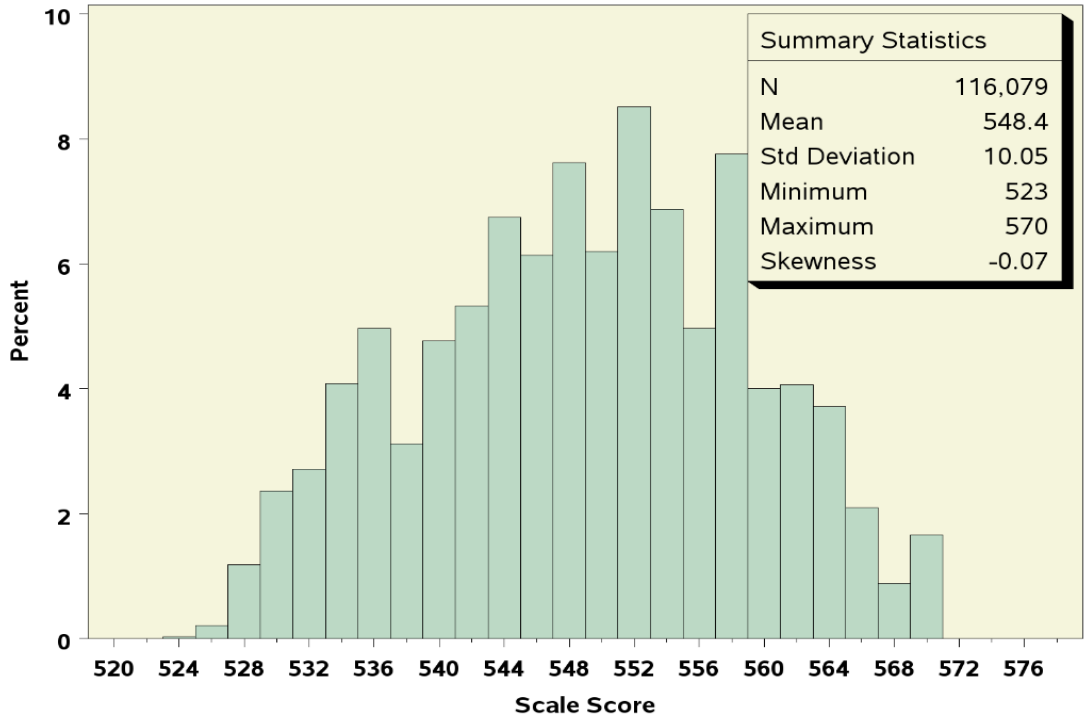



*Figure 8. 2* Grade 4 Mathematics Scale Score Distribution, Spring 2019

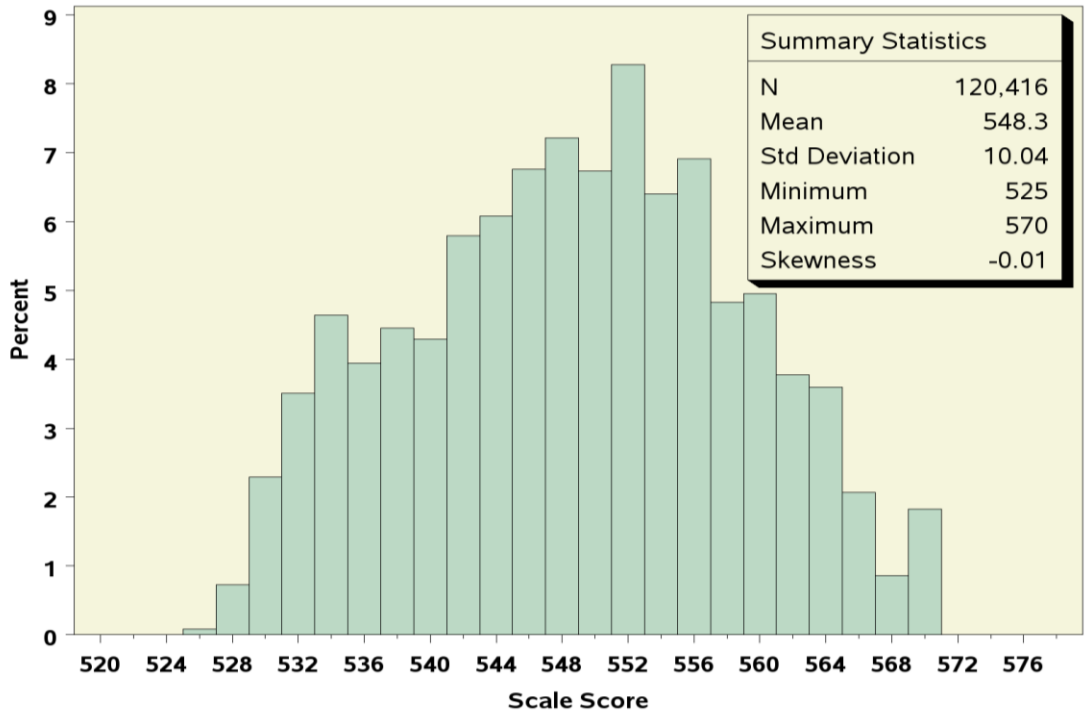


*Figure 8. 3* Grade 5 Mathematics Scale Score Distribution, Spring 2019

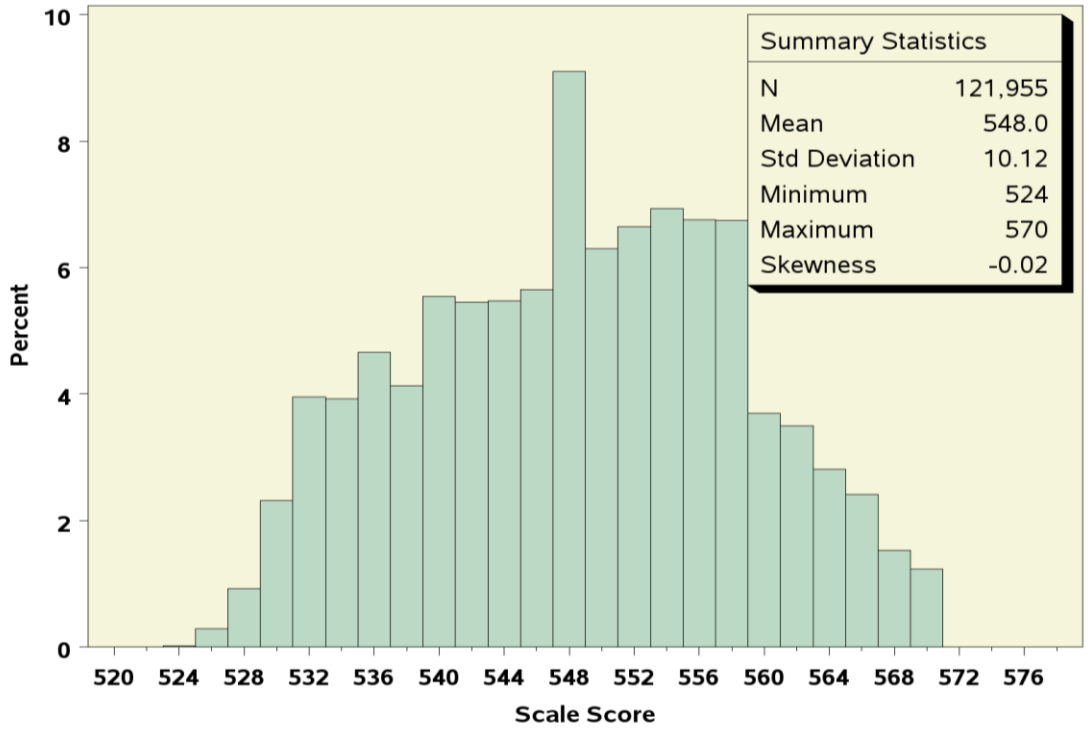



*Figure 8. 4* Grade 6 Mathematics Scale Score Distribution, Spring 2019

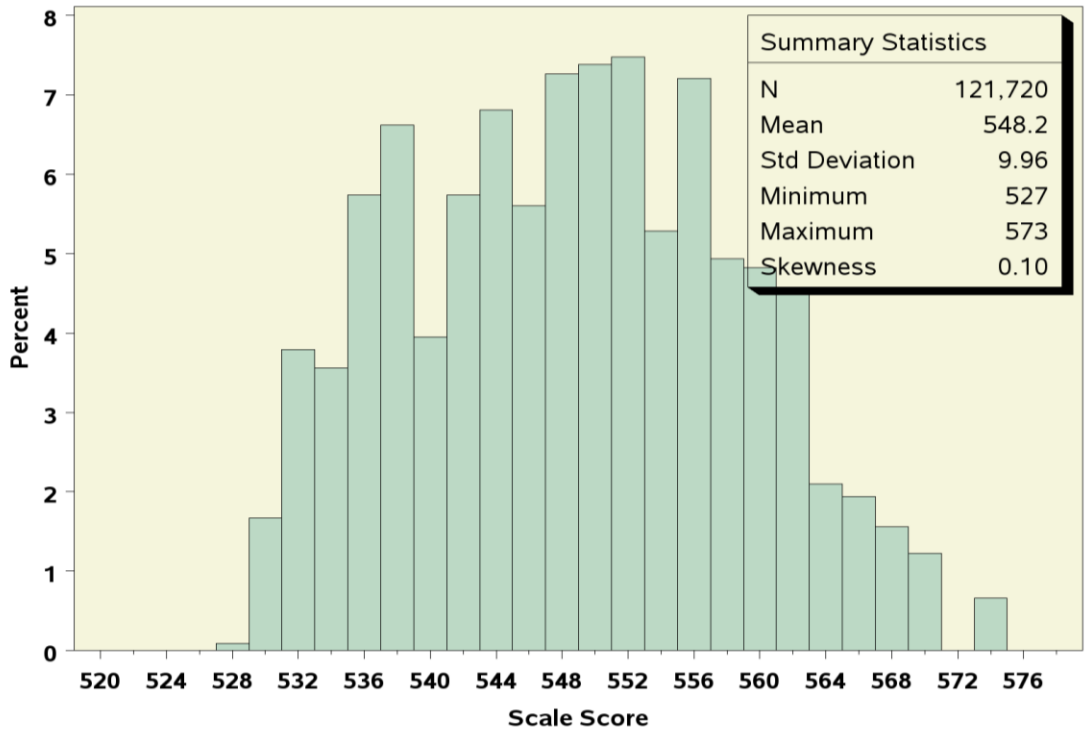


*Figure 8. 5* Grade 7 Mathematics Scale Score Distribution, Spring 2019

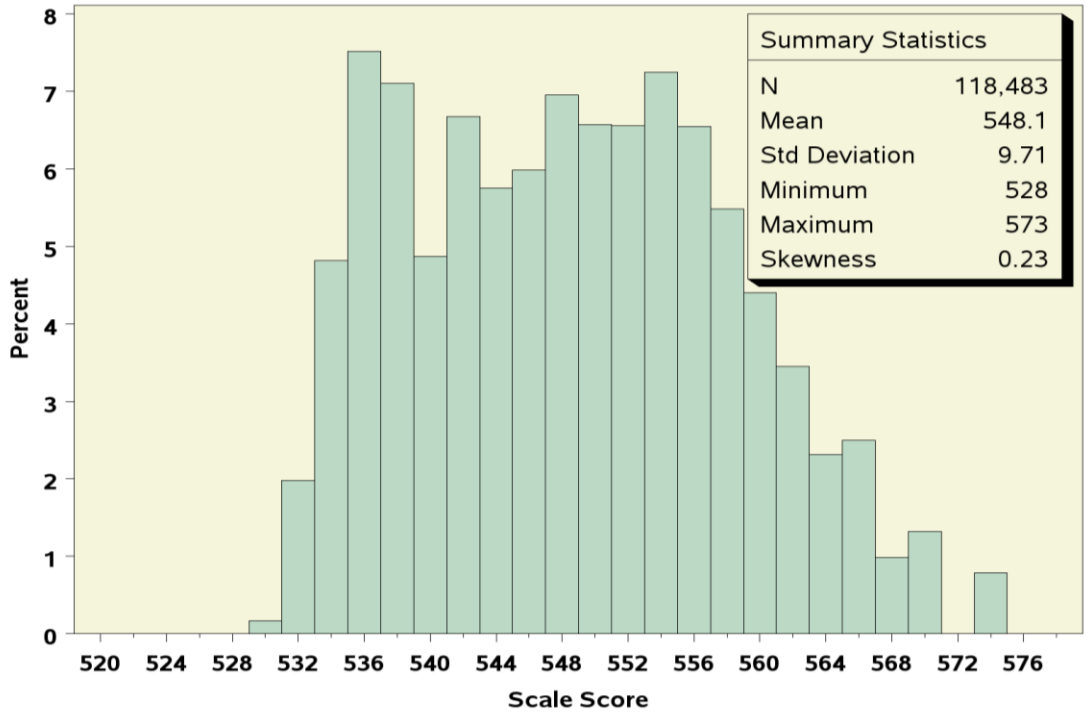



*Figure 8. 6* Grade 8 Mathematics Scale Score Distribution, Spring 2019

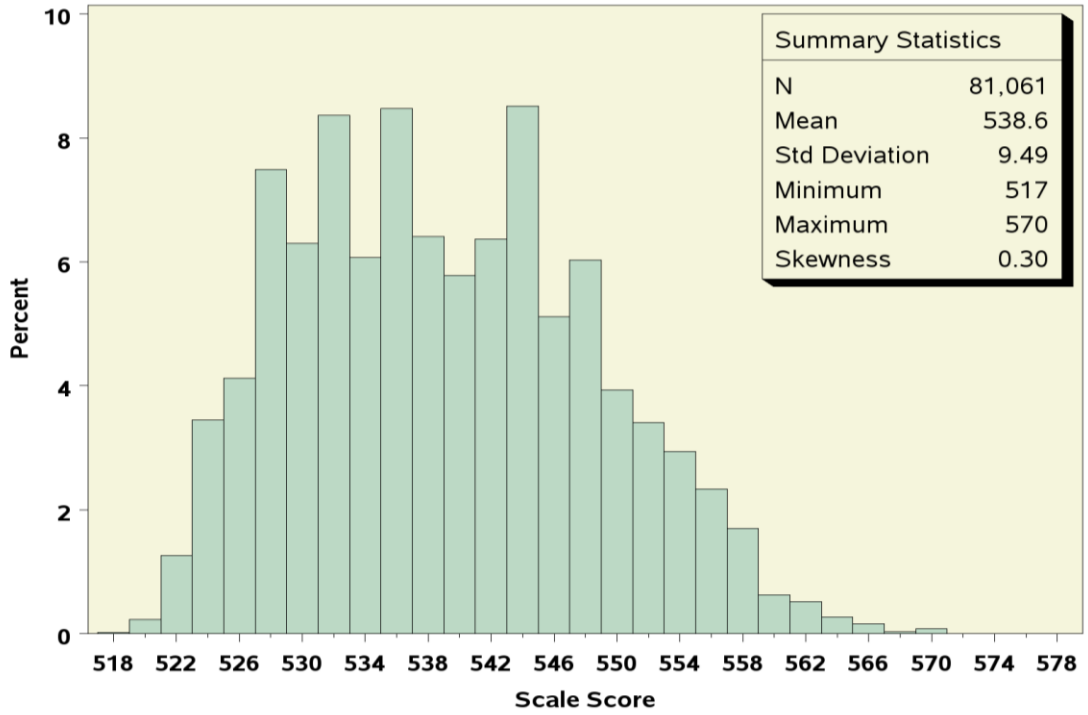


*Figure 8. 7* NC Math 1 Scale Score Distribution, Spring 2018-19

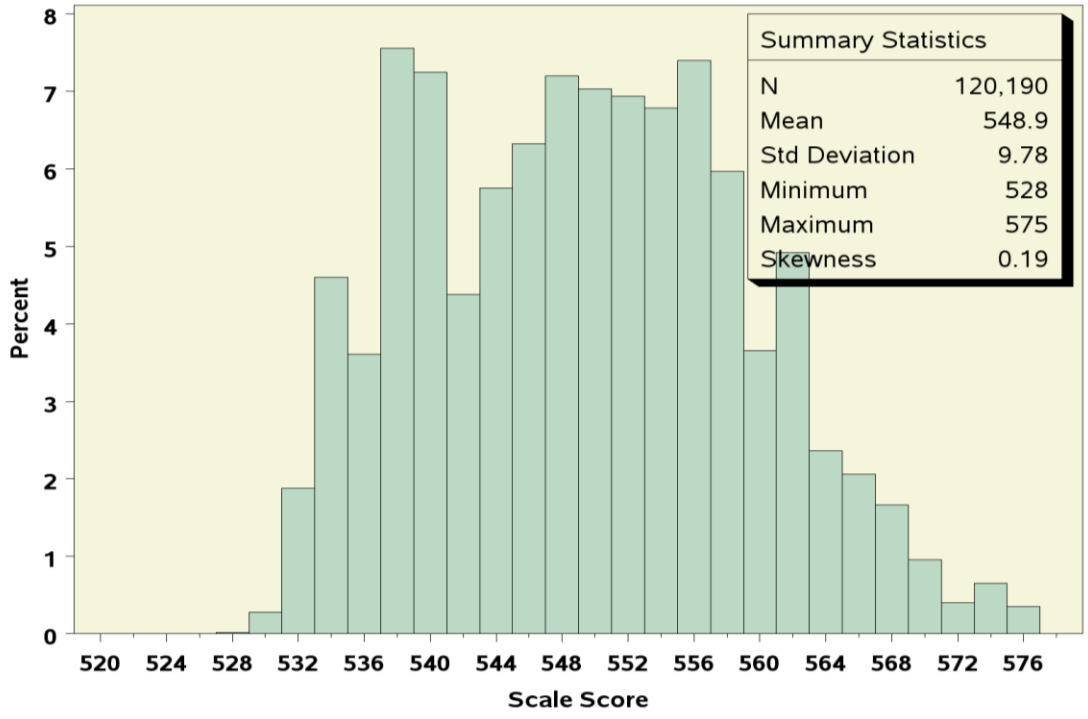
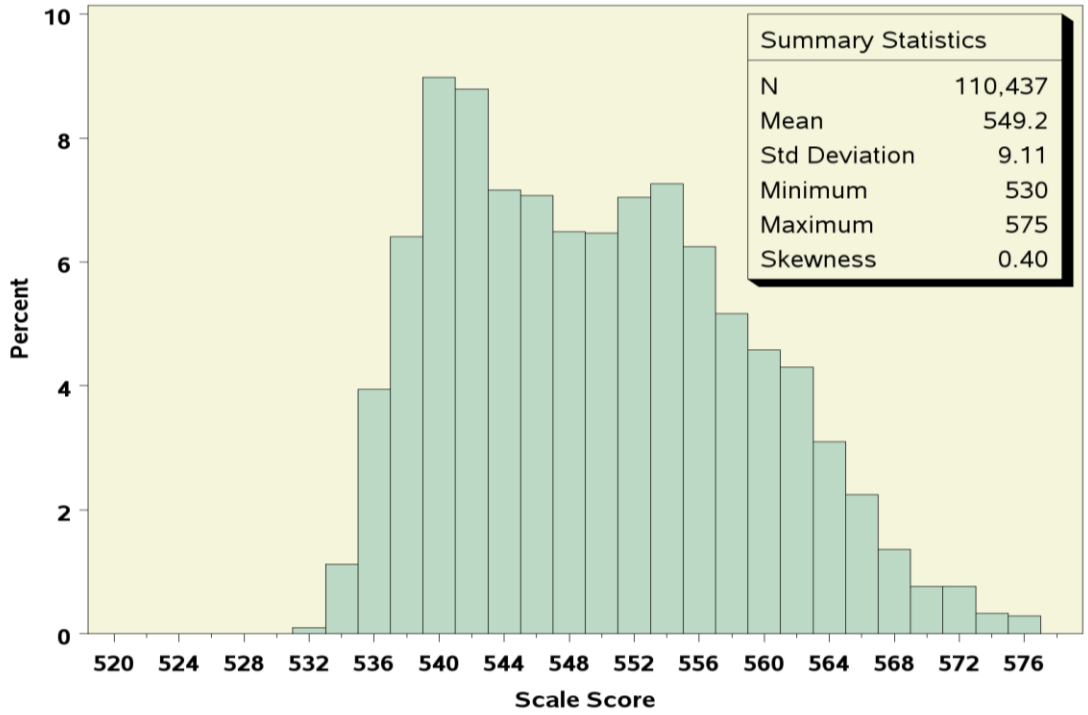



*Figure 8. 8* NC Math 3 Scale Score Distribution, Spring 2018-19



### 8.1.1 Scale Score by Accommodation Subgroups

The NCDPI allows the use of various types of accommodations with EOG and EOC assessments to ensure accessibility to all students. Students with IEPs can access their required accommodations described in Chapter 5 at any time during test administration. Research in measurement literature has demonstrated that these standard accommodations do not measure any significant construct irrelevant variance to students reported scores. Thus, results from students who received any of these approved accommodations are included in the general administration and the same inferences are made about student's performance. *Tables 8.1* through *Table 8.3* show the summary score distributions for EOG and EOC mathematics from 2018–19 administration by major accommodation subgroups.

*Table 8.1* and *Table 8.2* show the scale score summary results for Elementary and Middle Schools by accommodation subgroups and *Table 8.3* shows the results for NC Math 1 and NC Math 3. "Regular Administration" in these tables refers to students who did not receive any NCDPI approved accommodations. Each accommodation category includes all students who received one or more accommodation classified under this category in Section 5.5. For example, "Special Print" includes all students who received Braille or Large Print Edition (not for online) or One Test Item Per Page Edition (not for online).

*Table 8. 1 Grades 3–5 Mathematics Scale Score by Accommodation Subgroups, Spring 2019*

| Grade | Subgroups | N | Statistics | | Range | | Percentile | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | SD | Min | Max | 25th | Median | 75th |
| 3 | Regular Administration | 96,357 | 550 | 9.5 | 523 | 570 | 543 | 551 | 557 |
| | Assistive Devices | 17,021 | 540 | 8.2 | 523 | 570 | 533 | 538 | 545 |
| | Special Environment | 2,311 | 546 | 9.7 | 523 | 570 | 538 | 546 | 552 |
| | Special Print | 390 | 547 | 10.7 | 523 | 570 | 539 | 548 | 555 |
| | All | 116,079 | 548 | 10.1 | 523 | 570 | 541 | 549 | 556 |
| 4 | Regular Administration | 98,851 | 550 | 9.5 | 525 | 570 | 543 | 551 | 557 |
| | Assistive Devices | 18,248 | 539 | 7.7 | 525 | 570 | 533 | 538 | 545 |
| | Special Environment | 2,782 | 546 | 9.4 | 525 | 570 | 538 | 546 | 553 |
| | Special Print | 535 | 550 | 11.0 | 526 | 570 | 542 | 552 | 559 |
| | All | 120,416 | 548 | 10.0 | 525 | 570 | 541 | 549 | 556 |
| 5 | Regular Administration | 101,330 | 550 | 9.5 | 524 | 570 | 543 | 550 | 557 |
| | Assistive Devices | 17,380 | 538 | 7.5 | 524 | 570 | 532 | 536 | 543 |
| | Special Environment | 2,858 | 545 | 9.3 | 525 | 570 | 537 | 545 | 551 |
| | Special Print | 387 | 547 | 10.8 | 526 | 570 | 537 | 547 | 555 |
| | All | 121,955 | 548 | 10.1 | 524 | 570 | 540 | 548 | 556 |

*Table 8. 2 Grades 6–8* Mathematics *Scale Score by Accommodation Subgroups, Spring 2019*

| Grade | Subgroups | N | Statistics | | Range | | Percentile | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | SD | Min | Max | 25th | Median | 75th |
| 6 | Regular Administration | 103,813 | 550 | 9.5 | 527 | 573 | 543 | 550 | 557 |
| | Assistive Devices | 14,790 | 538 | 6.6 | 527 | 573 | 533 | 537 | 542 |
| | Special Environment | 2,973 | 545 | 9 | 528 | 573 | 537 | 544 | 551 |
| | Special Print | 144 | 543 | 10 | 528 | 570 | 535 | 541 | 552 |
| | All | 121,720 | 548 | 10 | 527 | 573 | 540 | 548 | 555 |
| 7 | Regular Administration | 101,838 | 550 | 9.4 | 528 | 573 | 542 | 550 | 556 |
| | Assistive Devices | 13,550 | 539 | 5.8 | 528 | 573 | 535 | 537 | 541 |
| | Special Environment | 2,942 | 544 | 8.8 | 529 | 573 | 537 | 543 | 550 |
| | Special Print | 153 | 544 | 9.1 | 529 | 573 | 536 | 543 | 550 |
| | All | 118,483 | 548 | 9.7 | 528 | 573 | 540 | 548 | 555 |
| 8 | Regular Administration | 66,859 | 540 | 9.3 | 517 | 570 | 533 | 540 | 547 |
| | Assistive Devices | 11,463 | 531 | 7.1 | 518 | 566 | 526 | 529 | 535 |
| | Special Environment | 2,573 | 536 | 9 | 518 | 569 | 529 | 535 | 543 |
| | Special Print | 166 | 538 | 9.7 | 522 | 558 | 531 | 538 | 545 |
| | All | 81,061 | 539 | 9.5 | 517 | 570 | 531 | 538 | 545 |

*Table 8. 3 NC Math 1 and NC Math 3 Scale Score by Accommodation Subgroups, Spring 2019*

| Grade | Subgroups | N | Statistics | | Range | | Percentile | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | SD | Min | Max | 25th | Median | 75th |
| NC Math 1 | Regular Administration | 104,428 | 550 | 9.5 | 528 | 575 | 543 | 550 | 557 |
| | Assistive Devices | 11,656 | 539 | 6 | 528 | 575 | 535 | 538 | 542 |
| | Special Environment | 3,902 | 544 | 8.3 | 529 | 575 | 537 | 542 | 549 |
| | Special Print | 204 | 546 | 9.7 | 530 | 568 | 538 | 544 | 554 |
| | All | 120,190 | 549 | 9.8 | 528 | 575 | 540 | 549 | 556 |
| NC Math 3 | Regular Administration | 101,876 | 550 | 9.1 | 530 | 575 | 542 | 549 | 556 |
| | Assistive Devices | 4,745 | 542 | 5.6 | 531 | 575 | 538 | 540 | 544 |
| | Special Environment | 3,626 | 545 | 7.6 | 532 | 575 | 539 | 543 | 549 |
| | Special Print | 190 | 549 | 8.7 | 534 | 572 | 541 | 549 | 555 |
| | All | 110,437 | 549 | 9.1 | 530 | 575 | 541 | 548 | 556 |

These results show that scale score distributions from regular administration have similar distributional properties to the scaling parameters with mean of 550 and standard deviation of approximately 10. For all grades, Assistive Devices, which includes all the read-aloud accommodation formats, was the, most used accommodation category. The average scale score for this category was about one standard deviation lower than the population average score. Preliminary results from data analysis and field investigations conducted by the NCDPI confirm that for the most part schools tend to encourage their lowest-performing students to access Read Aloud accommodation when available. Their intent is to remove non mathematics construct barriers that might otherwise impede access to EOG and EOC assessments for these students.

### 8.1.2 Scale Score by Gender

*Table 8.4* through *Table 8.6* summarize EOG and EOC mathematics scale score by gender. In all grade levels, there were slightly higher proportion of male students (about 51%) who took EOG and EOC mathematics in North Carolina during 2018–19 school year. Scale score distributions are similar between female and male students for the most part. In grades 3 and 4, male students on average performed slightly better than female students. However, for grades 5–8 and EOC NC Math 1 and NC Math 3, female students on average performed slightly better than male students.

*Table 8. 4 Grades 3–5 Mathematics Scale Score Descriptive Summary by Sex, Spring 2019*

| Grade | Gender | N | Statistics | | Range | | Percentile | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | SD | Min | Max | 25th | Median | 75th |
| 3 | Female | 56,716 | 548 | 9.8 | 523 | 570 | 541 | 549 | 555 |
| | Male | 59,340 | 549 | 10.3 | 523 | 570 | 541 | 549 | 557 |
| | All | 116,056 | 548 | 10.1 | 523 | 570 | 541 | 549 | 556 |
| 4 | Female | 58,523 | 548 | 9.8 | 525 | 570 | 541 | 548 | 555 |
| | Male | 61,766 | 549 | 10.3 | 525 | 570 | 541 | 549 | 556 |
| | All | 120,289 | 548 | 10.0 | 525 | 570 | 541 | 549 | 556 |
| 5 | Female | 59,682 | 548 | 9.8 | 524 | 570 | 541 | 548 | 555 |
| | Male | 62,251 | 548 | 10.4 | 524 | 570 | 540 | 548 | 556 |
| | All | 121,933 | 548 | 10.1 | 524 | 570 | 540 | 548 | 556 |

*Table 8. 5 Grades 6–8 Mathematics Scale Score Descriptive Summary by Sex, Spring 2019*

| Grade | Sex | N | Statistics | | Range | | Percentile | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | SD | Min | Max | 25th | Median | 75th |
| 6 | Female | 59,274 | 548 | 9.77 | 527 | 573 | 541 | 549 | 555 |
| | Male | 62,331 | 548 | 10.1 | 527 | 573 | 539 | 548 | 555 |
| | All | 121,605 | 548 | 9.96 | 527 | 573 | 540 | 548 | 555 |
| 7 | Female | 57,877 | 549 | 9.56 | 529 | 573 | 541 | 548 | 555 |
| | Male | 60,585 | 548 | 9.83 | 528 | 573 | 539 | 547 | 555 |
| | All | 118,462 | 548 | 9.71 | 528 | 573 | 540 | 548 | 555 |
| 8 | Female | 38,627 | 539 | 9.41 | 517 | 570 | 532 | 539 | 546 |
| | Male | 42,318 | 538 | 9.5 | 517 | 570 | 530 | 537 | 544 |
| | All | 80,945 | 539 | 9.49 | 517 | 570 | 531 | 538 | 545 |

*Table 8. 6 NC Math 1 and NC Math 3 Scale Score Descriptive Summary by Sex, Spring 2019*

| Grade | Sex | N | Statistics | | Range | | Percentile | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | SD | Min | Max | 25th | Median | 75th |
| NC Math 1 | Female | 57,988 | 550 | 9.5 | 528 | 575 | 542 | 550 | 556 |
| | Male | 62,026 | 548 | 9.9 | 528 | 575 | 540 | 548 | 555 |
| | All | 120,014 | 549 | 9.8 | 528 | 575 | 540 | 549 | 556 |
| NC Math 3 | Female | 55,092 | 550 | 8.9 | 530 | 575 | 542 | 549 | 556 |
| | Male | 55,203 | 549 | 9.3 | 530 | 575 | 541 | 548 | 556 |
| | All | 110,295 | 549 | 9.1 | 530 | 575 | 541 | 548 | 556 |

### 8.1.3 Scale Score by Major Ethnic Groups

*Table 8.7* through *Table 8.9* show the breakdown of EOG and EOC mathematics scale scores by major reportable ethnic groups from 2018–19 administration. For the purpose of this report, scale scores are summarized only for students self-reported to belong in one of these major ethnic groups: Black, Hispanic, and White. All students not self-identified in any of those three major groups are classified as Other. The distribution of North Carolina student population is very consistent across grade levels with White students representing about 46% of students across all levels and Black students representing about 25% with Hispanic students making about 20% of all student. Scale score distribution by these major ethnic groups show in all grades White students have the highest average scale scores compared to Black students with the lowest average scale scores. The average scale score difference ranged from 0.5 to 0.8 standard deviation across all EOG and EOC grades.

*Table 8. 7 Grades 3–4 Mathematics Scale Score Descriptive Summary by Ethnicity, Spring 2019*

| Grade | Race | N | Statistics | | Range | | Percentile | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | SD | Min | Max | 25th | Median | 75th |
| 3 | Black | 29,136 | 544 | 9.2 | 523 | 570 | 536 | 544 | 551 |
| | Hispanic | 22,341 | 546 | 9.5 | 523 | 570 | 539 | 546 | 553 |
| | White | 53,176 | 551 | 9.5 | 523 | 570 | 545 | 552 | 558 |
| | Others | 11,403 | 551 | 10.5 | 523 | 570 | 543 | 552 | 559 |
| | All | 116,056 | 548 | 10.1 | 523 | 570 | 541 | 549 | 556 |
| 4 | Black | 30,768 | 544 | 8.9 | 525 | 570 | 536 | 543 | 550 |
| | Hispanic | 23,125 | 546 | 9.4 | 525 | 570 | 539 | 547 | 553 |
| | White | 55,033 | 551 | 9.5 | 525 | 570 | 545 | 552 | 558 |
| | Others | 11,363 | 551 | 10.7 | 525 | 570 | 543 | 551 | 559 |
| | All | 120,289 | 548 | 10.0 | 525 | 570 | 541 | 549 | 556 |
| 5 | Black | 30,892 | 543 | 9.0 | 524 | 570 | 536 | 543 | 550 |
| | Hispanic | 23,425 | 546 | 9.4 | 524 | 570 | 539 | 546 | 553 |
| | White | 56,183 | 551 | 9.7 | 524 | 570 | 545 | 552 | 558 |
| | Others | 11,433 | 550 | 10.8 | 524 | 570 | 542 | 551 | 558 |
| | All | 121,933 | 548 | 10.1 | 524 | 570 | 540 | 548 | 556 |

*Table 8. 8 Grades 6–8 Mathematics Scale Score Descriptive Summary by Ethnicity, Spring 2019*

| Grade | Race | N | Statistics | | Range | | Percentile | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | SD | Min | Max | 25th | Median | 75th |
| 6 | Black | 30,370 | 543 | 8.6 | 527 | 573 | 536 | 543 | 550 |
| | Hispanic | 23,531 | 546 | 9.1 | 527 | 573 | 538 | 546 | 552 |
| | White | 56,426 | 551 | 9.5 | 527 | 573 | 544 | 552 | 558 |
| | Others | 11,278 | 551 | 10.8 | 528 | 573 | 542 | 551 | 559 |
| | All | 121,605 | 548 | 10.0 | 527 | 573 | 540 | 548 | 555 |
| 7 | Black | 29,567 | 543 | 8.2 | 529 | 573 | 536 | 542 | 549 |
| | Hispanic | 22,316 | 546 | 8.8 | 529 | 573 | 538 | 545 | 552 |
| | White | 55,746 | 551 | 9.4 | 528 | 573 | 544 | 551 | 558 |
| | Others | 10,833 | 551 | 10.8 | 528 | 573 | 542 | 551 | 559 |
| | All | 118,462 | 548 | 9.7 | 528 | 573 | 540 | 548 | 555 |
| 8 | Black | 23,437 | 536 | 8.5 | 517 | 566 | 529 | 534 | 542 |
| | Hispanic | 16,695 | 538 | 9.2 | 517 | 570 | 530 | 537 | 544 |
| | White | 34,749 | 541 | 9.5 | 517 | 570 | 534 | 541 | 548 |
| | Others | 6,064 | 539 | 9.6 | 520 | 570 | 531 | 538 | 546 |
| | All | 80,945 | 539 | 9.5 | 517 | 570 | 531 | 538 | 545 |

*Table 8. 9 EOC Mathematics Scale Score Descriptive Summary by Ethnicity, 2018-19*

| EOC | Race | N | Statistics | | Range | | Percentile | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | SD | Min | Max | 25th | Median | 75th |
| NC Math 1 | Black | 29,804 | 545 | 8.4 | 528 | 575 | 538 | 543 | 551 |
| | Hispanic | 21,431 | 547 | 9.1 | 528 | 575 | 539 | 546 | 553 |
| | White | 58,133 | 552 | 9.4 | 528 | 575 | 545 | 552 | 558 |
| | Others | 10,646 | 552 | 10.9 | 528 | 575 | 543 | 552 | 560 |
| | All | 120,014 | 549 | 9.8 | 528 | 575 | 540 | 549 | 556 |
| NC Math 3 | Black | 26,893 | 545 | 7.4 | 530 | 575 | 539 | 543 | 550 |
| | Hispanic | 17,859 | 547 | 8.2 | 531 | 575 | 541 | 546 | 553 |
| | White | 55,851 | 551 | 9.1 | 530 | 575 | 544 | 551 | 558 |
| | Others | 9,692 | 552 | 10.4 | 531 | 575 | 543 | 551 | 559 |
| | All | 110,295 | 549 | 9.1 | 530 | 575 | 541 | 548 | 556 |

The scale score differences represented in *Table 8.7* through *Table 8.9* are not an indication that EOG or EOC assessments are biased across ethnic groups. All EOG and EOC items were thoroughly vetted throughout several phases of item development, field test and item analysis by different experts to ensure operational EOG and EOC mathematics items did not exhibit any potential inference of bias or DIF for any student subgroup. The descriptive statistics of scale scores by other subgroups (EDS, SWD, and ELs) are shown in *Appendix 8-A*.

### 8.1.4 Scale Score by Mode

The NCDPI is in a transition period for EOG and EOC mathematics assessments from paper only administration to computer-based fixed form administration. During this period beginning with *Edition 5* mathematics assessments all EOG forms were designed to be administered in both paper and computer mode. Score comparability across mode is an important validity issue for score interpretation. It is important that the NCDPI demonstrate adequate empirical evidence that regardless of mode of administration students with the same expected ability level will have the same expected scale score across different administration modes.

During item analysis and form development, all items were checked for mode DIF to ensure any item exhibiting substantial mode DIF was not place on a final EOG or EOC form. Also, during scoring separate raw-to-scale score tables (see Chapter 4) were created for the same form across mode to statistically adjust for small differences on items attributed to administration mode. This ensured any item that functions differentially across mode is properly accounted for in the raw to scale conversion table and variability of test scores are not due to construct irrelevant variance from mode of administration. This is the important technical reason why the NCDPI no longer

publishes raw scores for EOG and EOC. These raw scores do not have the same interpretive meaning across forms and mode of administration and pose a serious validity concern to score interpretation. An unadjusted total raw score for students who took different forms across mode are not comparable.

Scale scores on the other hand have been adjusted to account for statistical differences across forms and mode so performance is on the same scale for all students. Scores can then be interpreted without distinction of form or mode administered. A difference of one scale score point has the same meaning for all students within a grade.

*Figure 8.9* though *Figure 8.16* show the scale score distributions from EOG and EOC by mode of administration. For the most part, the samples of students by test mode is not random. Schools have the option to move students to computer mode when they deem that they have the necessary technical capabilities and students are prepared for that transition. In 2018–19, EOC NC Math 1 and NC Math 3 were administered computer only. Exceptions were given to schools and students with technical challenges or special accommodation requests.

Overall, scale score distributions between modes are very similar. Noticeable observed differences in scale score distributions are explained by differences in student ability distribution between mode since schools and students were self-selected by mode.

*Figure 8. 9 Grade 3 Mathematics Scale Score Distributions by Mode, Spring 2019*

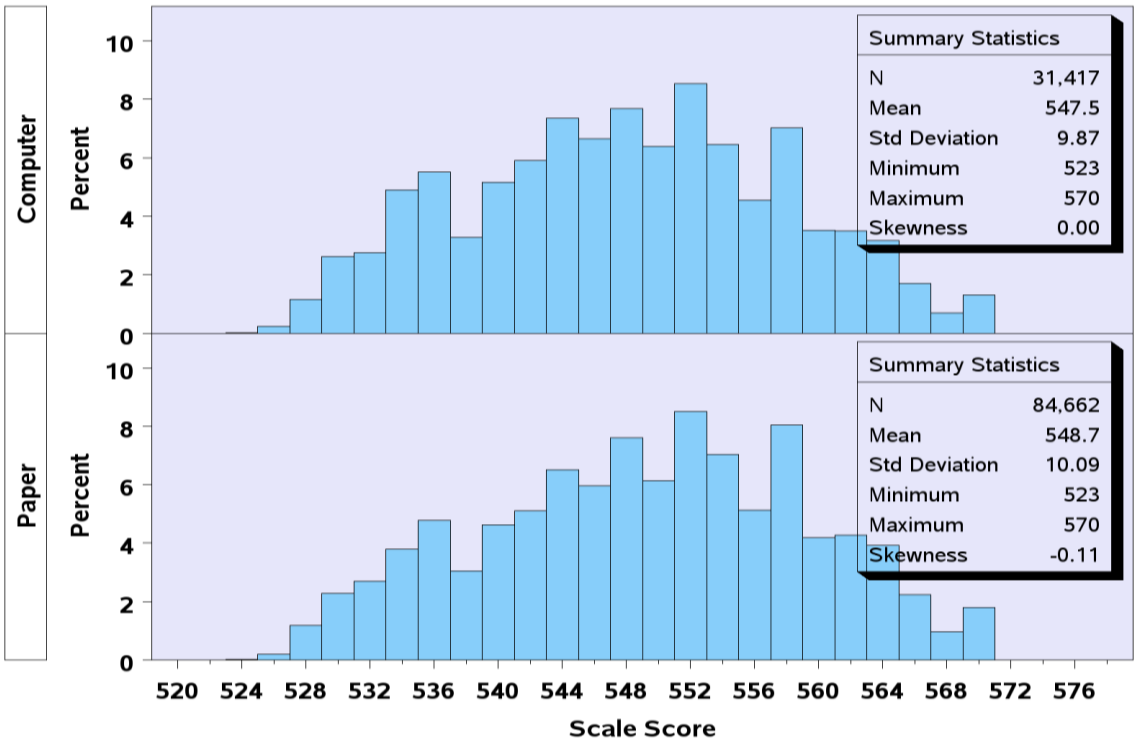



*Figure 8. 10 Grade 4 Mathematics Scale Score Distribution by Mode, Spring 2019*

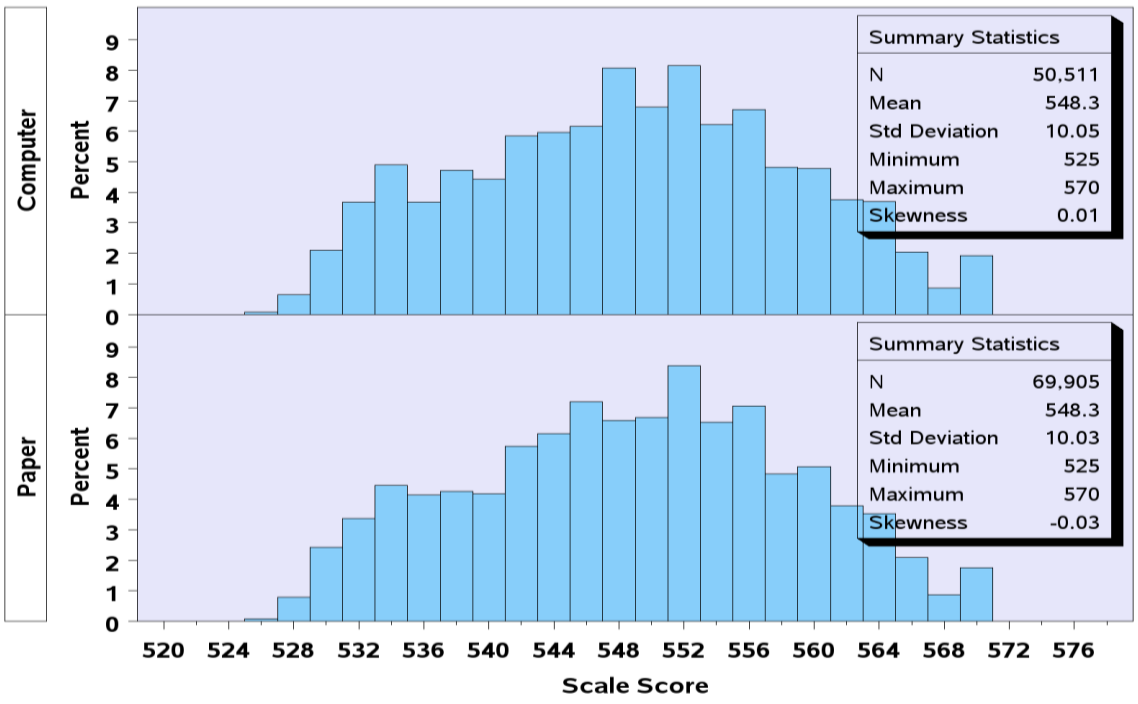


*Figure 8. 11 Grade 5 Mathematics Scale Score Distribution by Mode, Spring 2019*

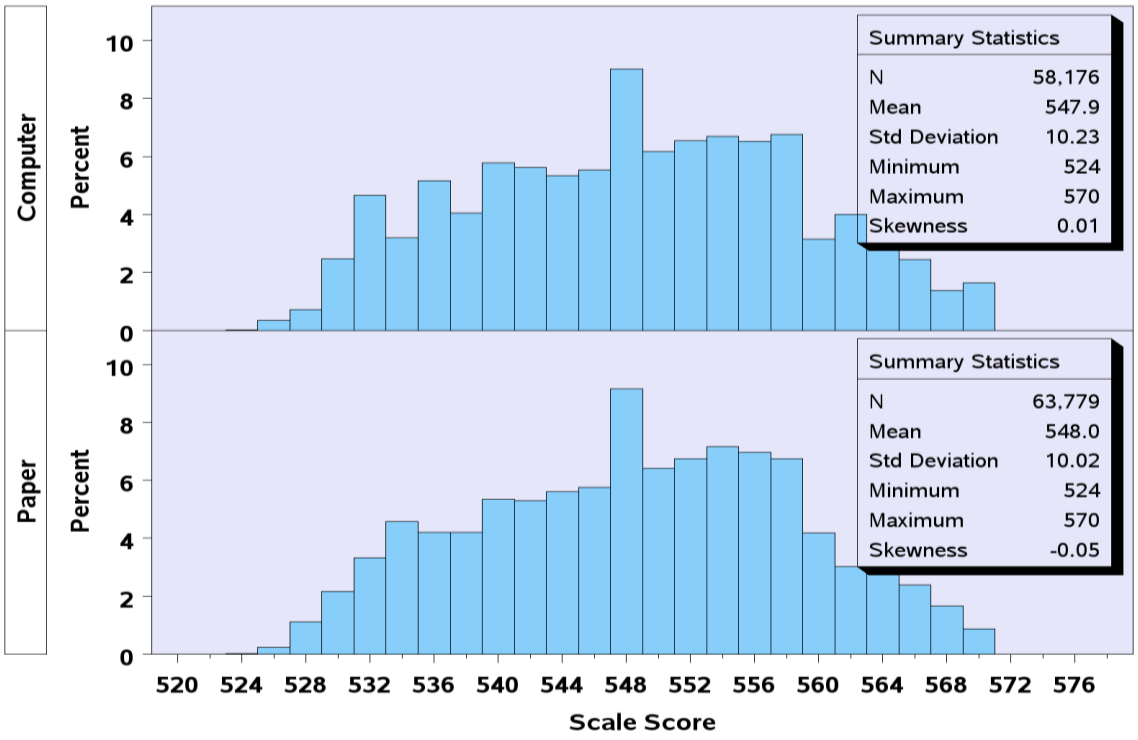



*Figure 8. 12 Grade 6 Mathematics Scale Score Distributions by Mode, Spring 2019*

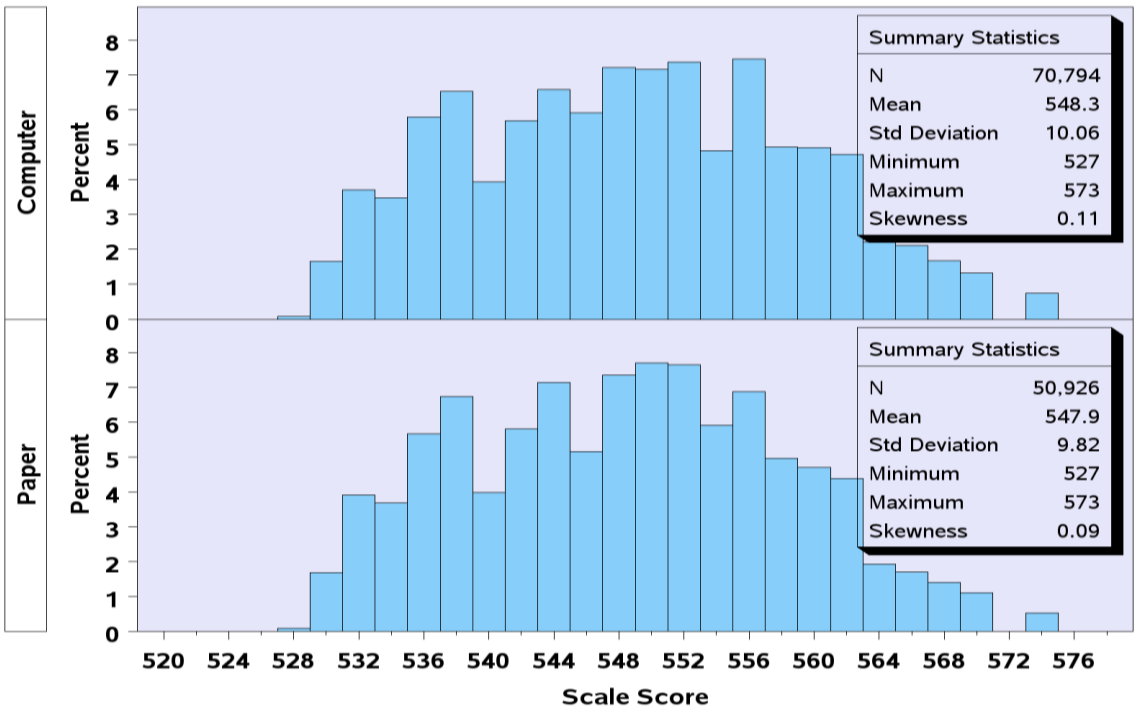


*Figure 8. 13 Grade 7 Mathematics Scale Score Distribution by Mode, Spring 2019*

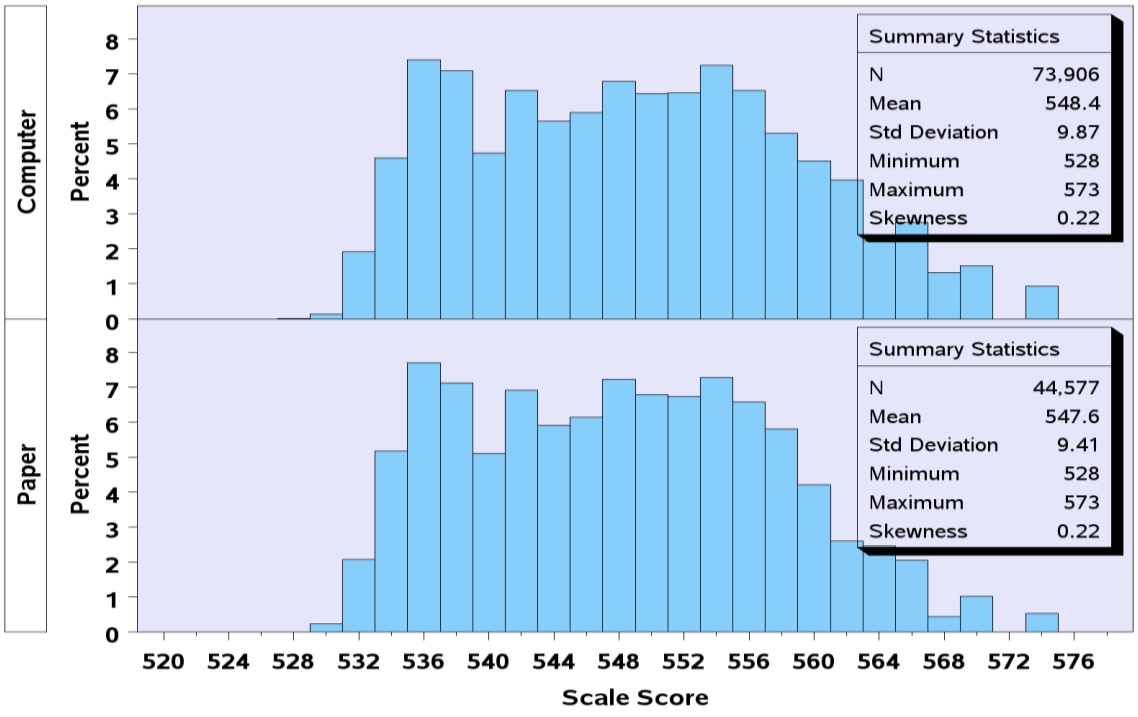



*Figure 8. 14 Grade 8 Mathematics Scale Score Distributions by Mode, Spring 2019*

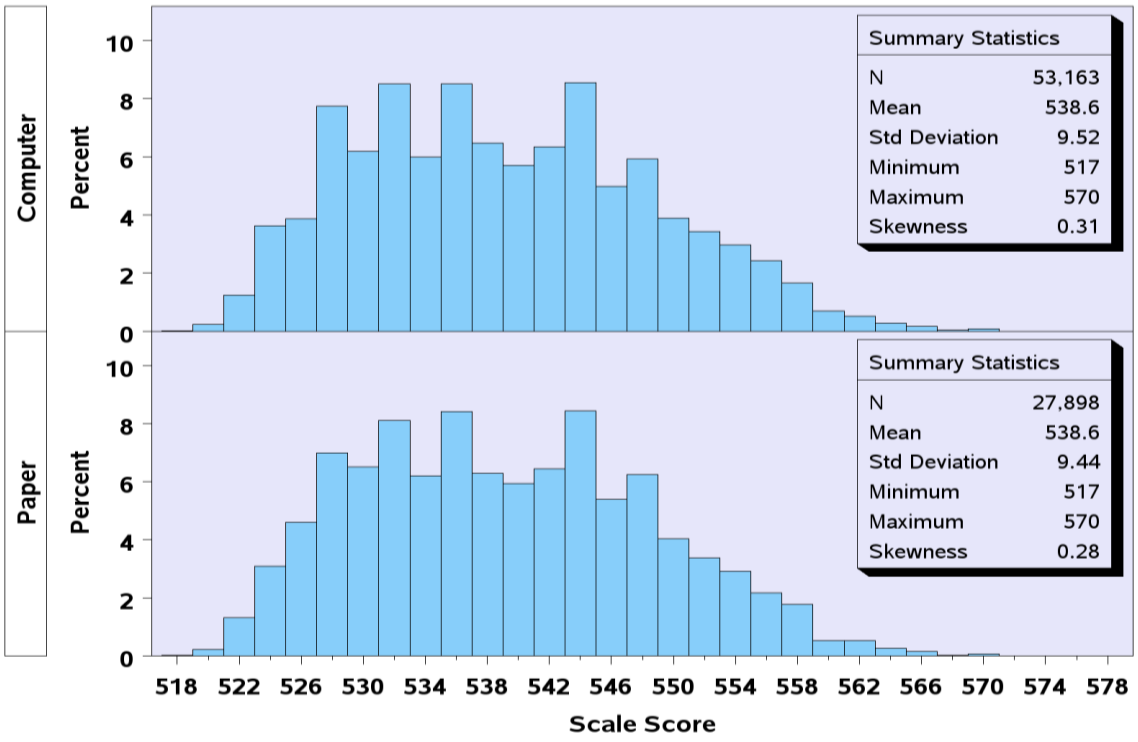


*Figure 8. 15 NC Math 1 Scale Score Distributions by Mode, 2018–19*

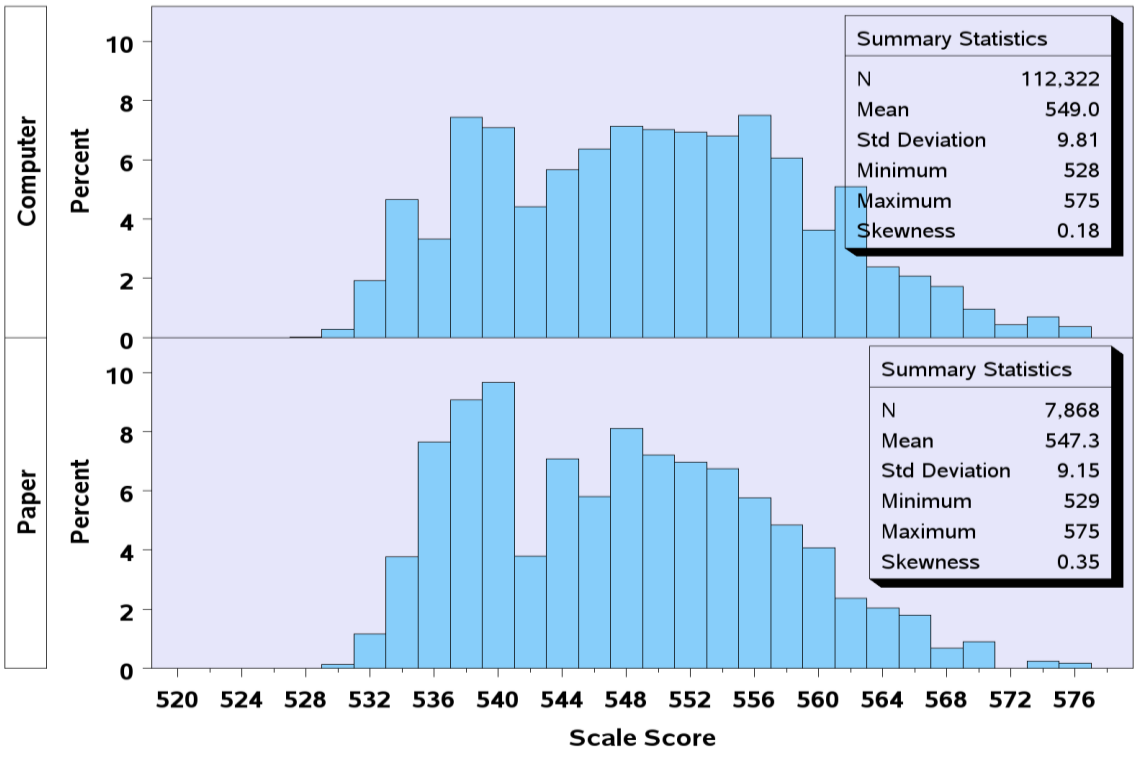
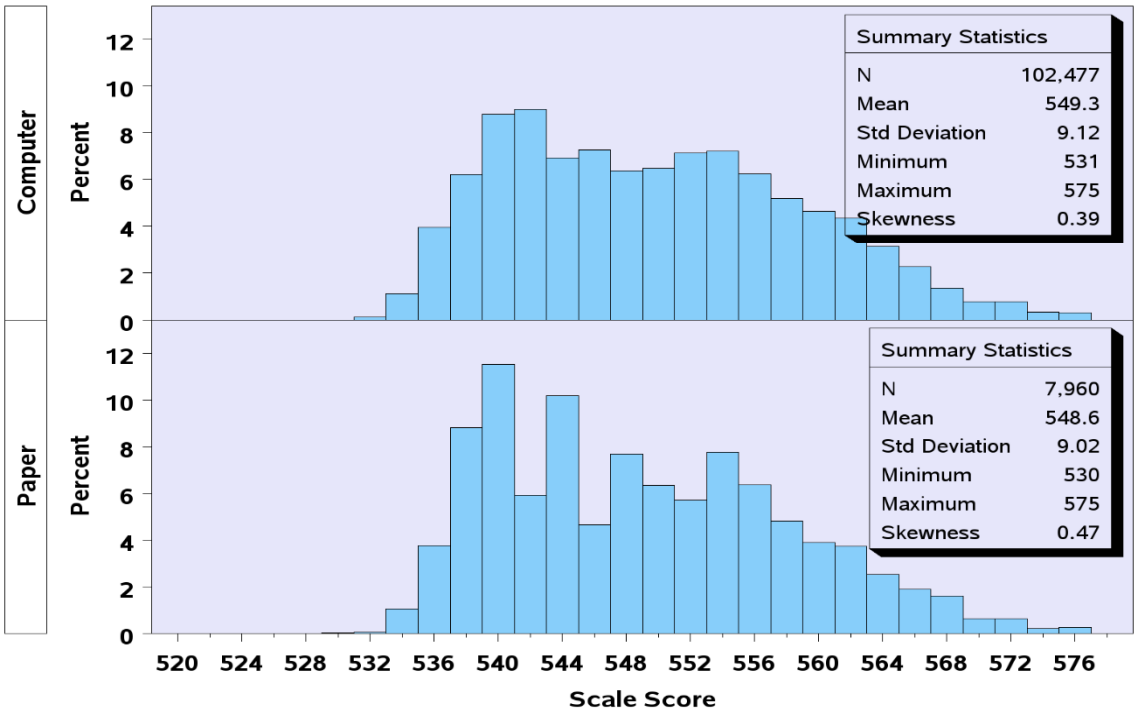



*Figure 8. 16 NC Math 3 Scale Score Distribution by Mode, 2018–19*



### 8.1.5 Achievement Levels Distributions

Beginning in 2018–19 with *Edition 5* of EOG and EOC, the NCDPI transitioned to classify and report student performance on EOG and EOC mathematics using four (4) performance or achievement levels aligned to grade level content standards and policy expectations. The four achievement levels presented in Chapter 7 are:

- **Not Proficient**: Students demonstrate inconsistent understanding of grade level content standards and will need support at the next grade/course.
- **Level 3**: Students demonstrate sufficient understanding of grade level content standards though some support may be needed to engage with content at the next grade/course.
- **Level 4**: Students demonstrate a thorough understanding of grade level content standards and are on track for career and college.
- **Level 5**: Students demonstrate comprehensive understanding of grade level content standards, are on track for career and college and are prepared for advanced content at the next grade/course.

These policy descriptors are used to summarize performance expectations for students at each level. For a detailed explanation of what students in each performance level are expected to be able to do refer to the full achievement level descriptors in *Appendix 8-B*. These achievement levels with their associated achievement level descriptors represent the principal standards-based claims that the NCDPI has sufficient validity evidence for interpreting students' EOG and EOC scores.

Based on NC state law prescribed in the state accountability model, all students with EOG or EOC performance levels of Level 3, Level 4 and Level 5 are considered and reported to have met grade level performance expectations. Students classified as Level 4 and Level 5 are further designated to be on track for CCR. This subset of Level 4 and Level 5 students is also used for federal accountability, to report the number of students proficient from state EOG assessment who are also on track for CCR.

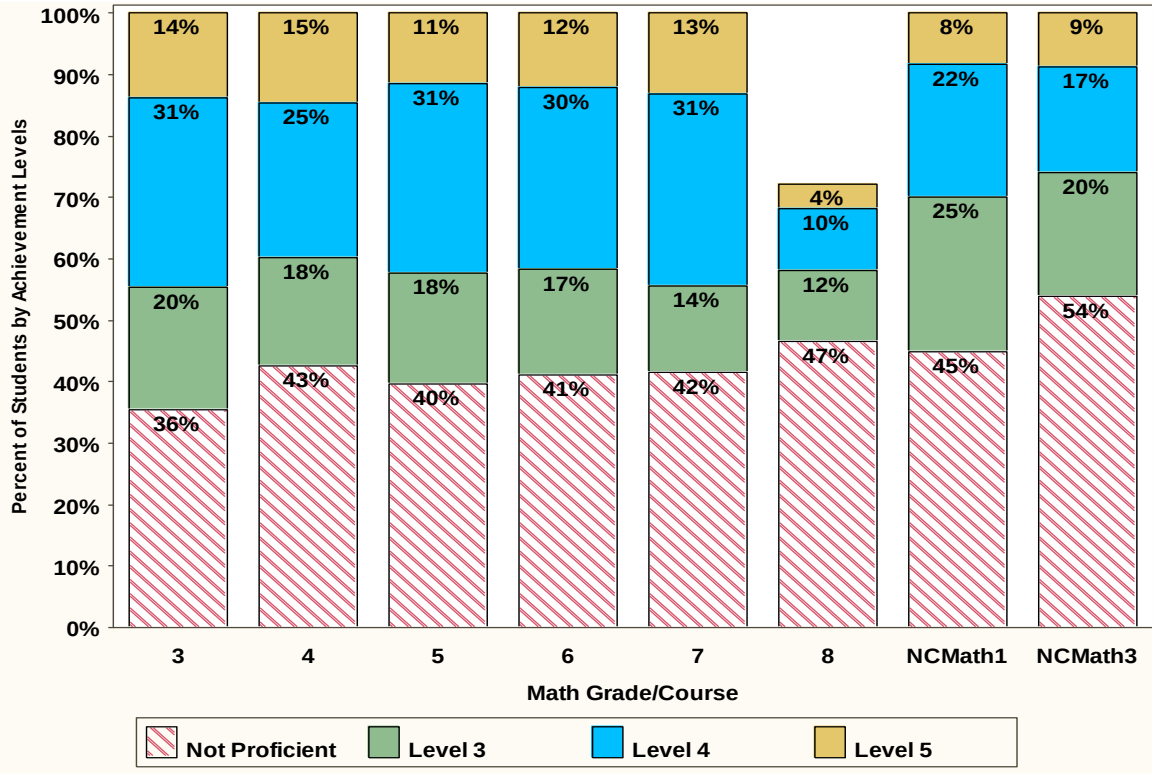
Additionally, NC state law and NC SBE policy require that all students classified as Level 5 based on previous year EOG or EOC results must be given the option in the following year to enroll in an advanced course at the next level.

*Figure 8.17* shows the summary of proportion of students by achievement level classifications from the 2018-19 North Carolina mathematics EOG and EOC assessments. The stacked bar graph shows the distribution by grade or course. For example, in EOG grade 3, 36% students are classified as Not Proficient, 20% Level 3, 31% Level 4 and 14% Level 5. Also, for state accountability reporting purposes, 65% of NC grade 3 students who took the EOG mathematics assessment are considered to have met grade level content expectations. While about 45% of

these students are considered proficient and on-track for CCR. The proficiency level classifications for other subgroups (SWD, EDS, and ELs) are shown in *Appendix 8-C*.

The stacked bar representing EOG grade 8 is an outlier. The NCDPI no longer double test grade 8 students who were also enrolled in NC Math 1 course in the same school year. Students with dual enrollment status are only required to take EOC NC Math 1 at the end of the year. These students are generally most of previous year's grade 7 students who were classified as Level 4 and Level 5 from EOG grade 7. As a result, the remaining subset of students who took EOG mathematics in grade 8 represent a truncated distribution that is skewed to the left.

*Figure 8. 17 State Level Achievement Level Classifications by Grade, 2018–19*

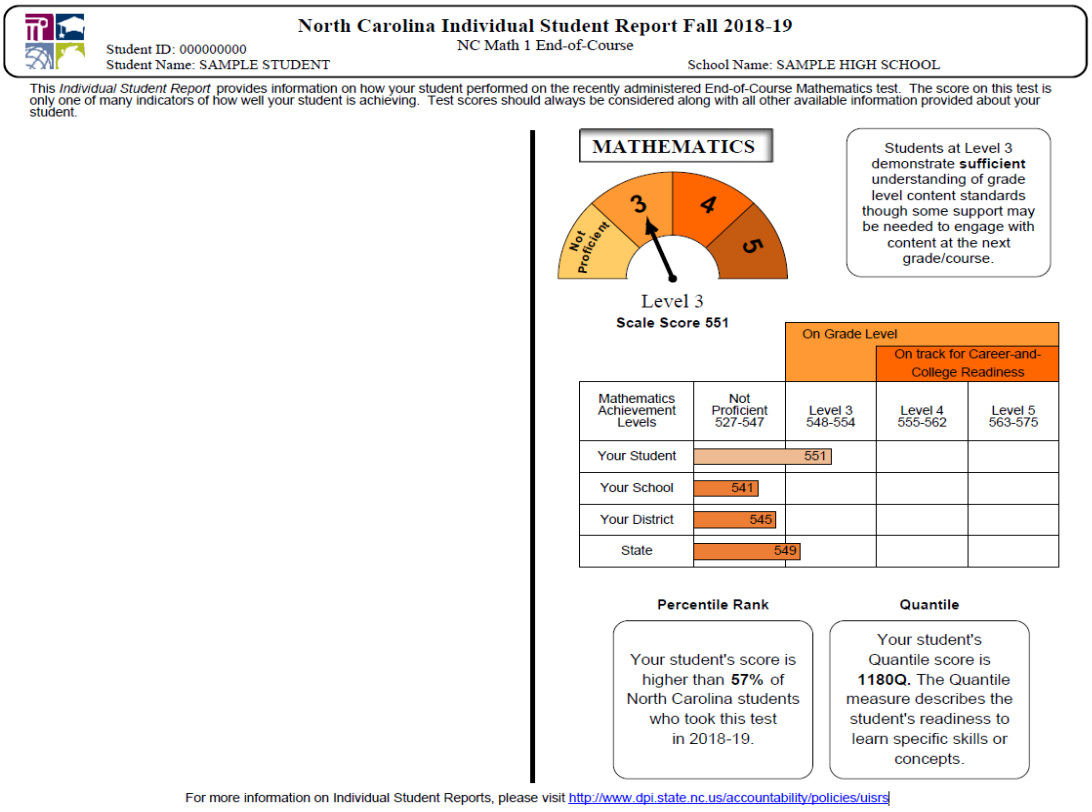



## 8.2 Score Reports

Consistent with Standard 1.1 (AERA, APA, & NCME, 2014) which states, "*Test developers should set forth clearly how test scores are intended to be interpreted and consequently used* " (p. 23), annual results from EOG and EOC assessments are compiled and reported in a variety of formats for two main audiences. The first audience reporting category is for individual students and their parents/guardians. The Individual Student Report (ISR) shown in *Figure 8.18* is designed to inform students and their parents/guardian on their overall performance based on

the EOC assessment as it relates to their standing on grade level content. The ISR highlights the achievement level and descriptor, with the associated scale score, the student is classified into based on performance. It also gives a quick comparative overview of the student's performance in relation to the school, district and all students in the state who took the EOC. More information and description of the ISR is available on the NCDPI website or through the link http://www.dpi.state.nc.us/accountability/policies/uisrs .

*Figure 8. 18 Individual Student Report (ISR)*




The second set of reports are generally generated for school and district audiences aimed to provide teachers and school administrators with in-depth and disaggregated data of their students and school performance to help inform instructional policies. In the current report format these reports are available as flat files that are pre-programmed in the reporting system and distributed to schools upon request. The goal, moving forward, is to have these reports in query database format so schools and districts, will be able to run custom reports, in real time. *Table 8.10* shows a summary list of the main pre-programmed static reports that are currently available to the different audiences for EOG and EOC mathematics assessments. The NCDPI also publishes on its website interpretive guides intended to help educators and decision makers at the classroom, school and district levels understand the content and uses of the various score reports (See

*Appendix 8-D*). These guides are also intended to help administrators and educators explain test results to parents and to the public.

*Table 8. 10 Reports by Audience*

| Report | Audience | | | | |
|---|---|---|---|---|---|
| | | Administrative | | | |
| | Parent/ Student | Teacher | School | District | State |
| Individual Student Report (ISRs) | ✓ | ✓ | ✓ | | |
| Class Roster Reports | | ✓ | ✓ | | |
| Score and Achievement Level Frequency | | ✓ | ✓ | ✓ | ✓ |
| Goal Summary Reports | | ✓ | ✓ | ✓ | ✓ |

## 8.3 Confidentiality of Student Information

State Board of Education policy GCS-A-010 (j)(1) states "*Educators shall maintain the confidentiality of individual students. Publicizing test scores or any written material containing personally identifiable information from the student's educational records shall not be disseminated or otherwise made available to the public by a member of the State Board of Education, any employee of the State Board of Education, the State Superintendent of Public Instruction, any employee of the North Carolina Department of Public Instruction, any member of a local board of education, any employee of a local board of education, or any other person, except as permitted under the provisions of the Family Educational Rights and Privacy Act of 1974, 20 U.S.C.§1232g*."

### 8.3.1 Confidentiality of Personal Information

The North Carolina Test Coordinators' Policies and Procedures Handbook instructs that while handling and transmitting personally identifiable information employees of Public School Units the NCDPI or other education institutions are legally and ethically obliged to safeguard the confidentiality of any private information they access while performing official duties. To protect the confidentiality of individuals from those who are not authorized to access individual-level data, Personally Identifiable Information (PII) is encrypted during transmission using one of the following methods, in order of preference:

- Secure FTP Server based on SFTP or FTPS protocols - Preferred method and most widely acceptable standard for transmitting encrypted data.
- Encrypted E-mail – If secure FTP capabilities do not exist, encrypted e-mail can be used.
- Password Protected E-mail – If compatible encryption is not available to both parties, data should be password protected. The password should be given to the recipient through a different medium, such as a phone call, never in notes or documents accompanying the actual data file, or another e-mail. In addition, the password should not be transferred via voicemail.

When sending e-mail, either encrypted or password protected, it is advised to ensure that it contains the least amount of Family Educational Rights and Privacy Act (FERPA) –protected information as possible. The subject line of an e-mail should not include FERPA–protected information; the body of an e-mail should not contain highly sensitive FERPA–protected information, such as a student's Social Security Number or full name. FERPA– protected data should always be in an attached encrypted/password protected file, never in the body of an email. Secure test questions, answer choices or portions of secure test questions or answer choices must not be sent via e-mail (use e-mail only if encrypted and/or password protected).

Fax machines and printers used to send and receive secure data must be located in areas that are secure. Public School Units should not use private or personal accounts to store students' PII. Public School Units wish to use the G suite for Education (previously called Google Apps for Education) should consult with their legal team to ensure compliance with FERPA and state security guidelines. Furthermore, it is recommended that the Data Leak Protection (DLP) feature of G Suite be used to protect data, even though FERPA compliance does not require DLP.

### 8.3.2 Confidentiality of Test Data

Confidential data must be transferred using secure methods (e.g., Secure File Transfer Protocol or receipted parcel delivery services, such as the U.S. Postal Service, UPS, or Federal Express). When placing confidential data on portable devices (e.g., laptops, thumb drives), the portable device must be protected by encryption or password protection. Some specific examples of confidential data that must not be released to anyone include the following:

- WinScan files contain data that are for test development and accountability purposes only, and their release would violate test security.
- The EDS data are property of the NCDPI and School Nutrition Services. Accountability Services has access to the data through a Memorandum of Understanding (MOU). Test coordinators are bound by the requirements of the MOU and FERPA to preserve the confidentiality of this data. Releasing this data to anyone in any manner that would allow the identification of the EDS status of an individual student would be a violation of federal law.

# CHAPTER 9 VALIDITY EVIDENCES

This chapter presents additional validity evidences collected in support of the interpretation of *Edition 5* mathematics EOG and EOC test scores. The first two sections present validity evidence in support of the internal structure of EOG and EOC assessments. Evidence presented in these sections include reliability, standard error estimates and classification consistency summary of reported achievement levels and an exploratory principal component analysis (PCA) to support the unidimensional interpretation of EOG and EOC mathematics scores. The penultimate sections of the chapter document content validity evidence summarized from the alignment study and evidence based on relation to other variables summarized from the EOG/EOC Quantile Framework linking study, while the last part presents summary of procedures used to ensure EOG and EOC assessments are accessible and fair for all students.

## 9.1 Reliability of Mathematics EOG and EOC Assessments

Internal consistency, as a reliability estimate, provides a sample base summary statistic that describes the proportion of the reported score variability that is attributed to true score variance. To justify valid use of test results in large-scale standardized assessments, evidence must be documented that shows test results are stable, consistent and dependable across all subgroups of the intended population. A reliable assessment produces scores that are expected to be relatively stable if the test is administered repeatedly under similar conditions to the same students. Scores from a reliable test reflect examinees' estimated expected ability in the construct being measured with very little error variance. Internal consistency reliability coefficients, measured by Cronbach alpha, range from 0.0 to 1.0, where a coefficient of 1.0 refers to a perfectly reliable measure with no measurement error. For high-stakes assessments, alpha estimates of 0.85 or higher are generally desirable. Cronbach's alpha (Cronbach, 1951) is calculated as:

$$\hat{\alpha} = \frac{\kappa}{\kappa-1}\left(1 - \frac{\Sigma\hat{\sigma}_i^2}{\hat{\sigma}_X^2}\right) \qquad (9\text{-}1)$$

Where *k* is the number of items on the test form, $\hat{\sigma}_i^2$ is the variance of item i and $\hat{\sigma}_X^2$ is the total test variance. It is worth noting that reliability estimates are less informative in describing the accuracy of individual students' scores, since they are sample based. *Table 9.1* shows reliability estimates (Cronbach alpha) for all EOG mathematics forms by grade and major demographic variables for 2018–19 administration. Similarly, *Table 9.2* shows alphas for EOC mathematics tests. Across all forms, overall reliability estimates based on the 2018–19 population ranged from 0.90 to 0.94. Subgroup reliabilities are also consistent across forms for the most part, they are consistently higher than the 0.85 threshold. Note that NC Math 1 and NC Math 3 forms are designed for computer-based administration with paper forms as accommodations. The sample size for the accommodation subgroups was small and variation was low.

*Table 9. 1 EOG Mathematics Reliabilities (Alpha) by Form and Subgroup*

| Grade | Mode | Form | Gender | | Ethnicity1 | | | Accommodations1 | | | All |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Female | Male | Black | Hispanic | White | EDS | SWD | ELs | |
| 3 | Both | A | 0.93 | 0.93 | 0.92 | 0.92 | 0.92 | 0.92 | 0.94 | 0.91 | 0.93 |
| | Both | B | 0.93 | 0.93 | 0.92 | 0.92 | 0.92 | 0.92 | 0.93 | 0.92 | 0.93 |
| | Both | C | 0.93 | 0.93 | 0.92 | 0.92 | 0.92 | 0.92 | 0.93 | 0.91 | 0.93 |
| 4 | PBT | A | 0.92 | 0.93 | 0.91 | 0.91 | 0.92 | 0.92 | 0.93 | 0.9 | 0.93 |
| | CBT | A | 0.92 | 0.93 | 0.91 | 0.9 | 0.92 | 0.91 | 0.92 | 0.9 | 0.92 |
| | PBT | B | 0.93 | 0.93 | 0.91 | 0.92 | 0.92 | 0.92 | 0.93 | 0.91 | 0.93 |
| | CBT | B | 0.92 | 0.92 | 0.91 | 0.91 | 0.91 | 0.91 | 0.93 | 0.9 | 0.92 |
| | PBT | C | 0.93 | 0.93 | 0.91 | 0.92 | 0.92 | 0.92 | 0.93 | 0.91 | 0.93 |
| | CBT | C | 0.93 | 0.93 | 0.91 | 0.92 | 0.93 | 0.92 | 0.94 | 0.91 | 0.93 |
| 5 | PBT | A | 0.91 | 0.92 | 0.9 | 0.9 | 0.9 | 0.9 | 0.92 | 0.87 | 0.91 |
| | CBT | A | 0.91 | 0.92 | 0.9 | 0.9 | 0.91 | 0.9 | 0.92 | 0.87 | 0.92 |
| | PBT | B | 0.92 | 0.92 | 0.91 | 0.91 | 0.91 | 0.91 | 0.92 | 0.89 | 0.92 |
| | CBT | B | 0.92 | 0.92 | 0.9 | 0.9 | 0.91 | 0.9 | 0.94 | 0.89 | 0.92 |
| | PBT | C | 0.91 | 0.92 | 0.9 | 0.9 | 0.91 | 0.91 | 0.92 | 0.89 | 0.92 |
| | CBT | C | 0.92 | 0.92 | 0.9 | 0.9 | 0.91 | 0.91 | 0.92 | 0.88 | 0.92 |
| 6 | PBT | A | 0.93 | 0.93 | 0.91 | 0.92 | 0.93 | 0.92 | 0.92 | 0.89 | 0.93 |
| | CBT | A | 0.93 | 0.94 | 0.91 | 0.92 | 0.93 | 0.91 | 0.91 | 0.89 | 0.93 |
| | PBT | B | 0.93 | 0.94 | 0.91 | 0.92 | 0.93 | 0.92 | 0.91 | 0.87 | 0.94 |
| | CBT | B | 0.93 | 0.94 | 0.91 | 0.92 | 0.93 | 0.92 | 0.93 | 0.91 | 0.94 |
| | PBT | C | 0.93 | 0.93 | 0.91 | 0.92 | 0.93 | 0.92 | 0.92 | 0.86 | 0.93 |
| | CBT | C | 0.93 | 0.94 | 0.91 | 0.92 | 0.93 | 0.92 | 0.93 | 0.91 | 0.93 |
| 7 | PBT | A | 0.93 | 0.94 | 0.91 | 0.93 | 0.93 | 0.92 | 0.9 | 0.9 | 0.94 |
| | CBT | A | 0.94 | 0.94 | 0.92 | 0.93 | 0.94 | 0.92 | 0.92 | 0.92 | 0.94 |
| | PBT | B | 0.94 | 0.94 | 0.91 | 0.92 | 0.93 | 0.92 | 0.9 | 0.87 | 0.94 |
| | CBT | B | 0.94 | 0.94 | 0.91 | 0.92 | 0.94 | 0.92 | 0.93 | 0.92 | 0.94 |
| | PBT | C | 0.94 | 0.94 | 0.91 | 0.92 | 0.93 | 0.92 | 0.9 | 0.87 | 0.94 |
| | CBT | C | 0.94 | 0.94 | 0.91 | 0.92 | 0.93 | 0.92 | 0.9 | 0.9 | 0.94 |
| 8 | CBT | A | 0.91 | 0.91 | 0.88 | 0.90 | 0.91 | 0.89 | 0.88 | 0.88 | 0.91 |
| | CBT | B | 0.90 | 0.90 | 0.87 | 0.89 | 0.90 | 0.88 | 0.87 | 0.88 | 0.90 |
| | CBT | C | 0.89 | 0.89 | 0.86 | 0.88 | 0.90 | 0.88 | 0.85 | 0.88 | 0.89 |

[1]Reliabilities estimates are displayed only for major ethnic groups and accommodations investigated in DIF analysis with acceptable sample size.

*Table 9. 2 EOC Mathematics Reliabilities (Alpha) by Form and Subgroup*

| Course | Mode | Form | Gender | | Ethnicity1 | | | Accommodations1 | | | All |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Female | Male | Black | Hispanic | White | EDS | SWD | ELs | |
| NC Math 1 | PBT | A | 0.92 | 0.92 | 0.89 | 0.91 | 0.92 | 0.9 | 0.86 | 0.87 | 0.92 |
| | CBT | M | 0.94 | 0.94 | 0.92 | 0.93 | 0.94 | 0.92 | 0.89 | 0.89 | 0.94 |
| | CBT | N | 0.93 | 0.94 | 0.91 | 0.92 | 0.93 | 0.91 | 0.9 | 0.89 | 0.94 |
| NC Math 3 | PBT | A | 0.91 | 0.92 | 0.84 | 0.89 | 0.92 | 0.86 | 0.78 | 0.66 | 0.92 |
| | CBT | M | 0.92 | 0.92 | 0.87 | 0.89 | 0.92 | 0.88 | 0.81 | 0.78 | 0.92 |
| | PBT | B | 0.91 | 0.92 | 0.84 | 0.88 | 0.92 | 0.86 | 0.77 | 0.65 | 0.92 |
| | CBT | N | 0.91 | 0.92 | 0.85 | 0.89 | 0.92 | 0.87 | 0.78 | 0.72 | 0.92 |
| | CBT | O | 0.92 | 0.93 | 0.86 | 0.90 | 0.93 | 0.88 | 0.82 | 0.79 | 0.92 |

[1]Reliabilities estimates are displayed only for major ethnic groups and accommodations investigated in DIF analysis with acceptable sample size.

## 9.2 Conditional Standard Errors at Scale Score Cuts

The information provided by the standard error (SE) for a given cut score is important because it helps in determining the accuracy of examinees' classifications. It allows a probabilistic statement to be made about an individual's test score. The conditional SEs at the lowest obtainable scale score (LOSS), highest obtainable scale score (HOSS) and scale score cuts at the achievement levels for the North Carolina EOG mathematics forms are shown in *Table 9.3* and EOC forms are shown in *Table 9.4*.

The conditional SE can be used to estimate a confidence band around any scale score or cut score where a decision must be precise. For example, the on-grade proficiency (Level 3) cut score for grade 3 mathematics is 545 (see *Table 9.3*). A student who took Form A and scored 545 with a SE of 2 has a 68% probability that his or her true score or ability ranges from 543 to 547 (545±1x2) when reported with a 1 standard error level of precision. Similarly, if an educator wants to estimate the students' true score with less precision say 2 standard error then the 95% confidence interval of the student predicted ability will be from 539 to 551 (545±2x3). For most of the EOG and EOC mathematics scale score cuts in the middle range, particularly at the Level 3 and Level 4, the conditional standard errors are between 2 and 3. Cuts at the LOSS have the conditional SEs between 5 and 6 and at the HOSS between 4 and 5. The higher SEs at the LOSS and HOSS are typical for extreme scores which allow less measurement precision because of a lack of informative items at those ability ranges.

*Table 9. 3 Conditional Standard Errors (SE) at Achievement Level Cuts for Grades 3–8 by Form*

| Grade | Mode | Form | Min | | Level 3 | | Level 4 | | Level 5 | | Max | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | LOSS | SE | Cut | SE | Cut | SE | Cut | SE | HOSS | SE |
| 3 | Both | A | 523 | 5 | 545 | 2 | 551 | 2 | 560 | 3 | 570 | 5 |
| | Both | B | 524 | 5 | 545 | 2 | 551 | 2 | 560 | 3 | 570 | 5 |
| | Both | C | 523 | 5 | 545 | 2 | 551 | 2 | 560 | 3 | 570 | 5 |
| 4 | Both | A | 525 | 5 | 547 | 2 | 552 | 2 | 560 | 3 | 570 | 5 |
| | Both | B | 525 | 5 | 547 | 2 | 552 | 2 | 560 | 3 | 570 | 5 |
| | Both | C | 525 | 5 | 547 | 2 | 552 | 2 | 560 | 3 | 570 | 5 |
| 5 | Both | A | 524 | 5 | 546 | 3 | 551 | 3 | 561 | 3 | 570 | 5 |
| | PBT | B | 524 | 5 | 546 | 2 | 551 | 2 | 561 | 3 | 570 | 5 |
| | CBT | B | 524 | 5 | 546 | 3 | 551 | 2 | 561 | 3 | 570 | 5 |
| | Both | C | 524 | 5 | 546 | 2 | 551 | 2 | 561 | 3 | 570 | 5 |
| 6 | Both | A | 527 | 5 | 546 | 2 | 551 | 2 | 561 | 3 | 573 | 5 |
| | Both | B | 527 | 5 | 546 | 2 | 551 | 2 | 561 | 3 | 573 | 5 |
| | Both | C | 527 | 5 | 546 | 2 | 551 | 2 | 561 | 2 | 573 | 5 |
| 7 | Both | A | 529 | 5 | 546 | 2 | 550 | 2 | 560 | 2 | 573 | 5 |
| | Both | B | 529 | 5 | 546 | 3 | 550 | 2 | 560 | 2 | 573 | 5 |
| | Both | C | 529 | 5 | 546 | 2 | 550 | 2 | 560 | 2 | 573 | 5 |
| 8 | CBT | A | 519 | 5 | 543 | 3 | 548 | 2 | 555 | 3 | 570 | 5 |
| | CBT | B | 517 | 6 | 543 | 3 | 548 | 2 | 555 | 3 | 570 | 5 |
| | CBT | C | 518 | 5 | 543 | 3 | 548 | 3 | 555 | 3 | 570 | 4 |

*Table 9. 4 Conditional Standard Errors (SE) at Achievement Level Cuts for NC Math 1 and NC Math 3 by Form*

| Course | Mode | Form | Min | | Level 3 | | Level 4 | | Level 5 | | Max | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | LOSS | SE | Cut | SE | Cut | SE | Cut | SE | HOSS | SE |
| NC Math 1 | PBT | A | 529 | 5 | 548 | 2 | 555 | 2 | 563 | 2 | 575 | 5 |
| | CBT | M | 528 | 5 | 548 | 2 | 555 | 2 | 563 | 2 | 575 | 5 |
| | CBT | N | 528 | 5 | 548 | 2 | 555 | 2 | 563 | 3 | 575 | 5 |
| NC Math 3 | PBT | A | 530 | 5 | 550 | 3 | 556 | 2 | 563 | 2 | 575 | 4 |
| | CBT | M | 531 | 6 | 550 | 3 | 556 | 2 | 563 | 2 | 575 | 4 |
| | Both | B/N | 531 | 5 | 550 | 3 | 556 | 2 | 563 | 2 | 575 | 5 |
| | CBT | O | 532 | 6 | 550 | 3 | 556 | 2 | 563 | 2 | 575 | 5 |

## 9.3 Classification Consistency

The No Child Left Behind Act of 2001 (USED, 2002) and subsequent Race to the Top Act of 2009 (2009) emphasized the measurement of adequate yearly progress (AYP) with respect to the percentage of students at or above performance standards set by states. With this emphasis on the achievement level classification, it is very important to provide evidence that shows all students are consistently and accurately classified into one of the four achievement levels. The importance of classification consistency as a measure of the categorical decisions when the test is used repeatedly has been recognized in Standard 2.16 (AERA, APA, & NCME, 2014), which states, "*When a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure*" (p. 46).

The methodology used for estimating the reliability of achievement level classification decisions as described in Hanson and Brennan (1990) and Livingston and Lewis (1995) provides estimates of decision accuracy and classification consistency. The classification consistency refers to "the agreement between classifications based on two non-overlapping, equally difficult forms of the test," and decision accuracy refers to "the extent to which the actual classifications of test takers (on the basis of their single-form scores) agree with those that would be made on the basis of their true scores, if their true scores could somehow be known" (Livingston & Lewis, 1995, p. 178). That is, classification consistency refers to the agreement between two observed scores, while classification accuracy refers to the agreement between observed and true scores.

The classification consistency analysis was conducted using the computer program BB-Class[2]. The program provides results for both the Hanson and Brennan, or HB, (1990) and Livingston and Lewis, or LL, (1995) procedures. Since the Hanson and Brennan (1990) procedures assume "test consists of n equally weighted, dichotomously-scored items," while the Livingston and Lewis (1995) procedures intends to handle situations when "a) items are not equally weighted and/or b) some or all of the items are polytomous scored" (Brennan, 2004, pp. 2–3), therefore the classification consistency analyses for the North Carolina mathematics EOG and EOC followed the HB procedures.

*Table 9.5* shows the decision accuracy and consistency indexes for achievement levels at each grade. Note that there is a range of classification accuracy and consistency in the table for some grades because cut scores corresponding to different raw scores in some forms of the grade are different. Overall, the values indicate good classification accuracy (ranging from 0.91 to 0.98) and consistency (from 0.87 to 0.97). For example, EOG grade 3 mathematics has an accuracy rate of 0.93 at Level 3 cut which means if a student who is classified as Level 3 were to take a

[2] BB-Class is an ANSI C computer program that uses the beta-binomial model (and its extensions) for estimating classification consistency and accuracy. It can be downloaded from https://www.education.uiowa.edu/centers/casma/computer-programs#de748e48-f88c-6551-b2b8-ff00000648cd.

non-overlapping, equally difficult form a second time, there is a 93% (**bolded**) probability that the student would still be classified as Level 3. The higher classification consistency also entails smaller standard error and higher reliability.

*Table 9. 5 Classification Accuracy and Consistency Results, EOG and EOC Mathematics*

| Grade | Level 3 | | Level 4 | | Level 5 | |
|---|---|---|---|---|---|---|
| | Acc. | Con. | Acc. | Con. | Acc. | Con. |
| 3 | **0.93** | 0.91 | 0.93 | 0.9 | 0.93 | 0.90–0.91 |
| 4 | 0.93 | 0.9 | 0.93 | 0.9 | 0.94 | 0.91–0.92 |
| 5 | 0.92 | 0.89–0.90 | 0.92 | 0.89 | 0.94 | 0.92 |
| 6 | 0.93 | 0.9 | 0.93 | 0.91 | 0.95 | 0.94 |
| 7 | 0.93 | 0.9 | 0.94 | 0.91–0.92 | 0.96 | 0.94 |
| 8 | 0.91 | 0.88–0.89 | 0.94 | 0.91–0.92 | 0.97–0.98 | 0.95–0.97 |
| NC Math 1 | 0.93 | 0.9 | 0.94 | 0.92 | 0.96–0.97 | 0.95 |
| NC Math 3 | 0.91 | 0.87–0.88 | 0.94 | 0.92 | 0.97 | 0.96 |

Note: Acc. = Accuracy; Con. = Consistency

## 9.4 Unidimensionality of EOG and EOC Assessments

North Carolina EOG and EOC mathematics assessments are designed base on a unidimensional assumption that total score represents an estimate of students' performance based on grade level content standards. It is therefore important that the NCDPI test design show relevant validity evidence to support the unidimensional use and interpretation of EOG test scores.

Empirical evidence of overall dimensionality for EOG and EOC mathematics assessments was explored using principal component analysis (PCA). PCA is an exploratory technique that seeks to summarize observed variables using fewer linear dimensions referred to as components. The primary hypothesis in a PCA is to determine the fewest reasonable dimensions or components that can explain most of the observed variance in the data. Two commonly used criteria to decide the number of meaningful dimensions for a set of observed variables are:

- retain components whose eigenvalues are greater than the average of all the eigenvalues, which is usually 1 and
- plot eigenvalues (scree plot) against components (factors) and count the number of components above the natural linear break.

It is very common to rely on both criteria when evaluating the number of possible dimensions for a given variable. PCA were extracted from the tetrachoric correlation matrix for dichotomized response data, or from the polychoric correlation matrix for categorical scored responses, to determine the number of meaningful components.

### 9.4.1 Eigenvalues and Variance

The eigenvalue for each component describes the amount of total variance accounted for by that component. A scree plot is used to show the graphical result from PCA showing the relations between main components and cumulative variance explained. *Figure 9.1* through *Figure 9.8* show the PCA results for all mathematics forms. The left vertical axis shows the actual eigenvalues of parallel forms and the right vertical axis displays the cumulative variance. The same information for the first three components with Eigenvalues greater than 1 are summarized in *Table 9.6* through *Table 9.8*. Based on the PCA results, the average ratio of the first to the second eigenvalue across grades is about 9. Also, on average the first principal component accounts for about 40% of the total variance with the exception of EOG grade 8 (32%).

Evaluation of the scree plots with the distinct break of the linear trend after the first dominant component present enough exploratory evidence in support of the assumption of unidimensionality with a single dominant component to explain a significant amount of the total variance of the North Carolina mathematics EOG and EOC assessments. The eigenvalues and proportion of variance explained by the first component are reasonably large supporting the assumption that each test form measures a single construct. The second main component accounts for less than 5% of total variance across all mathematics forms.

The two-factor exploratory factor analysis with simple structure showed that most items loaded positively to the first factor (see *Appendix 9-A*). These results further suggest that the North Carolina EOG and EOC mathematics items at each test measured an overall mathematics construct.

Based on the two evaluation criteria described above, scree plots and variance explained by the first component, a strong case can be made for one dominant component to explain a significant amount of the total variance in the observed correlation matrices for EOG and EOC forms. Evaluation of the scree graph with the distinct break of the linear trend after the first dominant component present sufficient exploratory evidence in support of the assumption of unidimensionality of the North Carolina EOG and EOC assessments. Thus, PCA results with one dominant component support interpreting EOG and EOC mathematics score using a unidimensional scale.

*Figure 9. 1 Grade 3 PCA Scree Plot and Cumulative Variance by Form*

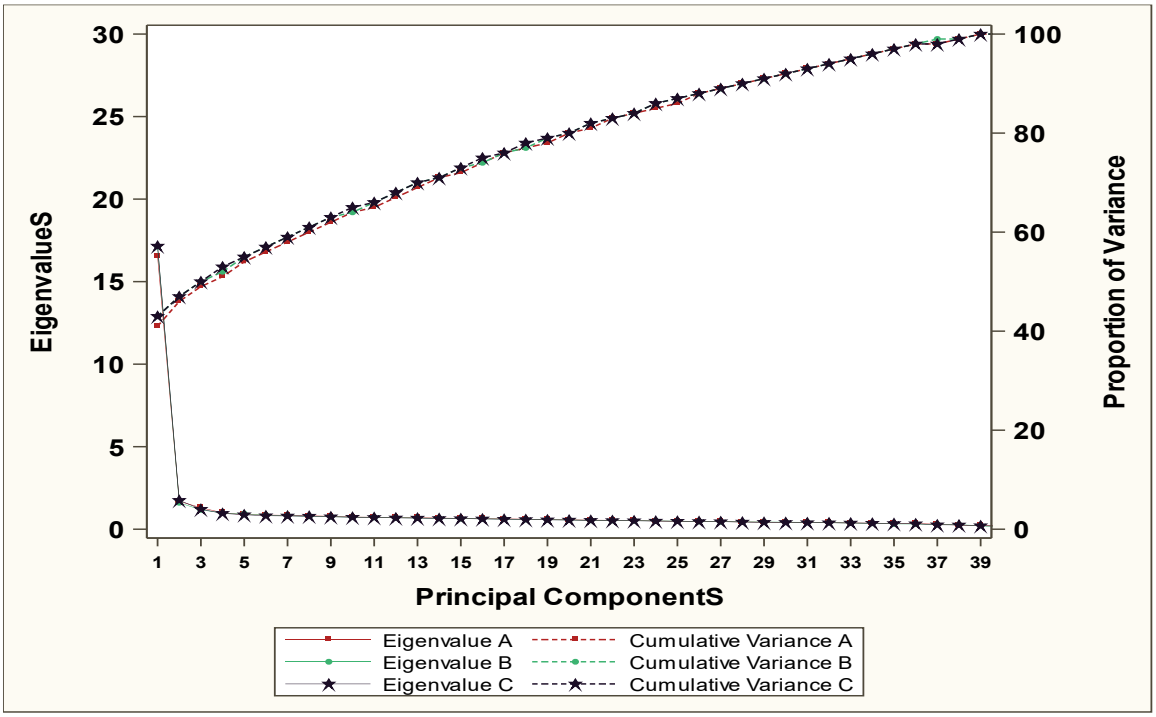



*Figure 9. 2 Grade 4 PCA Scree Plot and Cumulative Variance by Form*

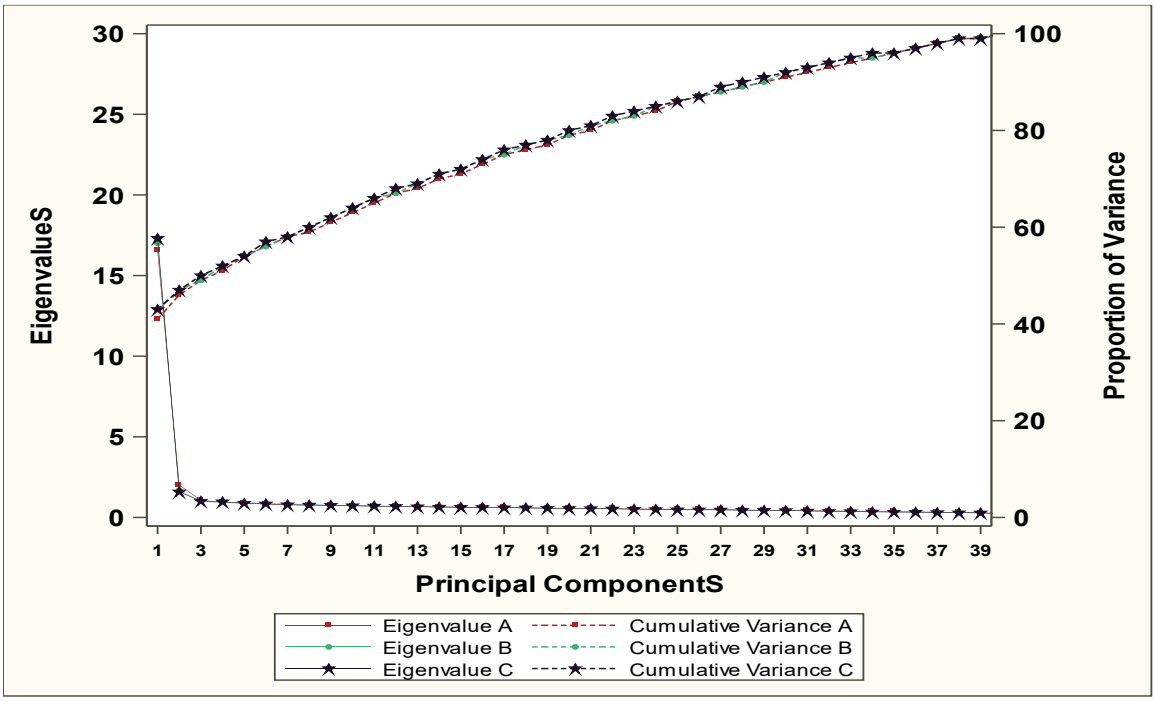


*Figure 9. 3 Grade 5 PCA Scree Plot and Cumulative Variance by Form*

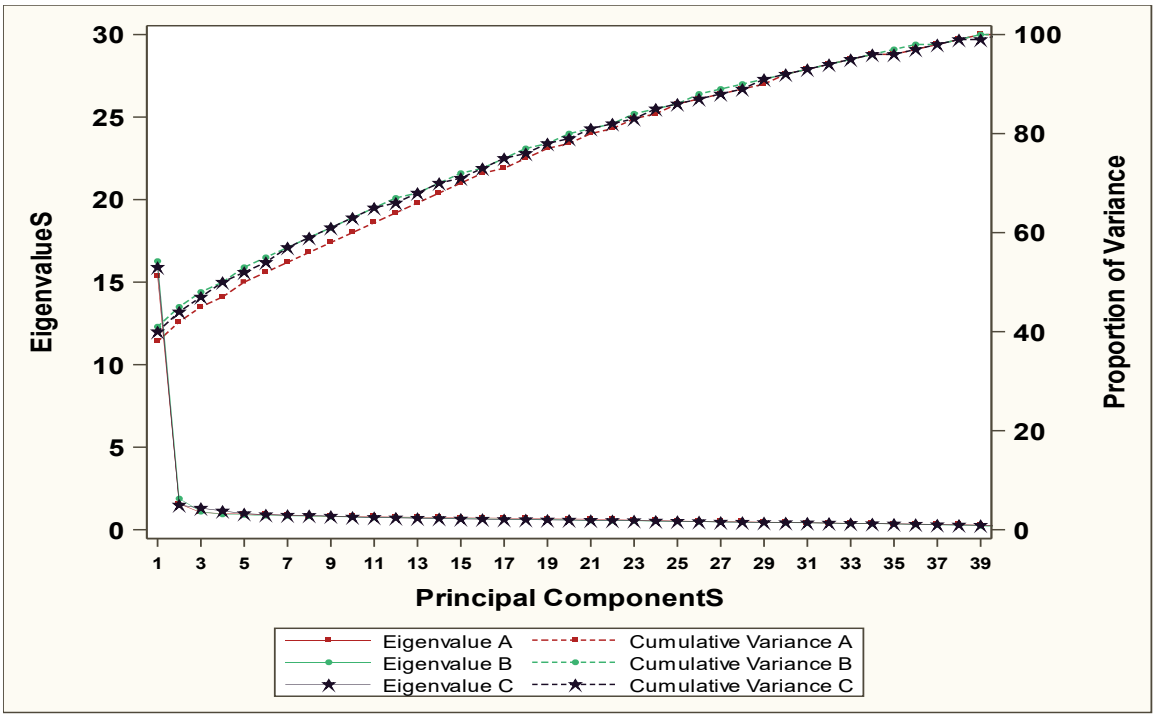



Figure 9. 4 Grade 6 PCA Scree Plot and Cumulative Variance by Form

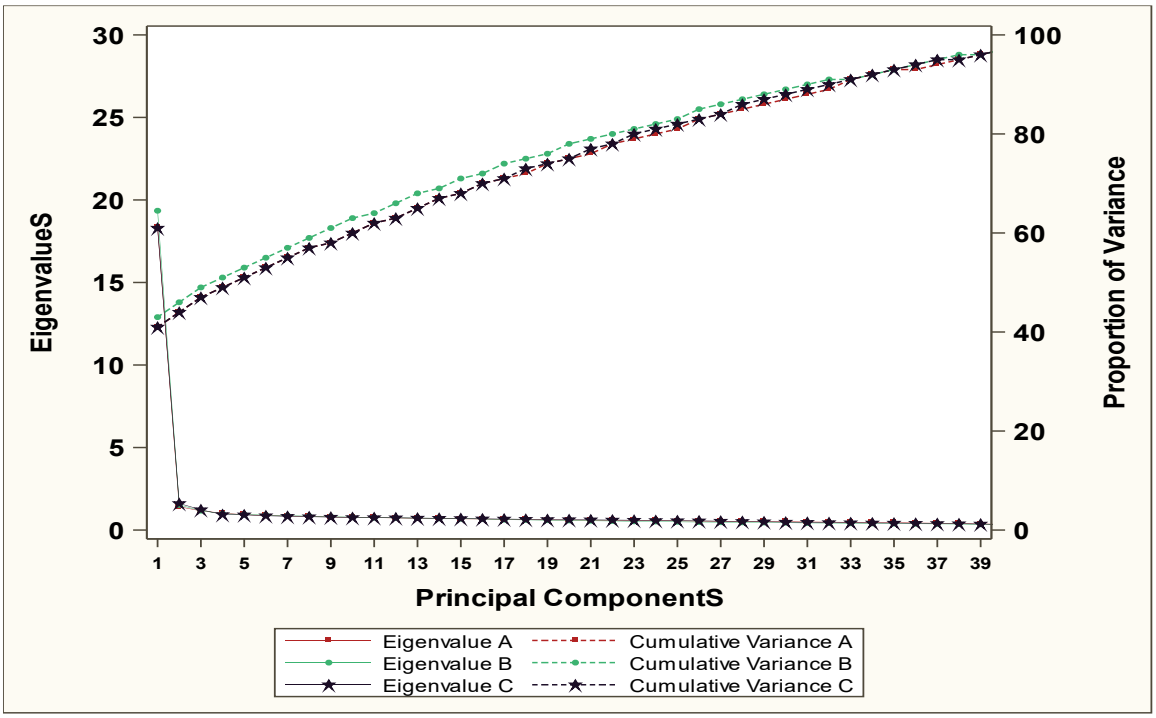


*Figure 9. 5 Grade 7 PCA Scree Plot and Cumulative Variance by Form*

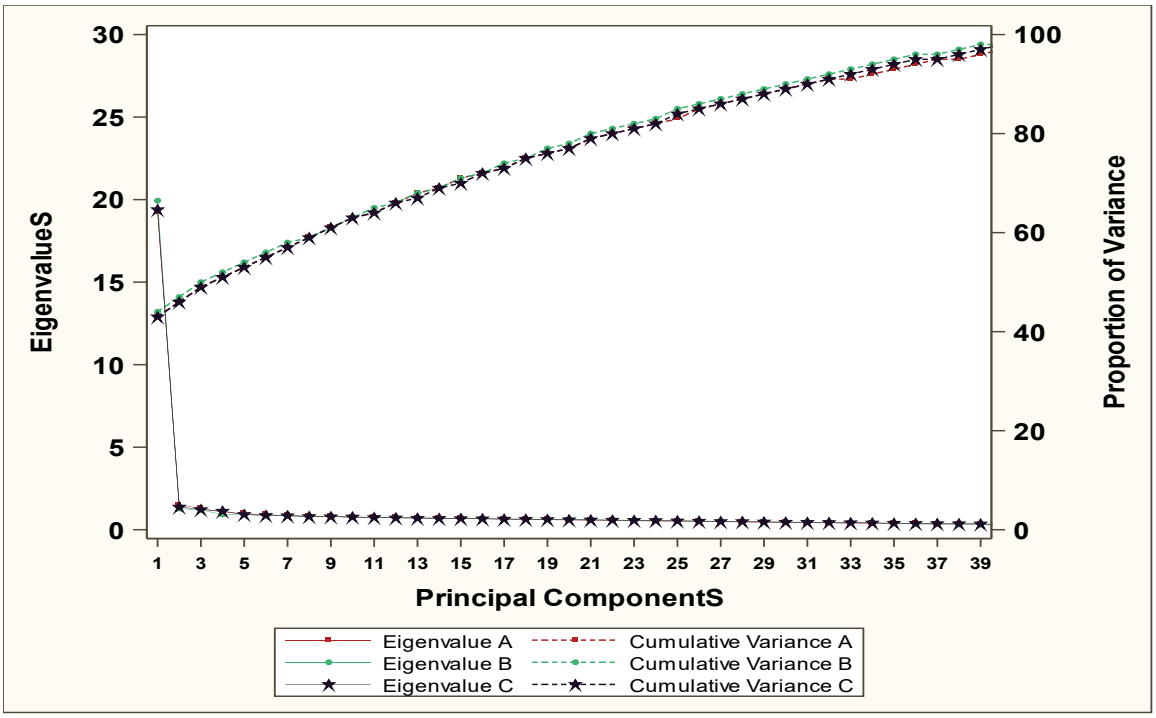



*Figure 9. 6 Grade 8 PCA Scree Plot and Cumulative Variance by Form*

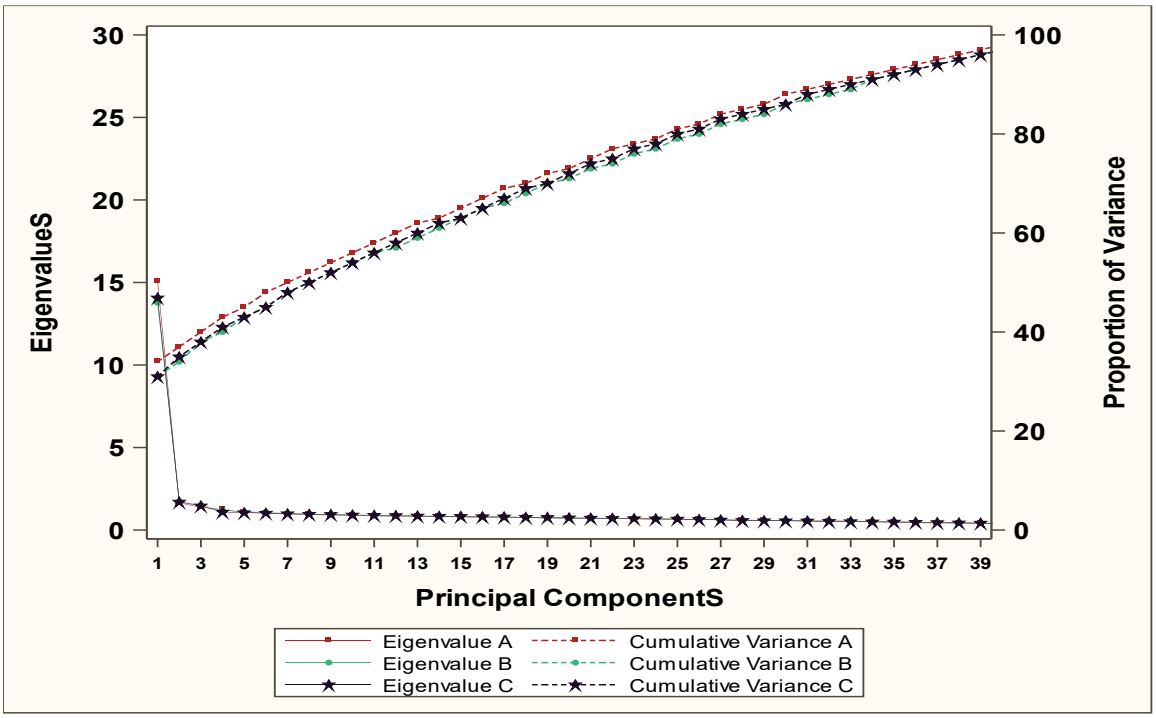


*Figure 9. 7 EOC NC Math 1 PCA Scree Plot and Cumulative Variance by Form*

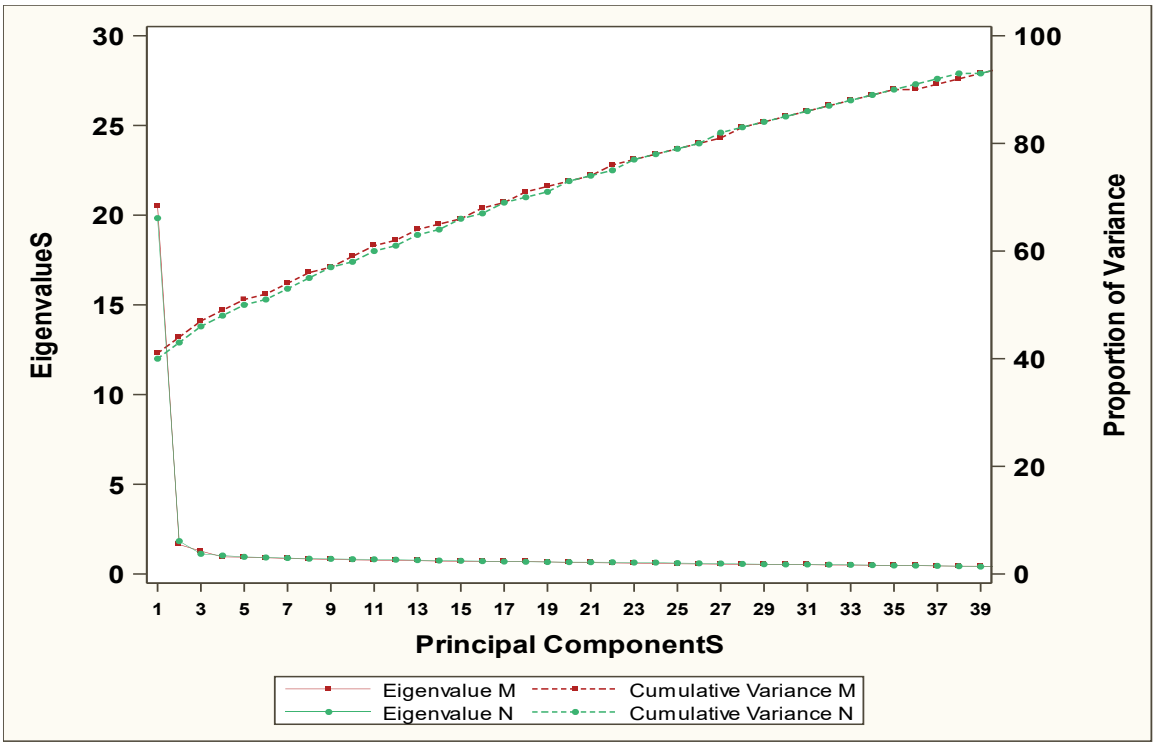



Figure 9. 8 EOC NC Math 3 PCA Scree Plot and Cumulative Variance by Form

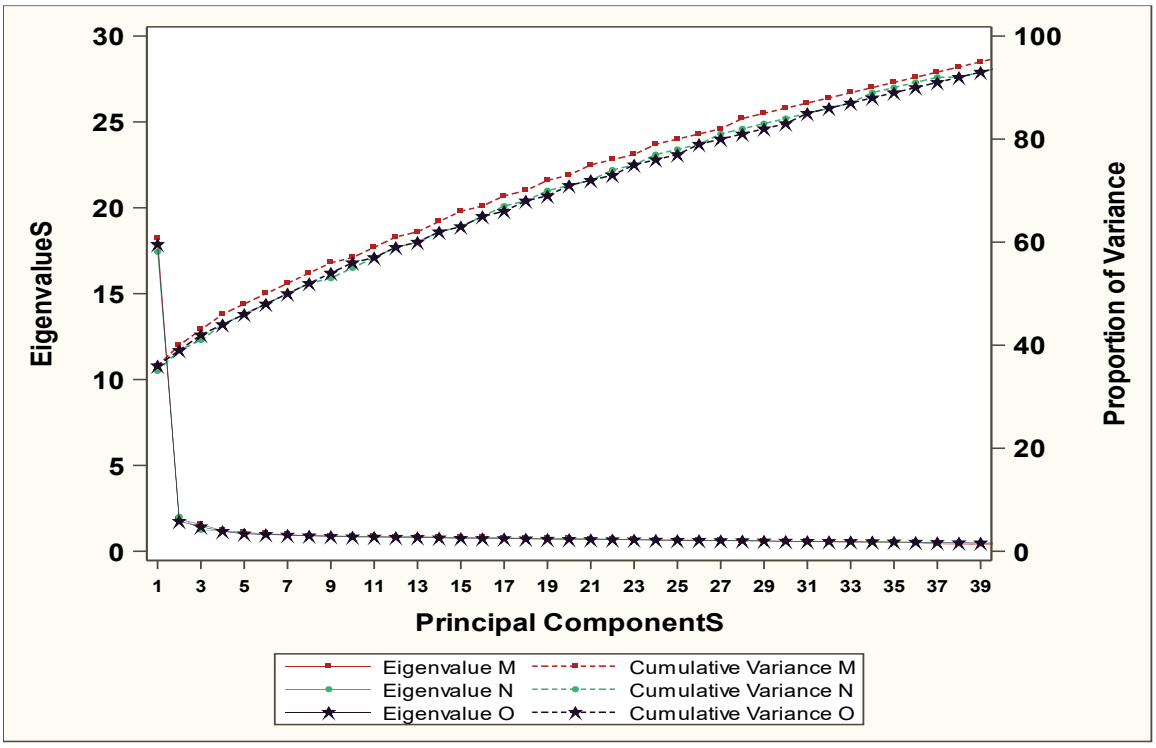


*Table 9. 6 Grades 3–5 Principal Component and Variance by Form*

| Grade | Form | Component | Computer | | | Paper | | |
|---|---|---|---|---|---|---|---|---|
| | | | Eigen | Variance | Cumulative Variance | Eigen | Variance | Cumulative Variance |
| 3 | A | 1 | 16.2 | 41% | 41% | 16.8 | 42% | 42% |
| | | 2 | 1.8 | 4% | 45% | 1.6 | 4% | 46% |
| | | 3 | 1.3 | 3% | 48% | 1.3 | 3% | 49% |
| | B | 1 | 16.7 | 42% | 42% | 17.5 | 44% | 44% |
| | | 2 | 1.6 | 4% | 46% | 1.6 | 4% | 48% |
| | | 3 | 1.2 | 3% | 49% | 1.2 | 3% | 51% |
| | C | 1 | 16.8 | 42% | 42% | 17.5 | 44% | 44% |
| | | 2 | 1.8 | 4% | 46% | 1.7 | 4% | 48% |
| | | 3 | 1.2 | 3% | 50% | 1.2 | 3% | 51% |
| 4 | A | 1 | 16.3 | 41% | 41% | 16.8 | 42% | 42% |
| | | 2 | 2.1 | 5% | 46% | 2.0 | 5% | 47% |
| | | 3 | 1.1 | 3% | 49% | 1.0 | 3% | 49% |
| | B | 1 | 16.8 | 42% | 42% | 17.2 | 43% | 43% |
| | | 2 | 1.8 | 4% | 46% | 1.6 | 4% | 47% |
| | | 3 | 1.0 | 2% | 49% | 1.0 | 3% | 49% |
| | C | 1 | 17.4 | 44% | 44% | 17.2 | 43% | 43% |
| | | 2 | 1.6 | 4% | 48% | 1.6 | 4% | 47% |
| | | 3 | 1.0 | 3% | 50% | 1.0 | 2% | 49% |
| 5 | A | 1 | 15.4 | 39% | 39% | 15.3 | 38% | 38% |
| | | 2 | 1.6 | 4% | 42% | 1.6 | 4% | 42% |
| | | 3 | 1.1 | 3% | 45% | 1.0 | 3% | 45% |
| | B | 1 | 16.2 | 41% | 41% | 16.3 | 41% | 41% |
| | | 2 | 1.9 | 5% | 45% | 1.9 | 5% | 46% |
| | | 3 | 1.1 | 3% | 48% | 1.1 | 3% | 48% |
| | C | 1 | 16.0 | 40% | 40% | 15.9 | 40% | 40% |
| | | 2 | 1.5 | 4% | 44% | 1.5 | 4% | 43% |
| | | 3 | 1.3 | 3% | 47% | 1.3 | 3% | 47% |

*Table 9. 7 Grades 6-8 Principal Component and Variance by Form*

| Grade | Form | Component | Computer | | | Paper | | |
|---|---|---|---|---|---|---|---|---|
| | | | Eigen | Variance | Cumulative Variance | Eigen | Variance | Cumulative Variance |
| 6 | A | 1 | 18.7 | 42% | 42% | 18.1 | 40% | 40% |
| | | 2 | 1.4 | 3% | 45% | 1.4 | 3% | 43% |
| | | 3 | 1.2 | 3% | 47% | 1.2 | 3% | 46% |
| | B | 1 | 19.4 | 43% | 43% | 19.3 | 43% | 43% |
| | | 2 | 1.5 | 3% | 46% | 1.6 | 3% | 46% |
| | | 3 | 1.2 | 3% | 49% | 1.2 | 3% | 49% |
| | C | 1 | 18.7 | 41% | 41% | 17.9 | 40% | 40% |
| | | 2 | 1.6 | 4% | 45% | 1.6 | 4% | 43% |
| | | 3 | 1.2 | 3% | 48% | 1.2 | 3% | 46% |
| 7 | A | 1 | 19.8 | 44% | 44% | 18.6 | 41% | 41% |
| | | 2 | 1.5 | 3% | 47% | 1.5 | 3% | 45% |
| | | 3 | 1.3 | 3% | 50% | 1.3 | 3% | 48% |
| | B | 1 | 20.3 | 45% | 45% | 19.6 | 43% | 43% |
| | | 2 | 1.3 | 3% | 48% | 1.3 | 3% | 46% |
| | | 3 | 1.2 | 3% | 51% | 1.2 | 3% | 49% |
| | C | 1 | 19.7 | 44% | 44% | 19.1 | 42% | 42% |
| | | 2 | 1.4 | 3% | 47% | 1.4 | 3% | 45% |
| | | 3 | 1.2 | 3% | 50% | 1.2 | 3% | 48% |
| 8 | A | 1 | 15.1 | 34% | 34% | . | . | . |
| | | 2 | 1.7 | 4% | 37% | . | . | . |
| | | 3 | 1.4 | 3% | 40% | . | . | . |
| | B | 1 | 13.8 | 31% | 31% | . | . | . |
| | | 2 | 1.7 | 4% | 34% | . | . | . |
| | | 3 | 1.5 | 3% | 38% | . | . | . |
| | C | 1 | 14.1 | 31% | 31% | . | . | . |
| | | 2 | 1.7 | 4% | 35% | . | . | . |
| | | 3 | 1.5 | 3% | 38% | . | . | . |

*Table 9. 8 EOC Mathematics Principal Component and Variance by Form*

| Grade | Form | Component | Computer | | |
|---|---|---|---|---|---|
| | | | Eigen | Variance | Cumulative Variance |
| NC Math 1 | M | 1 | 20.5 | 41% | 41% |
| | | 2 | 1.6 | 3% | 44% |
| | | 3 | 1.3 | 3% | 47% |
| | N | 1 | 19.8 | 40% | 40% |
| | | 2 | 1.8 | 4% | 43% |
| | | 3 | 1.1 | 2% | 46% |
| NC Math 3 | M | 1 | 18.2 | 36% | 36% |
| | | 2 | 1.9 | 4% | 40% |
| | | 3 | 1.6 | 3% | 43% |
| | N | 1 | 17.4 | 35% | 35% |
| | | 2 | 2.0 | 4% | 39% |
| | | 3 | 1.3 | 3% | 41% |
| | O | 1 | 17.9 | 36% | 36% |
| | | 2 | 1.8 | 3% | 39% |
| | | 3 | 1.4 | 3% | 42% |

## 9.5 Alignment Study

Alignment in large scale assessment refers to how well the assessment items and the assessment framework as a whole reflect the intended academic content and performance standards on which they are based. The collection of alignment evidences for the North Carolina assessments started from the item writing and test development phase where TMSs from the TOPS and the NCDPI as well as Psychometricians were responsible for training item writers for writing items aligned to academic content standards, selection of items representing test blueprint, performance expectations in terms of cognitive complexities or DOKs and creating a test reflecting target difficulty.

A formal alignment study quantifying the degree of alignments in the major outcome variables is planned for 2020–21 administration. The NCDPI has awarded contract to edCount[3] for the alignment study.

[3] Copyright 2013-19 edCount, LLC

## 9.6 Evidence Regarding Relationships with External Variables

One of the primary purposes of the EOG and EOC mathematics assessments is to provide data to measure students' achievement and progress relative to readiness as defined by college-and career-readiness standards. For the mathematics assessments to provide external evidence of this type of achievement, it is important to appropriately match students with materials at a level where the student has the background knowledge necessary to be ready for instruction on the new mathematical skills and concepts. To examine the mathematics achievement levels that can be matched with mathematics skills and concepts based on the North Carolina EOG and EOC mathematics assessments, the NCDPI commissioned MetaMetrics Inc.[4] to examine the relationship of the mathematics assessments to the Quantile Framework for mathematics (Request for Quote #: 40-RQ21164619, dated July 20, 2018). The primary purpose of this study was to:

- provide the NCDPI with Quantile measures on the North Carolina EOG mathematics and NC Math 1 assessments;
- provide tools (e.g., Quantile Math@Home, Quantile Teacher Assistant and Quantile Math Skills Database) and information that can be used to answer questions related to standards, student-level accountability, test score interpretation and test validation;
- develop tables for converting North Carolina EOG mathematics and NC Math 1 scale scores to Quantile measures; and
- produce a report that describes the linking analysis procedures.

The sections below summarize important evidences of the relationship and list findings from the report. The full report is included in *Appendix 9–B*. The report contains the North Carolina mathematics grades 3–8 and NC Math 1 Quantile Framework Linking Process, including design, sampling, item calibration and scoring, linking to quantile scale and characteristics of the linking items. NC Math 3 is not linked to Quantile Framework.

### 9.6.1 The Quantile Framework for Mathematics

The Quantile Framework is a scale that describes a student's mathematical achievement and uses a common metric—the Quantile—to scientifically measure a student's ability to reason mathematically, monitor a student's readiness for mathematics instruction and locate a student on its taxonomy of mathematical skills, concepts and applications. It was developed to assist teachers, parents and students in identifying strengths and weaknesses in mathematics and forecast growth in overall mathematical achievement. Items and mathematical content are calibrated using the Rasch IRT model. The Quantile Framework spans the developmental continuum from kindergarten mathematics through high school Algebra, Geometry,

[4] © 2020 MetaMetrics Inc.

Trigonometry and Precalculus. The Quantile scale ranges from below EM400Q to above 1600Q ("EM" – Emerging Mathematician, below 0Q).

The Quantile Framework was developed to assess how well a student 1) understands the natural language of mathematics, 2) knows how to read mathematical expressions and employ algorithms to solve decontextualized problems and 3) knows why conceptual and procedural knowledge is important and how and when to apply it. The Quantile Item Bank consists of multiple-choice items aligned with first-grade content through Geometry, Algebra II and Pre-calculus content and was field tested with a national sample of students during the winter of 2004.

For the Quantile Framework, which measures student understanding of mathematical skills and concepts, the most important aspect of validity that should be examined is construct-identification validity. The Quantile Framework evaluates content-description and criterion-prediction validity. MetaMetrics Inc. has collected a good amount of validity evidence to show how well Quantile measures relate to other measures of mathematics: 1) standardization set of items used with PASeries mathematics, 2) relationship of Quantile measures to other measures of mathematical ability, 3) Quantile Framework linked to other measures of mathematics understanding and 4) multidimensionality of the Quantile Framework items.

### 9.6.2 Linking the Quantile Framework to the NC Assessments

The Quantile linking test was constructed by aligning the items from the North Carolina EOG and EOC mathematics assessments for grades 3–8 and NC Math 1 with the Quantile Framework taxonomy of Quantile Skills and Concepts (QSCs). Based upon these target test reviews, previously tested items in quantile scale were embedded in each grade level linking test. To achieve this alignment, the content of the North Carolina EOG Mathematics and NC Math 1 assessment blueprints were matched to corresponding QSCs. Quantile linking items were all MC with known statistics and selected to maximize the alignment with the North Carolina content standard blueprints. The comparability of the material includes the number of operational items per test, the distribution of the content strands (which are closely matched to the distribution of the domains from the North Carolina Standard Course of Study) and the difficulty of the items. The linking study was conducted using linear equating. Separate linking functions were developed for each grade since EOG and EOC tests are not on a vertical scale.

As a part of the evaluation of the linking design, the classical statistics between the previously known statistics from the norm samples and statistics from the North Carolina samples were compared. Preliminary results indicated that the NC students were a more able group than the population used to develop the Quantile user norms, especially in grade 3 through grade 5. Given the unusually high distributions of Quantile measures in the linking samples, further examination was undertaken to see how the results compared with the 2013 linking study norms comparisons.

A clear distinction between the distributions appeared in Grades 3 through 5. The 2019 preliminary link values were markedly higher than the Quantile user norms as compared to the 2013 results. Therefore, for grade 3 through grade 5, the Quantile means and standard deviations from the current study were replaced with the Quantile means and standard deviations from the 2013 link (*see Appendix 9-B*). All other grades were well targeted and supported using the current data collection.

*Figure 9.9* shows the Quantile measures for the North Carolina EOG and EOC mathematics assessments from the linking sample and the Quantile norms before adjusting for grades 3–5. *Figure 9.10* shows the quantile measures after the grades 3–5 adjustment was made. A clear distinction between the distributions in *Figures 9.9* and *9.10* appears in Grades 3 through 5. As can be seen in *Figures 9.9 and 9.10*, the Quantile measures for the North Carolina EOG and EOC mathematics assessments are higher than the Quantile measure norms. This indicates that the EOG/EOC population in this study is more able than the samples used for the Quantile norms. Because the Quantile user norms consist of interim and summative assessment results, when comparing a summative assessment like North Carolina EOG Grades 3 through 8 mathematics and NC Math 1, higher Quantile measures would be expected with respect to the Quantile user norms.

*Figure 9. 9 Selected Percentiles (25th, 50th and 75th) Plotted for the North Carolina EOG Grades 3-8 Mathematics and NC Math 1 Quantile Measures for the linking Sample (N=661,766)*

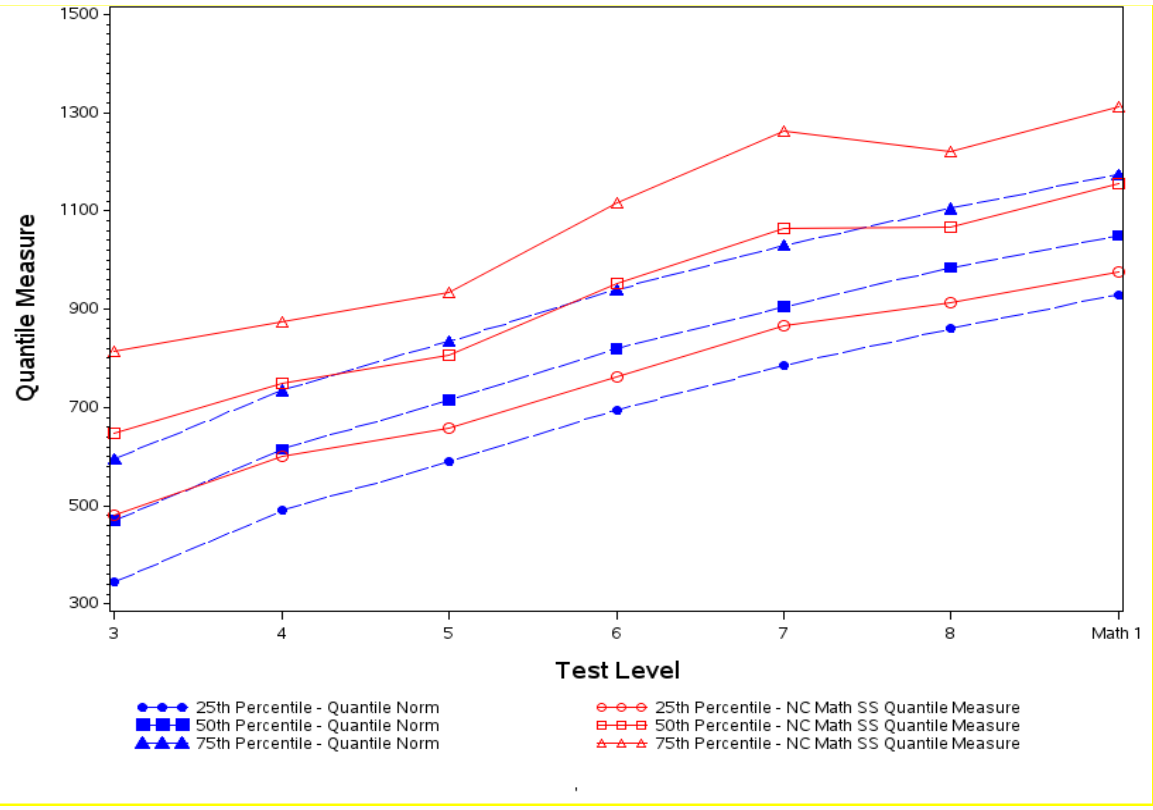


*Figure 9. 10 Selected Percentiles (25th, 50th, and 75th) Plotted for the NC READY EOG Mathematics/EOC Algebra I/Integrated I Quantile Measures for the Final Sample (N = 8,720) in Relation to the Quantile Norms (MetaMetrics, 2014).*

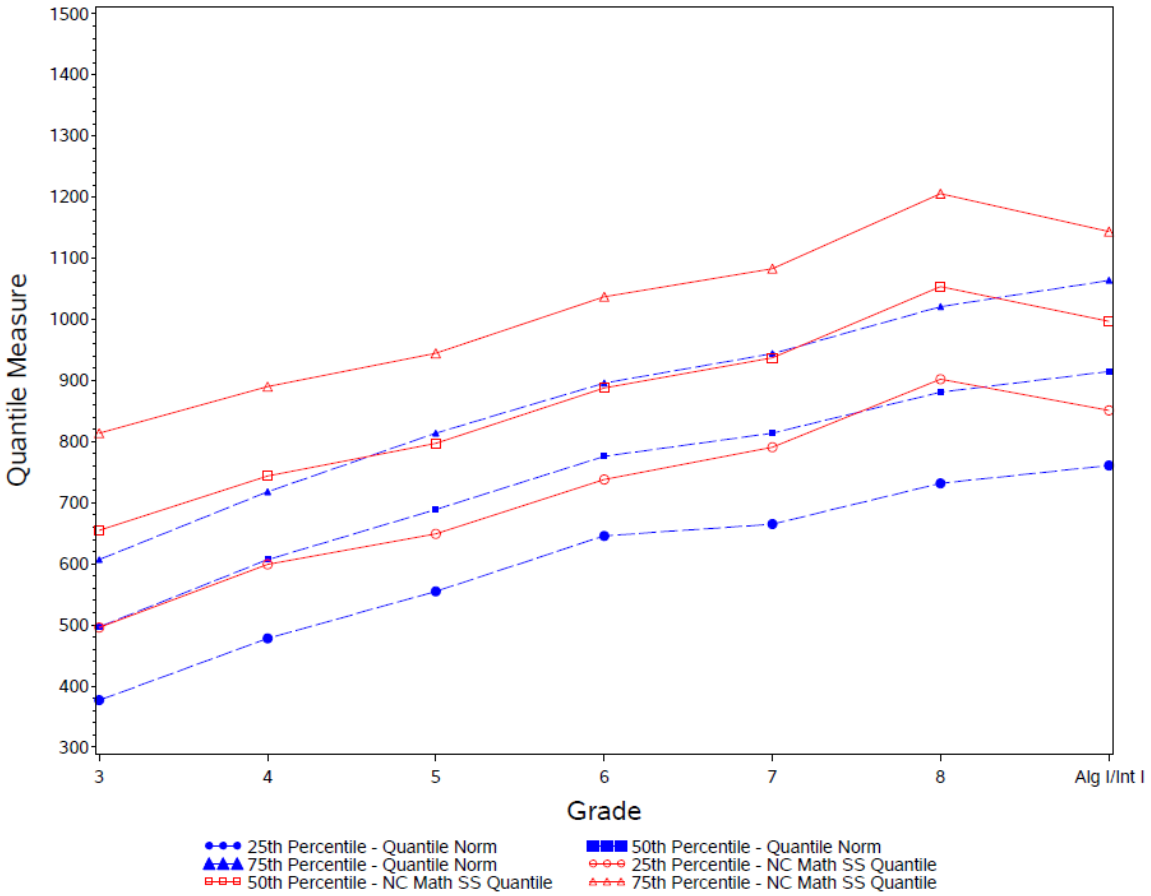



*Table 9.9* presents the achievement level cut scores on the North Carolina EOG mathematics and NC Math 1 assessments and the associated Quantile measures. The NCDPI established four achievement levels: Level 2& Below, Level 3, Level 4 and Level 5 (*see Appendix 7-A*). The values in the table are the cut scores associated with the bottom score of proficiency levels (2 & below, 3, 4 and 5) for each category.

*Table 9. 9 NC EOG Grades 3–8 and NC Math 1 Scale Scores and Quantile Measures for Achievement Levels*

| Grade | Not Proficient | | Level 3 | | Level 4 | | Level 5 | |
|---|---|---|---|---|---|---|---|---|
| | Scale Score Range | Quantile Measure Range | Scale Score Range | Quantile Measure Range | Scale Score Range | Quantile Measure Range | Scale Score Range | Quantile Measure Range |
| 3 | 520–544 | EM65Q–525Q | 545–550 | 530Q–665Q | 551–559 | 670Q–880Q | 560–570 | 885Q–975Q |
| 4 | 520–546 | 115Q–680Q | 547–551 | 685Q–785Q | 552–559 | 790Q–955Q | 560–570 | 960Q–1075Q |
| 5 | 520–545 | 165Q–715Q | 546–550 | 720Q–825Q | 551–560 | 830Q–1035Q | 561–570 | 1040Q–1125Q |
| 6 | 525–545 | 270Q–840Q | 546–550 | 845Q–975Q | 551–560 | 980Q–1250Q | 561–573 | 1255Q–1280Q |
| 7 | 525–545 | 385Q–975Q | 546–549 | 980Q–1085Q | 550–559 | 1090Q–1370Q | 560–573 | 1375Q–1430Q |
| 8 | 515–542 | 515Q–1130Q | 543–547 | 1135Q–1240Q | 548–554 | 1245Q–1390Q | 555–570 | 1395Q–1450Q |
| NC Math 1 | 525–547 | 510Q–1100Q | 548–554 | 1105Q–1280Q | 555–562 | 1285Q–1485Q | 563–575 | 1490Q–1510Q |

### 9.6.3 The Quantile Framework and College and Career Readiness

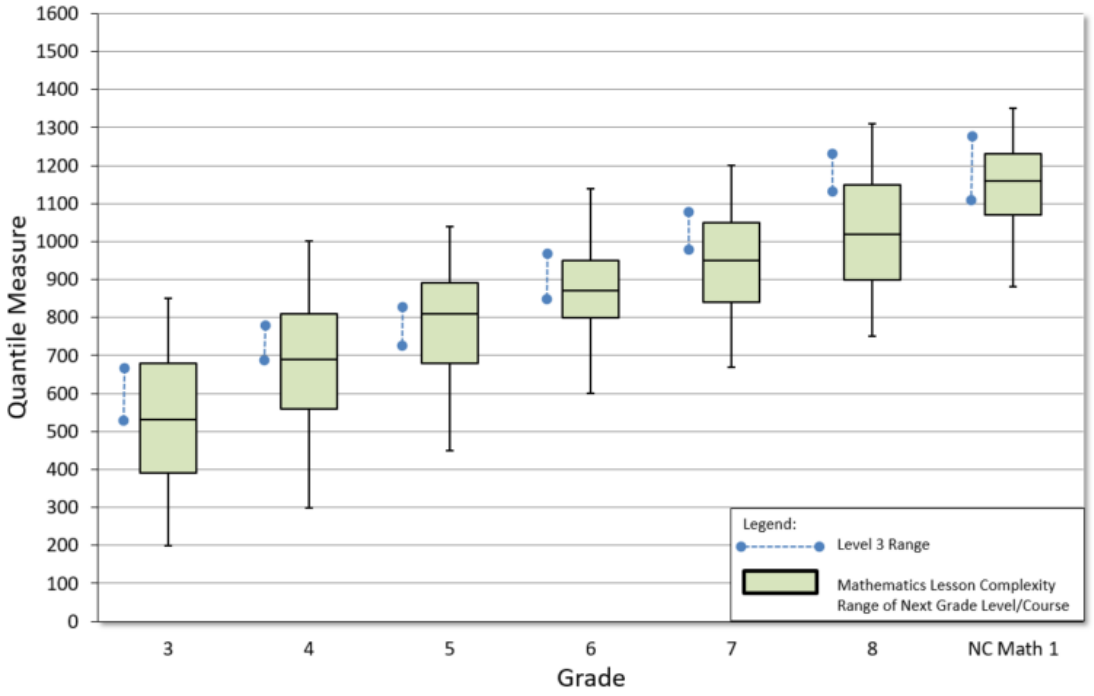
As noted above, one purpose of this study was to examine the mathematics level associated with the North Carolina EOG and EOC assessments. If these assessments are to provide information about CCR, then the mathematics level of the assessments must align to Quantile measures associated with CCR. It would undermine the credibility of the North Carolina assessments to measure CCR if the mathematics levels of the mathematics assessments were, say, below grade level.

Williamson, Sanford-Moore and Bickel (2016) began the examination of the mathematics demands of college and careers to answer the question, “What mathematics will a student likely encounter when entering college or a career?” To address this question, the mathematical concepts and skills that students are likely to encounter as they begin their postsecondary education and/or enter the workplace were examined. For college, being ready for instruction in the types of courses typical of those beyond high school graduation requirements and of first year college were examined (e.g., Precalculus, Trigonometry). For careers, competently performing the mathematics content required for a high school diploma (e.g., Algebra I content, Algebra II content) was examined. In this research, “competently perform” was defined as a 75% understanding of the mathematics skills and concepts. The range (interquartile range) of mathematical demands students are likely to encounter as they enter college- and career-readiness is 1220Q to 1440Q, with a median of 1350Q.

MetaMetrics research on the mathematical demand of college and career readiness can be used to compare achievement levels from the EOG Grades 3 through 8 mathematics and NC Math 1 assessments with the mathematics skills and concepts a student will likely encounter. *Figure 9.11* shows the relationship between the “Level 3” achievement level of the EOG Grades 3

through 8 mathematics and NC Math 1 Quantile measures and the mathematics lesson complexity ranges for the next grade level/course. For each grade/level, the box refers to the interquartile range. The line within the box indicates the median. The end of each whisker represents the 5th percentile at the low end of the distribution of mathematical demand distribution and the 95th percentile at the high end of the distribution. Level 3 achievement is within the mathematics lesson complexity ranges for the next grade level/course across all grades. This supports the interpretation that students at "Level 3" or above will be able to successfully engage with the material at the next grade level.

*Figure 9. 11 Comparison of NC EOG Grades 3-8 Mathematics and NC Math 1 Quantile Measures for the Level 3 Achievement Level and the Mathematical Demand at the Next Grade.*

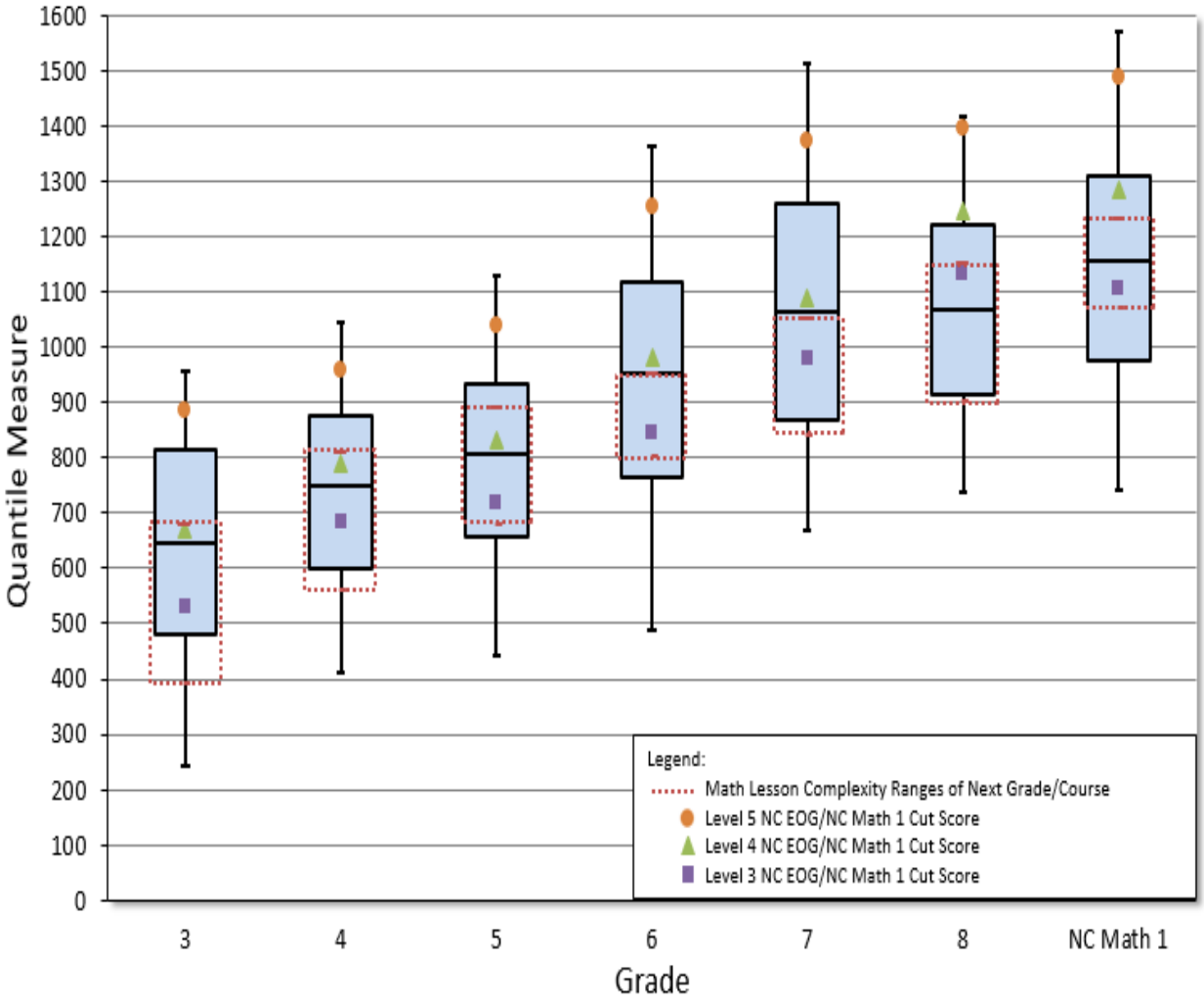



To better understand the results from the current Quantile linking study, student achievement levels from the EOG grades 3-8 mathematics and NC Math 1 assessments were compared with the distribution of student scores as Quantile measures and the mathematical demands of the instructional materials the students will likely encounter. `

*Figure 9.12* shows the spring 2019 student results from the EOG grades 3–8 mathematics and NC Math 1 assessments as Quantile measures. For each test level, the box refers to the interquartile range of student results. The line within the box indicates the median. The end of each whisker represents the 5th percentile at the low end of the distribution of scores and the 95th percentile at the high end of the distribution. The square, triangle and circle represent the EOG grades 3–8 mathematics and NC Math 1 achievement level cut scores as Quantile measures for "Level 3," "Level 4," and "Level 5," respectively. Additionally, the dotted box provides a

reference for the complexity of lessons students will encounter at the next grade level in mathematics. All grades show that the Level 3 cut point is within or above the range of the mathematical demands of the following school year's mathematics content. Ultimately, placing all the information on the same scale allows students to be matched with instructional materials targeted to the skills and concepts students will likely encounter as they enter the next grade level and, ultimately, as they enter college and careers.

*Figure 9. 12 NC EOG Grades 3 Through 8 Mathematics and NC Math 1 Student Achievement (Spring 2019) Expressed as Quantile Measures Compared to the NC EOG and NC Math 1 Student Achievement Levels and Mathematical Lesson Demand Distributions.*




In 2013, the NCReady data was linked with the Quantile scale for Grades 3 through 8 and Algebra I/Integrated Math 1. With the introduction of the new version of the North Carolina EOG grades 3–8 mathematics and NC Math 1 assessments in 2019, achievement levels for grades 3–8 and NC Math 1 were redefined, and new cut scores were identified (see *Table 9.10*).

*Table 9. 10 Comparison of North Carolina Achievement Levels*

| Grade | NC Ready Level 3 Quantile Cut Point | NC EOG/NC Math 1 Level 3 Quantile Cut Point |
|---|---|---|
| 3 | 610Q | 530Q |
| 4 | 725Q | 685Q |
| 5 | 775Q | 720Q |
| 6 | 910Q | 845Q |
| 7 | 960Q | 980Q |
| 8 | 1095Q | 1135Q |
| NC Math 1 | 1020Q | 1105Q |

### 9.6.4 Summary of Quantile Linking Framework

The North Carolina assessments were linked to the Quantile Framework as a means of collecting external evidence on the rigor of the NC assessments in relation to the demands of college- and career-readiness standards. This study showed that the mathematics levels of the NC assessments are aligned with expectations of college- and career-readiness as measured by the Quantile Framework. In addition, this study showed that the rigor of mathematics measured by the NC assessments has increased since the previous version of the assessments.

A caveat of the 2018–19 North Carolina Quantile linking study was that the classical and calibrated student Quantile measures in grades 3 through 5 were higher than both to the Quantile user norms and previous results from the 2013 North Carolina Quantile linking study. Therefore, the means and standard deviations of Quantile measures for grades 3 through 5 were replaced with those from the 2013 linking study. The primary purpose of the Quantile Framework is to provide appropriate instructional materials for students given their ability level. The results observed in Grades 3 through 5 necessitated the adjustment to avoid overestimation of Quantile measures and avoid assigning instructional materials that are too challenging for students. For grades 6 through 8, and NC Math 1, however, the means and standard deviation of the current study were reasonable.

## 9.7 Fairness and Accessibility

### 9.7.1 Accessibility in Universal Design

To ensure fairness and accessibility for all eligible students for NC assessments, the principle of universal design was embedded throughout the development and design of EOG and EOC assessments. The EOG and EOC assessments measure student's knowledge as defined in the *North Carolina State Content Standards*. Assessments must ensure comprehensible access to the content being measured to allow students to accurately demonstrate their standing in the content assessed. In order to ensure items and assessments were developed with universal design principles, the NCDPI train item writer and reviewers with "Plain English Principles".

Evidence of universal design principles applied in the development of EOG and EOC assessments (so that students could show what they know) has been documented throughout the item development and review, form review and test administration sections in the report. Some of the universal design principles used in the training include:

- Precisely defined constructs
    - Direct match to objective being measured
- Accessible, nonbiased items[5]
    - Accommodations included from the start (Braille, large–print, oral presentation etc.)
    - Ensuring that quality is retained in all items
- Simple, clear directions and procedures
    - Presenting in understandable language,
    - Using simple, high frequency and compound words,
    - Using words that are directly related to content the student is expected to know,
    - Omitting words with double meanings or colloquialisms,
    - Consistency in procedures and format in all content areas.
- Maximum legibility
    - Simple fonts
    - Use of white space
    - Headings and graphic arrangement
    - Direct attention to relative importance
    - Direct attention to the order in which content should be considered
- Maximum readability:
    - plain language
    - Increases validity to the measurement of the construct
    - Increases the accuracy of the inferences made from the resulting data

[5] See discussions on fairness review in Chapter 4

    - Active instead of passive voice
    - Short sentences
    - Common, everyday words
    - Purposeful graphics to clarify what is being asked
- Accommodations
    - One item per page
    - Extended time for ELs Students
    - Test in a separate room
- Computer–based Forms
    - All students receive one item per test page,
    - All students may receive larger font and different background colors.

### 9.7.2 Fairness in Access

Alignment evidence, presented throughout Chapter 2 through Chapter 6, demonstrated that the NCDPI commitment that all assessment blueprints are aligned to content domains that are also aligned to the NCSCOS. Assessments' content domain specifications and blueprints are published on the NCDPI public website with other relevant information regarding the development of EOG and EOC assessments. This ensures schools and students have exposure to content being targeted in the assessments and thus provides them with an opportunity to learn.

Prior to the administration of the first operational form of EOG and EOC assessments, the NCDPI also published released forms for every grade level, which were constructed using the same blueprint as the operational forms. These released forms provided students, teachers and parents with sample items and a general practice form that is similar to the operational assessment. These released forms also served as a resource to familiarize students with the various response formats in the new assessments.

### 9.7.3 Fairness in Administration

Chapter 5 of this report documents the procedures put in place by the NCDPI to assure that the administration of the EOG and EOC assessments are standardized, fair and secured for all students across the state. For each assessment, the NCDPI publishes an *North Carolina Test Coordinators' Policies and Procedures Handbook* that is the main training material for all test administrators across the state. These guides provide comprehensive details of policies and procedures for each assessment including general overview of each assessment that covers the purpose of the assessment, student eligibility, testing window and makeup testing options. Assessment guides also cover all preparations and steps that should be followed the day before testing, on test day and after testing. Samples of answer sheets are also provided in the assessment guide. In addition to assessment guides used to train test administrators, the NCDPI also publishes a *Proctor Guide* that is used by test coordinators for training proctors.

Computer-based assessments are available to all students in regular or large font and in alternate background colors; however, the NCDPI recommends these options be considered only for students who routinely use similar tools (e.g., color acetate overlays, colored background paper and large print text) in the classroom. It is recommended that students be given the opportunity to view the large font and/or alternate background color versions of the online tutorial and released forms of the assessment (with the device to be used on test day) to determine which mode of administration is appropriate.

Additionally, the NCDPI recommends that the Online Assessment Tutorial should be used to determine students' appropriate font size (i.e., regular or large) and/or alternate background color for test day. These options must be entered in the student's interface questions (SIQ) before test day. The Online Assessment Tutorial can assist students, whose IEP or Section 504 Plan designates the Large Print accommodation in determining, whether the large font will be adequate for the student on test day. If the size of the large font is insufficient for a student because of his/her disability, this accommodation may be used in conjunction with the *Magnification Devices* accommodation, or a *Large Print Edition* of the paper-and-pencil assessment may be ordered.

In order to prepare students for gridded response items in their upcoming EOG mathematics grades 5–8, NC Math 1, and NC Math 3 assessments, the NCDPI produced practice activities for using the grids. The NCDPI requires students take the gridded response practice activity before the administration of the EOG grades 5–8, NC Math 1, and NC Math 3 assessments. Schools must ensure that every student participating in the grades 5–8 EOG mathematics assessments complete the grade-appropriate gridded response practice activity at least one time at the school before test day.

### 9.7.4 Fairness Across Forms and Modes

The AERA, APA, & NCME (2014) states, "*When multiple forms of a test are prepared, the same test specifications should govern all of the forms.*" It is imperative that when multiple forms are created from the same test blueprint, the resulting test scores from parallel forms are comparable; and it should make no difference to students which form was administered. For EOG and EOC assessments, parallel forms were created based on the same content and statistical specifications. As shown in Chapter 4 and Chapter 6, all parallel forms were constructed and matched to have the same CTT and IRT properties of average p-value, reliability and closely aligned TCCs as well as CSEM. Meeting these criteria ensured that the test forms are essentially parallel. Moreover, these forms were spiraled within class to obtain equivalent samples for calibration and scaling. This ensured that each form was administered to a random-equivalent sample of students across the state. Any difference in form difficulty was accounted for during separate group calibration as the random-group data design ensured all parameters were placed

onto the same IRT scale and separate raw-to-scale tables were created to adjust for any form differences.

To ensure that scores from forms administered across mode (paper and computer) were comparable, the DIF sweep procedure was implemented during item analysis. The DIF sweep procedure flags items that show a significant differential item parameter between computer and paper modes. These items, though identical, are treated as unique items during joint calibration of computer and paper forms. The process involved two steps: in step 1, items were calibrated in each mode separately and their estimated item parameters were evaluated. If the estimated parameters are within the set threshold showing no evidence of a mode effect then the two sets of responses were concurrently calibrated to estimate the final item parameters. If the estimated parameters are outside the set threshold showing a sign of mode effect, then in step 2 those items that exhibited no DIF were considered anchors and a separate set of item parameters were estimated for each item by mode that exhibited DIF. This process ensured that the item parameters and test scores were on a common IRT scale and that mode effects were accounted for. Finally, the resulting item parameters were used to create a separate raw-to-scale score table for each form by modes.

To ensure equitable access for students taking computer-based forms, the NCDPI has set minimum device requirements that will guarantee all items and forms will exhibit acceptable functionality as intended. These requirements are based on a review of industry standards and usability studies and research findings conducted with other national testing programs. The NCDPI device requirements for EOG and EOC computer-based assessments include:

- A minimum screen size of 9.5 inches
- A minimum screen resolution of 1024 x 768
- iPads must use Guided Access or a Mobile Device management system to restrict the iPad to only run the NCTest iPad App.
- Screen capture capabilities must be disabled.
- Chrome App on desktops and laptops requires the Chrome Browser version 43 or higher.
- Windows machines must have a minimum of 512 MB of RAM.
- A Pentium 4 or newer processor for Windows machines and Intel for MacBooks

In addition to the technical specification of devices, the NCDPI also conducts a review of each sample item across devices (i.e., laptops, iPads and desktops) to make sure items are rendered as intended. Reviews also check functionalities of the test platform, such as audio files, large font and high contrast versions.

## Glossary of Key Terms

The terms below are defined by their application in this document and their common uses in the North Carolina Testing Program. Some of the terms refer to complex statistical procedures used in the process of test development. In an effort to avoid excessive use of technical jargon, definitions have been simplified; however, they should not be considered exhaustive.

| Key Terms | Definition |
|---|---|
| Accommodations | Changes made in the format or administration of the test to provide options to test takers who are unable to take the original test under standard test conditions. |
| Achievement Levels | Descriptions of a test taker's competency in a particular area of knowledge or skill, usually defined as ordered categories on a continuum classified by broad ranges of performance. |
| Asymptote | An item statistic that describes the proportion of examinees who endorsed a question correctly but did poorly on the overall test. Asymptote for a theoretical four-choice item is 0.25 but can vary somewhat by test. |
| Biserial Correlation | The relationship between an item score (right or wrong) and a total test score. |
| Cut Scores | A specific point on a score scale, such that scores at or above that point are interpreted or acted upon differently from scores below that point. |
| Dimensionality | The extent to which a test item measures more than one ability. |
| Embedded Field-Test Design | Using an operational test to FT new items or sections. The new items or sections are "embedded" into the new test and appear to examinees as being indistinguishable from the operational test. |
| Equivalent Forms | The differences between forms are not statistically significant. |
| Field-Test | A collection of items to approximate how a test form will work. Statistics produced will be used in interpreting item behavior/performance and allow for the calibration of item parameters used in equating tests. |
| Foil Counts | Number of examinees that endorse each foil (e.g., number who answer "A," number who answer "B," etc.). |

| Key Terms | Definition |
|---|---|
| Item Response Theory | A method of test item analysis that takes into account the ability of the examinee and determines characteristics of the item relative to other items in the test. The NCDPI uses the 3-parameter model, which provides slope, threshold and asymptote. |
| | |
| Mantel-Haenszel | A statistical procedure that examines the differential item functioning (DIF) or the relationship between a score on an item and the different groups answering the item (e.g. gender, race). This procedure is used to identify individual items for further fairness review. |
| Operational Test | Test administered statewide with uniform procedures, full reporting of scores and stakes for examinees and schools. |
| P–value | Difficulty of an item defined by using the proportion of examinees who answered an item correctly. |
| Parallel Forms | Forms that are developed with the same content and statistical specifications. |
| Percentile | The score on a test below which a given percentage of scores fall. |
| | |
| Raw Score | The unadjusted score on a test determined by counting the number of correct answers. |
| Scale Score | A score to which raw scores are converted by numerical transformation. Scale scores allow for comparison of different forms of the test using the same scale. |
| Slope | The ability of a test item to distinguish between examinees of high and low ability. |
| Standard Error of Measurement | The standard deviation of individuals' observed scores, usually estimated from group data. |
| Test Blueprint | The testing plan, which includes the numbers of items from each objective that are to appear on a test and the arrangement of objectives. |
| Threshold | The point on the ability scale where the probability of a correct response is fifty percent. Threshold for an item of average difficulty is 0.00. |

## References

AERA, APA, & NCME (2014). Standards for educational and psychological testing. Washington, D.C.: Author.

Anastasi, A., & Urbina, S. (1997). Psychological testing. (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Brennan, R. L. (2004). Manual for BB-CLASS: A computer program that uses the Beta-Binomial model for classification consistency and accuracy. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment (CASMA).

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Lincolnwood, IL: Scientific Software International.

Camilli, G. & Shepard, L.A. (1994). Methods for Identifying Biased Test Items. Thousand Oaks, CA: Sage Publications, Inc.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 22(3), 297–334.

Hambleton, R. K. (2000). Advances in Performance Assessment Methodology. Applied Psychological Measurement, 24(4), 291-293: © 2000 Sage Publication, Inc.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp. 89-116). Mahwah, NJ: Erlbaum. Hambleton, R.K. & Swaminathan, H. (1985). Item Response Theory: Principles and Applications. Kluwer-Nijhoff Publishing.

Hambleton, R. & Swaminathan, H. (1985). Item response theory. Boston: Kluwer Nijhoff.

Hanson, B.A. & Brennan, R.L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. Journal of Educational Measurement, 27(4), 345–359.

Hess, K. (2013). A Guide for Using Webb's Depth of Knowledge with Common Core State Standards. © 2013 Common Core Institute.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.

Karantonis, A. & Sireci, S. G. (2006). The Bookmark Standard-Setting Method: A Literature Review. Educational Measurement Issues and Practice, March 2006.

Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (1998). The Bookmark procedure: Methodology and recent implementations. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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 Edition) (pp. 225–253). New York, NY: Routledge

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.

MetaMetrics, Inc. (2014). Linking the NC READY EOG Math/EOC Algebra I/Integrated I with the Quantile® Framework: A study to link the NC READY EOG Math/EOC AlgebraI/Integrated I with the Quantile Framework for Mathematics. Durham, NC: Author.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp. 249–281). Mahwah, NJ: Erlbaum.

Shyyan, V. V., Thurlow, M. L., Larson, E. D., Christensen, L. L., & Lazarus, S. S. (2016). White paper on common accessibility language for states and assessment vendors. Minneapolis, MN: University of Minnesota, Data Informed Accessibility—Making Optimal Needs-based Decisions (DIAMOND).

Thissen, D., Nelson, L., Rosa, K., & McLeod, L. D. (2001). Item response theory for items scored in more than two categories. In D. Thissen & H. Wainer (Eds.), Test scoring (pp. 141–186). Mahwah, NJ: Erlbaum.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds), Test Scoring (pp. 73–140). Mahwah, NJ: Lawrence Erlbaum Associates.

USDE (2002). No Child Left Behind Act of 2001. U.S. Department of Education.

Williamson, G.L., Sanford-Moore, E.E., & Bickel, L. (2016). The Quantile® Framework for Mathematics quantifies the mathematics ability needed for college and career readiness. (MetaMetrics Research Brief). Durham, NC: MetaMetrics, Inc.

Yen, W. M., & Fitzpatrick, A. R. (2006, p112). Item Response Theory. In R. L. Brennan (Ed.), Educational Measurement (4th Ed.). Westport, CT: American Council on Education and Praeger Publishers.

# Appendix 1

# Appendix 1-A

## Session Law 2014-78 Senate Bill 812

**GENERAL ASSEMBLY OF NORTH CAROLINA**
**SESSION 2013**

# SESSION LAW 2014-78
# SENATE BILL 812

AN ACT TO EXERCISE NORTH CAROLINA'S CONSTITUTIONAL AUTHORITY OVER ALL ACADEMIC STANDARDS; TO REPLACE COMMON CORE; AND TO ENSURE THAT STANDARDS ARE ROBUST AND APPROPRIATE AND ENABLE STUDENTS TO SUCCEED ACADEMICALLY AND PROFESSIONALLY.

Whereas, the North Carolina Constitution, Article IX, Section 5, directs the State Board of Education to supervise and administer a free public school system and make all needed rules and regulations in relation thereto, subject to laws enacted by the General Assembly; and

Whereas, the North Carolina General Statutes direct the State Board of Education to adopt and modify academic standards for the public schools; and

Whereas, the North Carolina General Statutes also grant local boards of education broad discretion and authority with respect to specific curricular decisions and academic programs, as long as they align with the standards adopted by the State Board of Education; and

Whereas, North Carolina desires its academic standards to be among the highest in the nation; and

Whereas, the adoption and implementation of demanding, robust academic standards is essential for providing high-quality education to our students and for fostering a competitive economy for the future of our State; and

Whereas, North Carolina's standards must be age-level and developmentally appropriate; Now, therefore,

The General Assembly of North Carolina enacts:

**SECTION 1.(a)** The State Board of Education shall:

(1) Continue to exercise its authority under the North Carolina Constitution and G.S. 115C-12(9c) to adopt academic standards for the public schools.

(2) Conduct a comprehensive review of all English Language Arts and Mathematics standards adopted under G.S. 115C-12(9c) and propose modifications to ensure that those standards meet all of the following criteria:
   a. Increase students' level of academic achievement.
   b. Meet and reflect North Carolina's priorities.
   c. Are age-level and developmentally appropriate.
   d. Are understandable to parents and teachers.
   e. Are among the highest standards in the nation.

(3) Not enter into any agreement, understanding, or contract that would cede control of the Standard Course of Study and related assessments. This requirement does not prohibit the use of national or international curricula, such as the Advanced Placement or International Baccalaureate programs.

(4) Involve and survey a representative sample of parents, teachers, and the public to help determine academic content standards that meet and reflect North Carolina's priorities and the usefulness of the content standards.

(5) Prior to making changes to the standards, consult with the Academic Standards Review Commission, which is established in Section 2 of this act.

**SECTION 1.(b)** Academic standards adopted by the State Board of Education under G.S. 115C-12(9c) shall continue to be named and referred to as the "North Carolina

Standard Course of Study," reflecting emphasis on North Carolina's needs and priorities. The State Board of Education shall maintain and reinforce the independence of the North Carolina Standard Course of Study and related student assessments, rejecting usurpation and intrusion from federally mandated national or standardized controls.

**SECTION 2.(a)** There is established the Academic Standards Review Commission. The Commission shall be located administratively in the Department of Administration but shall exercise all its prescribed powers independently of the Department of Administration.

**SECTION 2.(b)** The Commission shall be composed of 11 members as follows:

(1) Four members appointed by the President Pro Tempore of the Senate. The President Pro Tempore shall consider, but is not limited to, appointing representatives from the following groups in these appointments: parents of students enrolled in the public schools; Mathematics and English Language Arts teachers; Mathematics and English Language Arts curriculum experts; school leadership to include principals and superintendents; members of the business community; and members of the postsecondary education community who are qualified to assure the alignment of standards to career and college readiness.

(2) Four members appointed by the Speaker of the House of Representatives. The Speaker of the House of Representatives shall consider, but is not limited to, appointing representatives from the following groups in these appointments: parents of students enrolled in the public schools; Mathematics and English Language Arts teachers; Mathematics and English Language Arts curriculum experts; school leadership to include principals and superintendents; members of the business community; and members of the postsecondary education community who are qualified to assure the alignment of standards to career and college readiness.

(3) Two members of the State Board of Education as follows: (i) the Chair or the Chair's designee and (ii) a member appointed by the Chair, representing the State Board's Task Force on Summative Assessment.

(4) One member appointed by the Governor.

No individual serving in a statewide elected office or as a member of the General Assembly shall be appointed to the Commission. The Commission shall meet on the call of the Chair of the State Board of Education no later than September 1, 2014. The cochairs of the Commission shall be elected during the first meeting from among the members of the Commission by the members of the Commission.

**SECTION 2.(c)** The Commission shall:

(1) Conduct a comprehensive review of all English Language Arts and Mathematics standards that were adopted by the State Board of Education under G.S. 115C-12(9c) and propose modifications to ensure that those standards meet all of the following criteria:
   a. Increase students' level of academic achievement.
   b. Meet and reflect North Carolina's priorities.
   c. Are age-level and developmentally appropriate.
   d. Are understandable to parents and teachers.
   e. Are among the highest standards in the nation.

(2) As soon as practicable upon convening, and at any time prior to termination, recommend changes and modifications to these academic standards to the State Board of Education.

(3) Recommend to the State Board of Education assessments aligned to proposed changes and modifications that would also reduce the number of high-stakes assessments administered to public schools.

(4) Consider the impact on educators, including the need for professional development, when making any of the recommendations required in this section.

The Commission shall assemble content experts to assist it in evaluating the rigor of academic standards. The Commission shall also involve interested stakeholders in this process and otherwise ensure that the process is transparent.

**SECTION 2.(d)** The Commission shall meet upon the call of the cochairs. A quorum of the Commission shall be nine members. Any vacancy on the Commission shall be filled by the appointing authority. The Commission shall hold its first meeting no later than September 1, 2014.

**SECTION 2.(e)** To the extent that funds are available, the Commission may contract for professional, clerical, and consultant services. Professional and clerical staff positions for the Commission may be filled by persons whose services are loaned to the Commission to fulfill the work of the Commission.

**SECTION 2.(f)** The Department of Administration shall provide meeting rooms, telephones, office space, equipment, and supplies to the Commission and shall be reimbursed from the Commission's budget, to the extent that funds are available.

**SECTION 2.(g)** To the extent that funds are available, the Commission members shall receive per diem, subsistence, and travel allowances in accordance with G.S. 138-5, 138-6, or 120-3.1, as appropriate.

**SECTION 2.(h)** Upon the request of the Commission, all State departments and agencies and local governments and their subdivisions shall furnish the Commission with any information in their possession or available to them.

**SECTION 2.(i)** The Commission shall make a final report of its findings and recommendations to the State Board of Education, the Joint Legislative Education Oversight Committee, and the 2016 Session of the 2015 General Assembly. The Commission shall terminate on December 31, 2015, or upon the filing of its final report, whichever occurs first.

**SECTION 3.(a)** G.S. 115C-174.11(c)(3) is repealed.

**SECTION 3.(b)** The State Board of Education shall continue to develop and update the North Carolina Standard Course of Study in accordance with G.S. 115C-12(9c), including a review of standards in other states and of national assessments aligned with those standards, and shall implement the assessments the State Board deems most aligned to assess student achievement on the North Carolina Standard Course of Study, in accordance with Section 9.2(b) of S.L. 2013-360 and Section 5 of this act.

**SECTION 4.** G.S. 115C-12(39) reads as rewritten:

"(39) Power to Accredit Schools. – Upon the request of a local board of education, the State Board of Education shall evaluate schools in local school administrative units to determine whether the education provided by those schools meets acceptable levels of quality. The State Board shall adopt rigorous and appropriate academic standards for accreditation after consideration of (i) the standards of regional and national accrediting agencies, (ii) ~~the Common Core Standards adopted by the National Governors Association Center for Best Practices and the Council of Chief State School Officers,~~ the academic standards adopted in accordance with subdivision (9c) of this section, and (iii) other information it deems appropriate.

The local school administrative unit shall compensate the State Board for the actual costs of the accreditation process."

**SECTION 5.** The State Board of Education shall report to the Joint Legislative Education Oversight Committee by July 15, 2015, on the acquisition and implementation of a new assessment instrument or instruments to assess student achievement on the academic standards adopted pursuant to G.S. 115C-12(9c). The State Board shall not acquire or implement the assessment instrument or instruments without the enactment of legislation by the General Assembly authorizing the purchase. The assessment instrument or instruments shall be nationally normed, aligned with the North Carolina Standard Course of Study, and field-tested. Examples of appropriate assessment models would include, but not be limited to, the Iowa Test of Basic Skills (ITBS), the Scholastic Aptitude Test (SAT), ACT Aspire, and the National Assessment of Educational Progress (NAEP).

**SECTION 6.** Local boards of education shall continue to provide for the efficient teaching of the course content required by the Standard Course of Study as provided under G.S. 115C-47(12). The current Standard Course of Study remains in effect until official notice is provided to all public school teachers, administrators, and parents or guardians of students enrolled in the public schools of any changes made in the Standard Course of Study by the State Board of Education.

**SECTION 7.** This act becomes effective July 1, 2014.
In the General Assembly read three times and ratified this the 16th day of July, 2014.

s/ Philip E. Berger
President Pro Tempore of the Senate

s/ Thom Tillis
Speaker of the House of Representatives

s/ Pat McCrory
Governor

Approved 12:07 p.m. this 22nd day of July, 2014

## Appendix 1-B

## The North Carolina Academic Standards Review Commission ReportDec2015

https://www.ednc.org/wp-content/uploads/2016/01/NC-Academic-Standard-Review-Commission.pdf

## Appendix 1-C

## EOG Standards Review Revision and Implementation Materials

https://simbli.eboardsolutions.com/Meetings/Attachment.aspx?S=10399&AID=55709&MID=2422

## Appendix 1-D

## EOG Standards Review Revision and Implementation Materials

https://simbli.eboardsolutions.com/Meetings/Attachment.aspx?S=10399&AID=94022&MID=3366

# Appendix 2

## Appendix 2-A

## Math Test Specification Meeting Agendas, Survey Form, and Demographic Information of Participants

# Test Specification Meeting

## DAY 1—Meeting Agenda
## North Carolina Department of Public Instruction/Room 150 North

| | |
|---|---|
| 8:30am | **Registration—Room 150 North**<br>Betty Barbour, Josh Griffin |
| 9:00am | **Welcome and Introductions**<br>Josh Griffin, Hope Lung, Betty Barbour<br>• Internet Access, Restrooms and Café (cash only)<br>• Substitute Teacher Form, Stipend Form, Demographics Form<br>• Testing Code of Ethics and Test Security Agreement<br>• Travel Reimbursement |
| 9:35am | **Summative Assessment Psychometric Overview**<br>Dr. Kinge Mbella, Lead Psychometrician, NCDPI/Test Development |
| 10:20am | **Break** |
| 10:30am | **Overview of Revised Standards for Math Grades 3-5**<br>Kitty Rutherford and Denise Shultz, NCDPI/K-12 Mathematics Curriculum and Instruction |
| 11:45am | **Lunch** (on your own) |
| 12:45pm | **Prioritize Standards—ROUND 1** (Breakout Groups: Grade 3, Grade 4, Grade 5)<br>Josh Griffin, Math Test Measurement Specialist, NCDPI/Test Development<br>• Prioritize Assessable Standards<br>• Recommend Weighting by Domain |
| | **Break** (on your own) |
| 2:00pm | **Prioritize Standards—ROUND 2** (Large Group)<br>Josh Griffin<br>• Prioritize Assessable Standards<br>• Recommend Weighting by Domain |
| 3:00pm | **Recommend Percent by Item Type, Calculator Use—Discussion** (Large Group)<br>Josh Griffin |
| 3:45pm | **Summary of Recommendations and General Considerations**<br>Josh Griffin |
| 4:00 pm | **Meeting Adjourned** (Bring snacks/drinks for Day 2)<br>Josh Griffin |

## DAY 2—Meeting Agenda

| | |
|---|---|
| 9:00am | **Collect Travel Reimbursement Documentation**<br>Betty Barbour |
| 9:15am | **Overview of Cognitive Complexity (Webb's Depth of Knowledge)**<br>Josh Griffin, Math Test Measurement Specialist, NCDPI/Test Development |
| | **Break** (on your own) |
| 11:15am | **Cognitive Complexity—ROUND 1** (Breakout Groups: Grade 3, Grade 4, Grade 5) |

| | |
|---|---|
| | Josh Griffin<br>• Recommend Percent by DOK Level |
| | **Break** (on your own) |
| 12:30pm | **Cognitive Complexity—ROUND 2** (Large Group)<br>Josh Griffin<br>• Recommend Percent by DOK Level |
| 1:00 pm | **Distribution of Certificates and Meeting Adjourned**<br>Josh Griffin |

# Demographic Form
## Test Specifications Meeting

Purpose: The completion of this form is voluntary. We are requesting information from each individual because it will provide a description of this group. This information will be used by the North Carolina Department of Public Instruction for aggregate data analysis only. Thank you for your consideration!

**Information**

(Optional) Print your Name: _____

Gender: Male Female

Ethnicity: _____

**Education**

Highest Degree Earned: B.A/B.S M.A./M.S./M.Ed. Ed.D/Ph.D Other: _____

Approximate Year Highest Degree Received:_____

**Experience**

(Active teachers only) What grade level(s) or course(s) did you teach in 2016–17? _____

_____

National Board Certified (circle one): Yes No

If Yes, list your National Board Certification Fields: _____

_____

North Carolina Teacher Certification Fields:_____

_____

Number of Years Employed in Education:_____

Grade Levels Taught (include your entire teaching career; circle all that apply):

K 1 2 3 4 5 6 7 8 9 10 11 12

Experience Teaching the Following (circle all that apply):

EL Students Students with Disabilities Gifted Students Extended Content Standards

**Employment**

Employment Classification (circle one): Full-Time Part-Time Retired

If Full-Time or Part-Time, what is the title of your position?_____

Are you employed by a charter school (circle one)? Yes No

If YES, what is the name of the charter school? _____

Are you employed by a school district (circle one)? Yes No

If YES, what is the name of the school district? _____

If you work at the school-level, what is the name of the school? _____

Compared to other school districts in North Carolina, which of the following best describes the size of your district (meaning the number of students attending schools in your district)?

Large Medium Small

Compared to other school districts in North Carolina, which of the following best describes the community setting of your district (circle one)?

Urban Suburban Rural

*Table 2.1 Demographic Characteristics of the Test Specification Meeting Participants*

| Category | Sub-Category | NC Math 2 & 3 (N=24) | | Grade 3-5 (N=40) | | Grade 6-8 (N=39) | | NC Math 1 (N=13) | |
|---|---|---|---|---|---|---|---|---|---|
| | | N | % | N | % | N | % | N | % |
| Gender | Female | 20 | 83% | 36 | 90% | 34 | 87% | 11 | 85% |
| | Male | 4 | 17% | 4 | 10% | 5 | 13% | 2 | 15% |
| Ethnicity | Asian | | 0% | 1 | 3% | | 0% | | 0% |
| | Black | 1 | 4% | 5 | 13% | 7 | 18% | 4 | 31% |
| | Native American | | 0% | | 0% | 1 | 3% | | 0% |
| | Hispanic | 1 | 4% | | 0% | | 0% | | 0% |
| | White | 8 | 33% | 32 | 80% | 25 | 64% | 9 | 69% |
| | Mixed | | 0% | 1 | 3% | 1 | 3% | | 0% |
| Highest Degrees Earned | BA/BS | 7 | 29% | 14 | 35% | 15 | 38% | 5 | 38% |
| | J.D./Ed.D/Ph.D | 2 | 8% | 4 | 10% | 1 | 3% | 2 | 15% |
| | MA/MS/M.Ed | 15 | 63% | 22 | 55% | 23 | 59% | 6 | 46% |
| District Size | Large | 10 | 42% | 13 | 33% | 16 | 41% | 6 | 46% |
| | Large/Medium | | | | | | | 1 | 8% |
| | Medium | 5 | 21% | 14 | 35% | 13 | 33% | 2 | 15% |
| | Small | 5 | 21% | 8 | 20% | 5 | 13% | 2 | 15% |
| Urbanicity | Rural | 8 | 33% | 14 | 35% | 8 | 21% | 3 | 23% |
| | Suburban | 5 | 21% | 11 | 28% | 12 | 31% | 5 | 38% |
| | Suburban/Rural | 1 | 4% | 2 | 5% | 2 | 5% | | 0% |
| | Urban | 4 | 17% | 6 | 15% | 5 | 13% | 2 | 15% |
| | Urban/Suburban | | 0% | 1 | 3% | 2 | 5% | 1 | 8% |
| | Urban/Suburban/Rural | | 0% | 1 | 3% | 1 | 3% | | 0% |

*Some participants did not declare some of the demographic characteristics

## Appendix 2-B

## General Definition of Mathematics DOK Level

**Mathematics Depth-of-Knowledge Levels**

Level 1 (Recall) includes the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. That is, in mathematics a one-step, well-defined, and straight algorithmic procedure should be included at this lowest level. Other key words that signify a Level 1 include "identify," "recall," "recognize," "use," and "measure." Verbs such as "describe" and "explain" could be classified at different levels depending on what is to be described and explained.

***Level 2 (Skill/Concept)***

Level 2 includes the engagement of some mental processing beyond a habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity, whereas Level 1 requires students to demonstrate a rote response, perform a well-known algorithm, follow a set procedure (like a recipe), or perform a clearly defined series of steps. Keywords that generally distinguish a Level 2 item include "classify," "organize," "estimate," "make observations," "collect and display data," and "compare data." These actions imply more than one step. For example, to compare data requires first identifying characteristics of the objects or phenomenon and then grouping or ordering the objects. Some action verbs, such as "explain," "describe," or "interpret" could be classified at different levels depending on the object of the action. For example, if an item required students to explain how light affects mass by indicating there is a relationship between light and heat, this is considered a Level 2. Interpreting information from a simple graph, requiring reading information from the graph, also is a Level 2. Interpreting information from a complex graph that requires some decisions on what features of the graph need to be considered and how information from the graph can be aggregated is a Level 3. Caution is warranted in interpreting Level 2 as only skills because some reviewers will interpret skills very narrowly, as primarily numerical skills, and such interpretation excludes from this level other skills such as visualization skills and probability skills, which may be more complex simply because they are less common. Other Level 2 activities include explaining the purpose and use of experimental procedures; carrying out experimental procedures; making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts.

***Level 3 (Strategic Thinking)***

Level 3 requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is a Level 3. Activities that require students to make conjectures are also at this level. The cognitive demands at Level 3 are complex and abstract. The complexity does not result from the fact that there are multiple answers, a possibility for both Levels 1 and 2, but because the task requires more demanding reasoning. An activity, however, that has more than one possible answer and requires students to justify the response they give would most likely be a Level 3. Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and using concepts to solve problems.

***Level 4 (Extended Thinking)***

Level 4 requires complex reasoning, planning, developing, and thinking most likely over an extended period of time. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as a Level 2. However, if the student is to conduct a river study that requires taking into consideration a number of variables, this would be a Level 4. At Level 4, the cognitive demands of the task should be high and the work should be very complex. Students should be required to make several connections—relate ideas *within* the content area or *among* content areas—and have to select one approach among many alternatives on how the situation should be solved, in order to be at this highest level. Level 4 activities include designing and conducting experiments; making connections between a finding and related concepts and phenomena; combining and synthesizing ideas into new concepts; and critiquing experimental designs.

## Appendix 2-C

## A Guide for Using Webb's DOK

http://aimc.alpineschools.org/wp-content/uploads/sites/6/2014/06/Webs-Depth-of-Knowledge-Flip-Chart1.pdf

## Appendix 2-D

**North Carolina Testing Program Test Development Process**
https://files.nc.gov/dpi/documents/accountability/testing/testdvprcss17.pdf

# Appendix 4

# Appendix 4-A

# Test Information Functions and Conditional Standard Error of Measurement

*Figure 4.1 TIFs and CSEMs Based on Field Test Item Parameters, Grade 3*

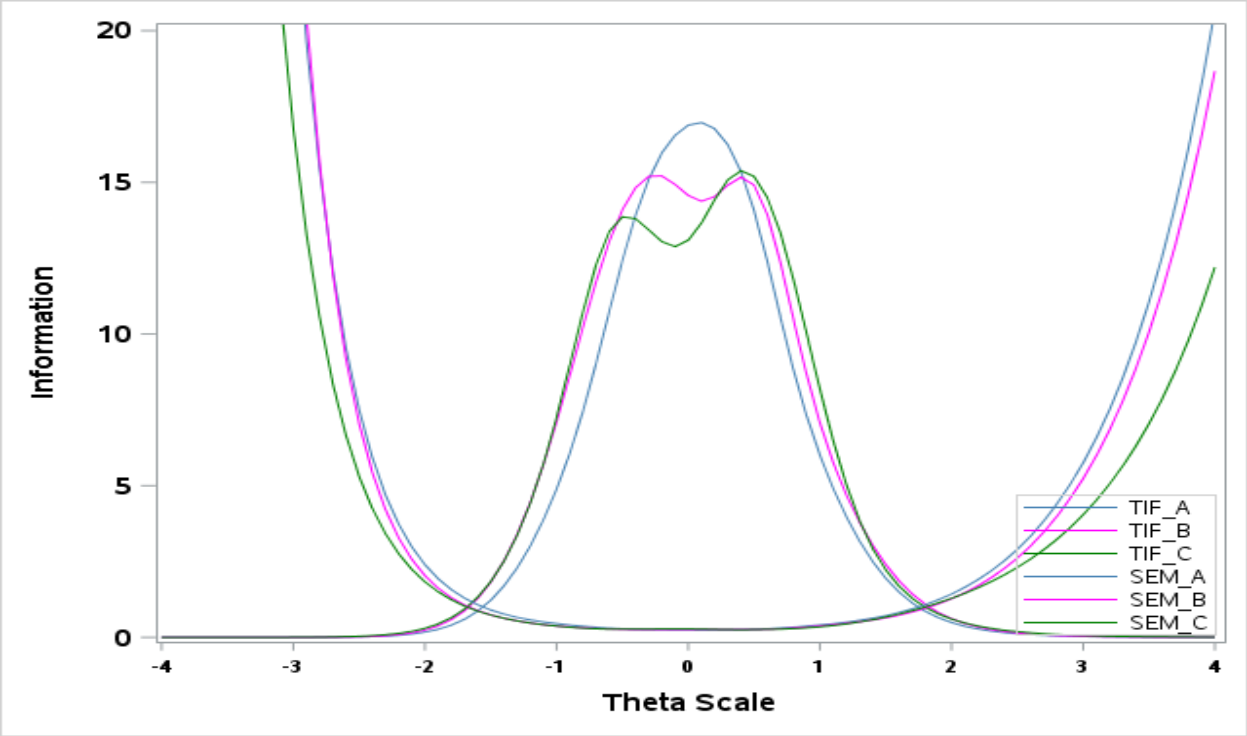



*Figure 4.2 TIFs and CSEMs Based on Field Test Item Parameters, Grade 4*

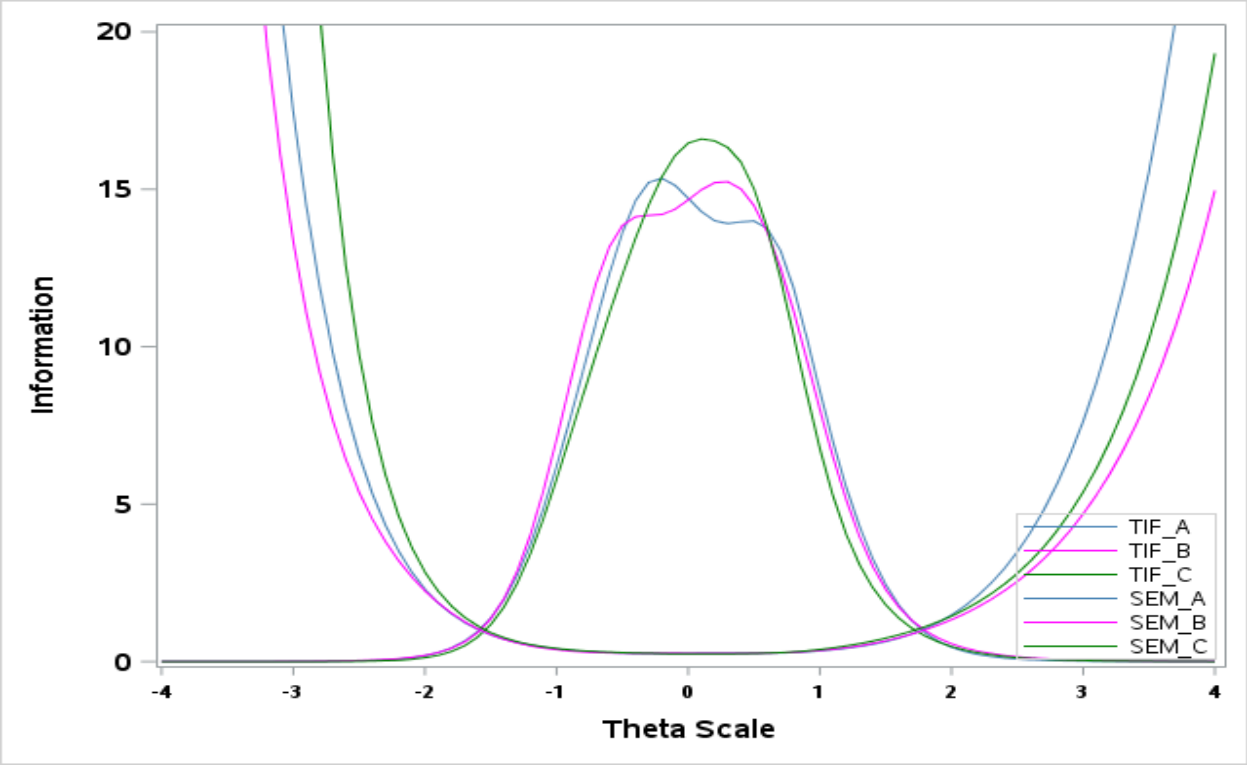


*Figure 4.3 TIFs and CSEMs Based on Field Test Item Parameters, Grade 5*

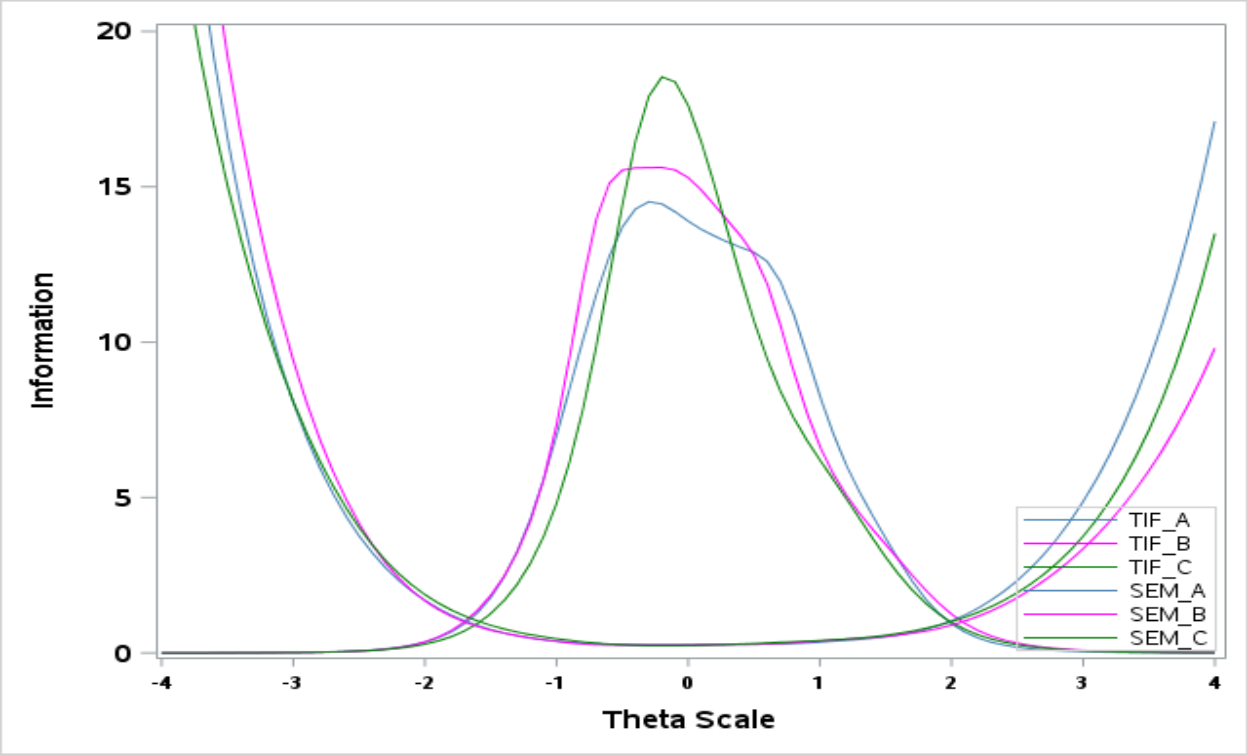



*Figure 4.4 TIFs and CSEMs Based on Field Test Item Parameters, Grade 6*

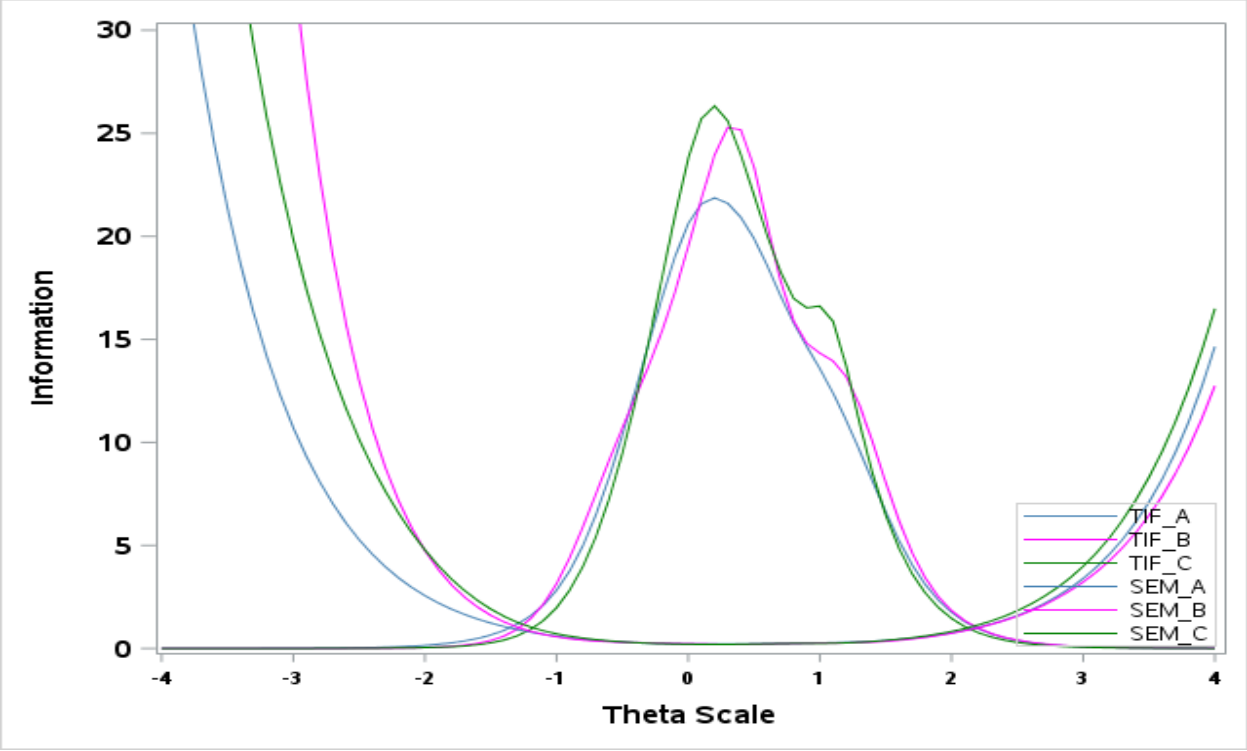


*Figure 4.5 TIFs and CSEMs Based on Field Test Item Parameters, Grade 7*

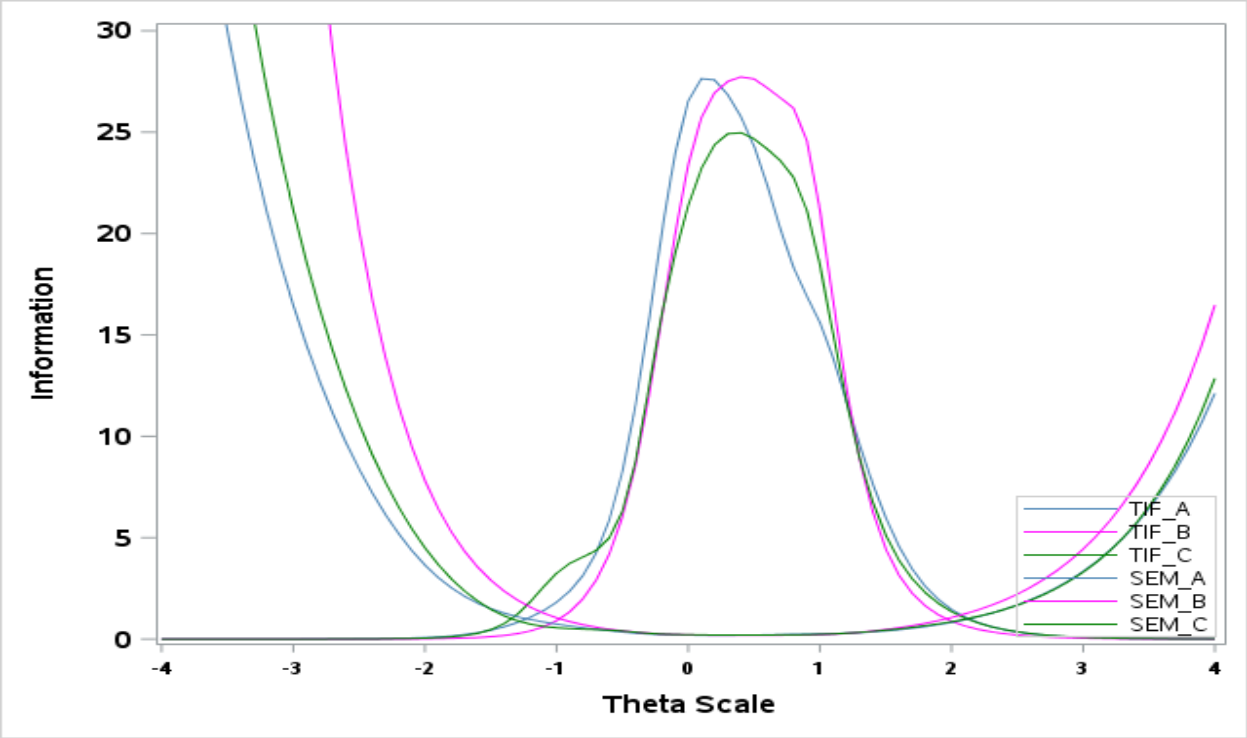



*Figure 4.6 TIFs and CSEMs Based on Field Test Item Parameters, Grade 8*

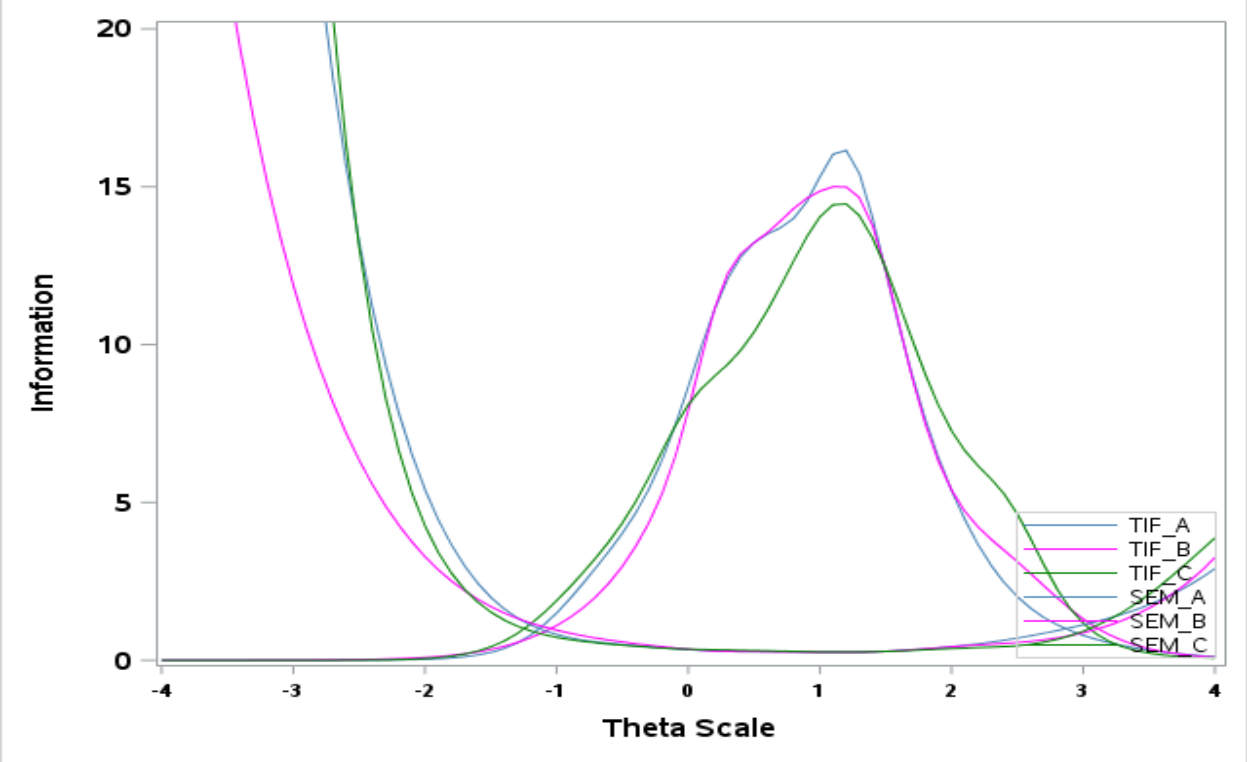


*Figure 4.7 TIFs and CSEMs Based on Field Test Item Parameters, NC Math 1*

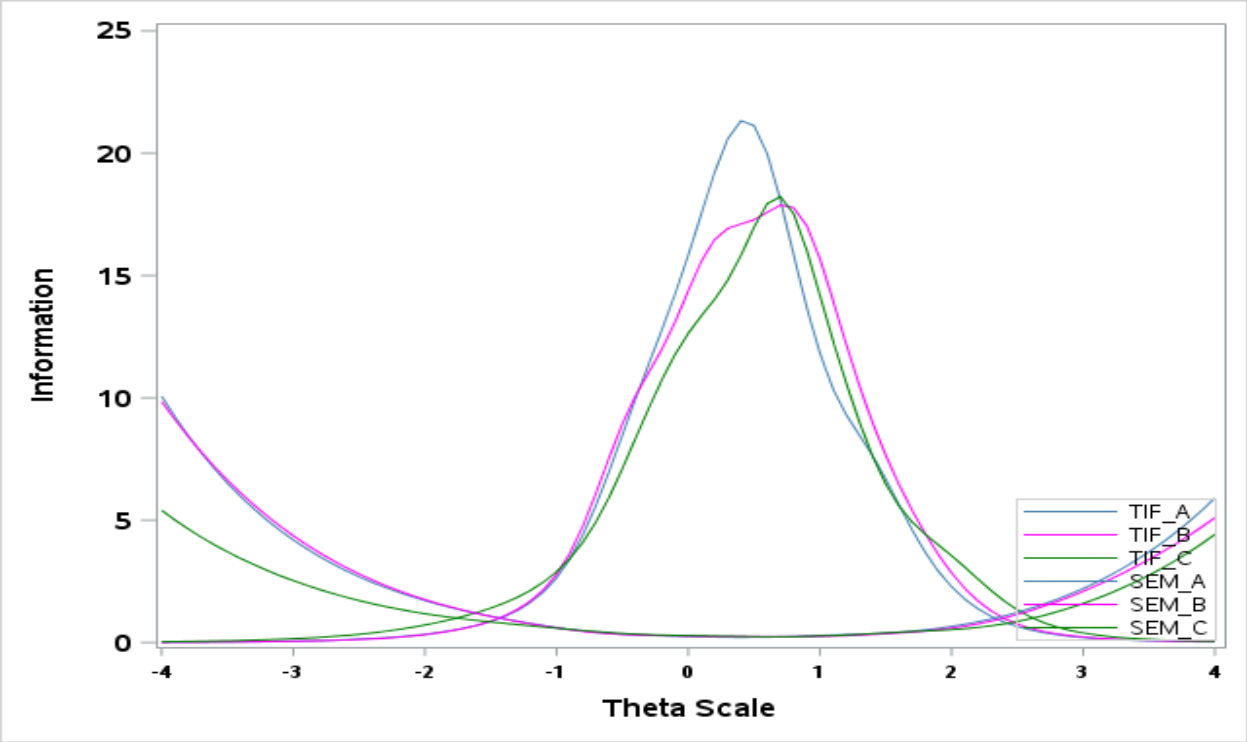
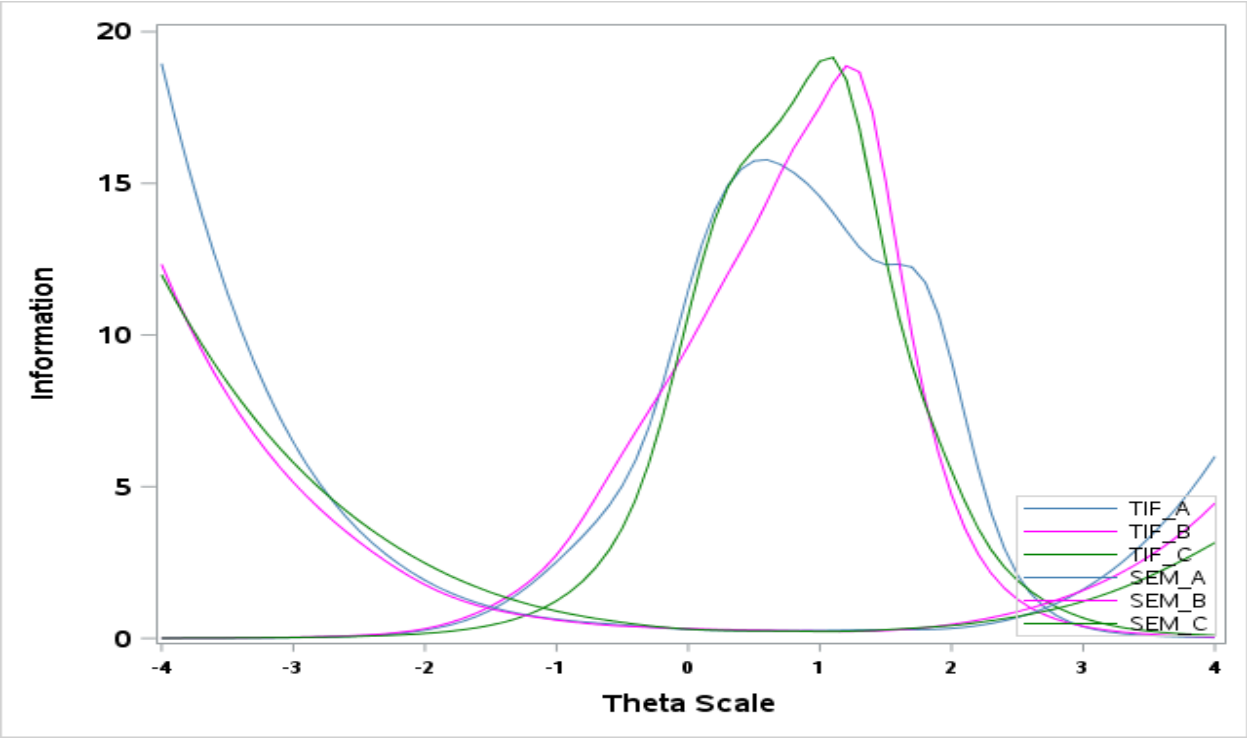



*Figure 4.8 TIFs and CSEMs Based on Field Test Item Parameters, NC Math 3*



# Appendix 4-B

# Fairness and DIF Review Process

# Item Writing and Review for Bias and Sensitivity and Differential Item Functioning (DIF)

## Including processes for EC, ESL, VI reviews

## Defined

Item creation for the North Carolina Testing Program has an established history of inclusion of consideration for bias and sensitivity, and this has been considered as an integrated part of the development process prior to field testing. Vetting steps that specifically involve the EC/ESL/VI Specialists look for content that may present a bias or insensitivity issue such as contexts that might elicit an emotional response and inhibit students' ability to respond and contexts that students may be unfamiliar with for cultural or socio-economic reasons.

## Participant Requirements

Teachers in North Carolina are the principal target population, but participants can be augmented with retired teachers and or those holding undergraduate degrees in the content area. The number of item writers and reviewers required during any item development period is determined by the need and the time allotted. All item writers and reviewers must be trained for bias and sensitivity.

## Training Requirements

Item writers and reviewers must be trained on the standards and content being measured. All item writers and reviewers are subjected to extensive training on proper item design and they are also trained to consider bias and sensitivity of item content. Additionally, since the vetting process includes specific steps for EC, ESL, and VI check, training is required for these reviewers. Depending on the event and the experience of the group that is being asked to write and review, training may be best applied in a face-to-face session. However, the majority of training is designed to be delivered in self-directed online training modules.

## Process and Timeline

Item writing can begin any time a change in standards has been initiated for any content that is required to be measured with a standardized test administration. See the flowcharts in the appendices for the process of writing and review that items must go through in order to be considered candidates for inclusion on either stand-alone field tests or as embedded experimental items on operational tests. Quantities and type of items per targeted standard and the time frame set by leadership of when operational tests are to exist helps determine the timeline for when items must be ready and how many item writers and reviewers are needed.

# DIF Review

## Defined

Per step 14 in the official SBE approved Test Development Process Flow Chart (http://www.ncpublicschools.org/docs/accountability/latestflowchart.pdf) bias reviews occur after items have been field tested and have data that supports further inspection of the items for bias or insensitivity. This is processed in steps within the online test development system (TDS) that are titled DIF Review.

The methodology used for the North Carolina Testing Program to identify items that show differential item functioning (DIF, sometimes called "statistical bias", is a concept that is different from the non-technical notion of "bias") is the Mantel-Haensel Delta-DIF method.

## Calculating Statistical Bias using Mantel-Haensel Delta-DIF Method

Since the method depends on sample size, there is no single number or range of numbers that identifies an item as having moderate or more significant levels of DIF. Rather, the statistical methodology takes the sample size into account and determines whether an item should be rated as A, B, or C, according to whether it displays no significant DIF (A level), significant but still low level of DIF (B level), or more pronounced DIF (C level). A minimum number of 300 per subgroup is necessary in order to produce DIF values that are stable and do not exaggerate the counts of DIF in the B and C levels.

The current operational strategy is to reduce or eliminate the need for DIF Review by choosing not to use any item that has any significant degree of differential item functioning (C level DIF). In the rare case where an item is needed to fill test form design parameters and no A level DIF item exists, then an item in B (first choice) or C (last resort) DIF is put through an additional bias review process that content specialists coordinate.

The current subgroup analyses conducted are: Male/Female, White/Black, White/Hispanic, Urban/Rural, EDS/non-EDS.

This is the same system that the National Assessment of Educational Progress uses. For each analysis of DIF, there is a focal group and a reference group. For example in the male-female analysis, the focal group is females and the reference group is males. A plus (+) or minus (-) sign is used to indicate the direction of DIF. For example, if an item has a B- rating for the male-female analysis that means that the item slightly disfavors (minus sign) females (or slightly favors males). There may be many reasons for a B rating, and such a rating is by no means regarded as a reason to forbid the item to be on a test.

Below are some relevant links that describe the DIF methodology and related topics. The last link shows that NAEP sometimes does use items that have been flagged as having certain levels of DIF (click the individual links for the tests in the various NAEP content areas), provided that those items receive approval following the bias panel review and the subsequent content review. Ultimately, in NAEP's process, the final decision of whether to use an item is made by human beings based on all available info. It is not an automated decision produced purely by computer analyses.

- https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_checks_dif_proced.aspx
- https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_checks_dif_categ.aspx

- https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_avoidviolat_results.aspx

## Participant Requirements

DIF Review participants collectively must model the dimensions that are subject to the DIF parameters which match the Bias Review Panel participants. Since the volume of items that typically get flagged for non-A level values in the analysis that need to go through DIF Review is very small, the number of participants can likewise be a minimum set of five or six.

## Training Requirements

DIF Review participants are required to go through the same training provided to the item writers and reviews and the Bias Review panel participants.

## Review Process and Timeline

Tests are administered both fall and spring and the DIF analyses is done after the spring administration on combined data (fall and spring).

February through May:

- DIF reviews of DIF flagged items from the Fall

June through September:

- DIF reviews of DIF flagged items from the Spring

October through February:

- Spring base forms are assembled and embedded items are placed

# DIF Review Questions

1. Does the item contain language that is not commonly used statewide or has different connotations in different parts of the state or in different cultural or gender groups?
   No
   Yes - Explain

2. Does the item contain any local references that are not a part of the statewide curriculum?
   No
   Yes - Explain

3. Does the item portray anyone in a stereotypical manner? (These could include activities, occupations, or emotions.)
   No
   Yes - Explain

4. Does the item contain any demeaning or offensive materials?
   No
   Yes - Explain

5. Does the item have offensive, stereotyping, derogatory, or proselytizing religious references?
   No
   Yes - Explain

6. Does the item assume that all students come from the same socioeconomic background?
(e.g., a suburban home with two-car garage)
   No
   Yes - Explain

7. Does the artwork adequately reflect the diversity of the student population?
   Yes
   N/A
   No - Explain

8. Is there any source of bias detected in this item?
   No
   Yes - Explain

Additional Comments:

## Sample Bias and Sensitivity Training Materials

# Instructions for Review

### What is the purpose of this review?

After items are field tested, statistics are gathered on each item based on examinees' responses. Sometimes, the statistics indicate the possibility of Construct-Irrelevant Variance – "noise" in the item that prevents us from knowing something about the student's abilities and is measuring something else instead. Your part in this review is to judge whether the content of the item is in fact measuring something about the student other than his or her ability or knowledge in the content area that the question was intended to measure.

### How were these items identified for review?

Through a statistical technique called "Differential Item Functioning" (DIF). After controlling for students' ability, are there differences in performance on the item between groups? If an item behaves differently statistically for one group of examinees than it does for another group of examinees, it is flagged for review.

The content of the items was not considered during the statistical analysis. So, these items were flagged for review because we need to determine if there is anything about these items that may be a source of bias.

### What is bias?

TRUE Bias is when

- An item measures membership in a group more than it measures a content objective.
- An item contains information or ideas that are unique to the culture of one group AND this information or idea is not part of the course of study (North Carolina Essential Standards or North Carolina Common Core Standards).
- The item cannot be answered by a person who does not possess some certain background knowledge.

Sensitivity is another issue that could occur in an item. Sensitivity issues occur when

- An item contains information or ideas that some people will find objectionable or raise strong emotions AND this information or idea is not part of the course of study.
- Assumptions are made within the item that all examinees come from the same background.

Bias is NOT

- Just having a boy's name or a girl's name in the item
- Just mentioning a part of the state, country, or world
- Just mentioning an activity that is variably familiar to certain groups (e.g., vacations, using a bank)
- Just mentioning a "boy" activity (e.g., sports) or a "girl" activity (e.g., cooking) Think about: Jackee Joyner-Kersee or Babe Zaharias; Emeril or The Cajun Chef

**DIF versus Bias**

There is, then, a distinction between DIF and bias. DIF is a statistical technique whereas bias is a qualitative judgment. It is important to know the extent to which an item on a test performs differently for different students. DIF analyses examine the relationship between the score on an item and group membership, while controlling for ability, to determine if an item may be behaving differently for a particular group. While the presence or absence of true bias is a qualitative decision, based on the content of the item and the curriculum context within which it appears, DIF can be used to quantitatively identify items that should be subjected to further scrutiny.

# Guidelines for Bias Review

All groups of society should be portrayed accurately and fairly without reference to stereotypes or traditional roles regarding gender, age, race, ethnicity, religion, physical ability, or geographic setting. Presentations of cultural or ethnic differences should neither explicitly nor implicitly rely on stereotypes nor make moral judgments. All group members should be portrayed as exhibiting a full range of emotions, occupations, activities, and roles across the range of community settings and socioeconomic classes. No one group should be characterized by any particular attribute or demographic characteristic.

The characterization of any group should not be at the expense of that group. Jargon, slang, and demeaning characterizations should not be used, and reference to ethnicity, marital status, or gender should only be made when it is relevant to the context. For example, gender neutral terms should be used whenever possible.

In writing items, an item-writer, in an attempt to make an item more interesting, may introduce some local example about which only local people have knowledge. This may (or may not) give an edge to local people and introduce an element of bias into the test. This does not mean, however, that no local references should be made if such local references are a part of the curriculum (in North Carolina history, for example). The test of bias is this: Is this reference to a cultural activity or geographic location something that is taught as part of the curriculum? If not, it should be examined carefully for potential bias.

**Name of Reviewer: _____ Date: _____**

## When reviewing testing materials for bias, consider the following:

1. Does the item contain language that is not commonly used statewide or has different connotations in different parts of the state or in different cultural or gender groups?
2. Does the item contain any local references that are not a part of the statewide curriculum?
3. Does the item portray anyone in a stereotypical manner? (These could include activities, occupations, or emotions.)
4. Does the item contain any demeaning or offensive materials?
5. Does the item have offensive, stereotyping, derogatory, or proselytizing religious references?
6. Does the item assume that all students come from the same socioeconomic background? (e.g., a suburban home with two-car garage)
7. Does the artwork adequately reflect the diversity of the student population?
8. Other comments
9. No source of bias detected in the item

# Appendix 6

# Appendix 6-A

## Test Character Curves, Information Functions and Conditional Standard Error of Measurement

### Test Characteristic Curves (TCCs)

Figure 6.1 Grade 3 TCCs Math 2018-19 Operational Forms

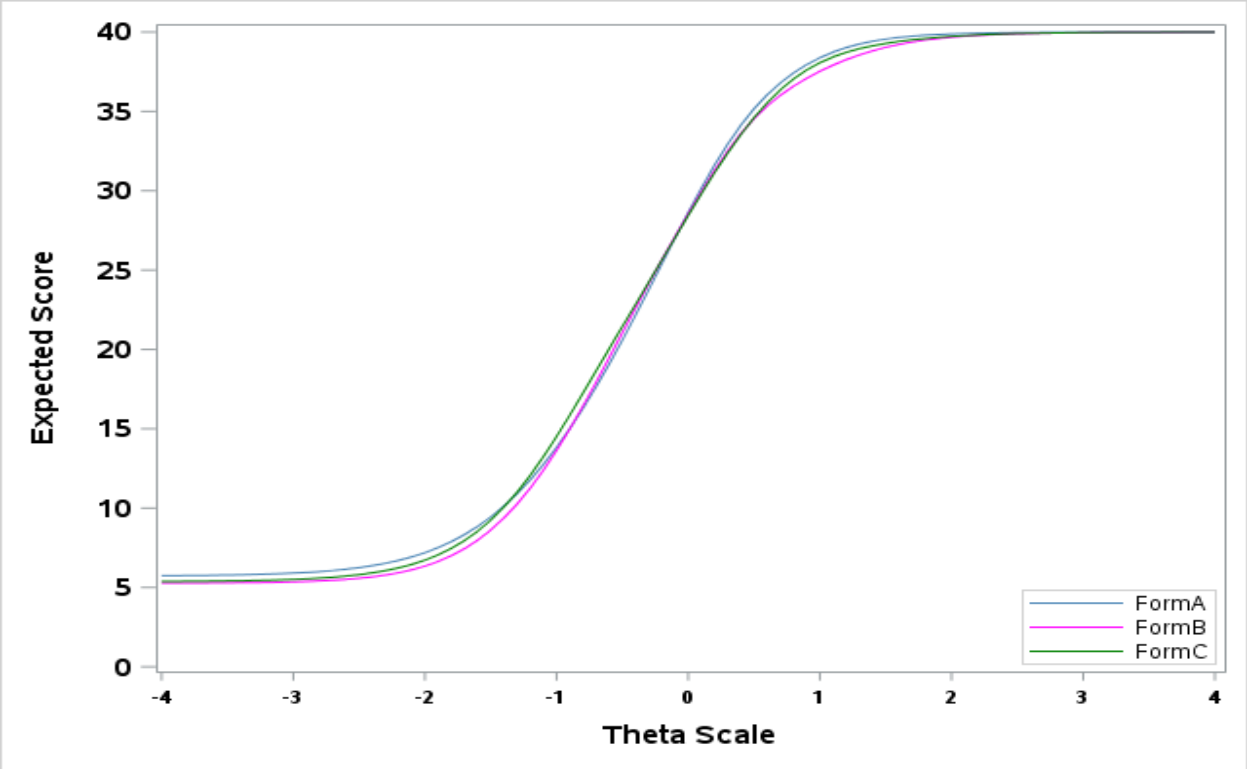



Figure 6.2 Grade 4 TCCs Math 2018-19 Operational Forms

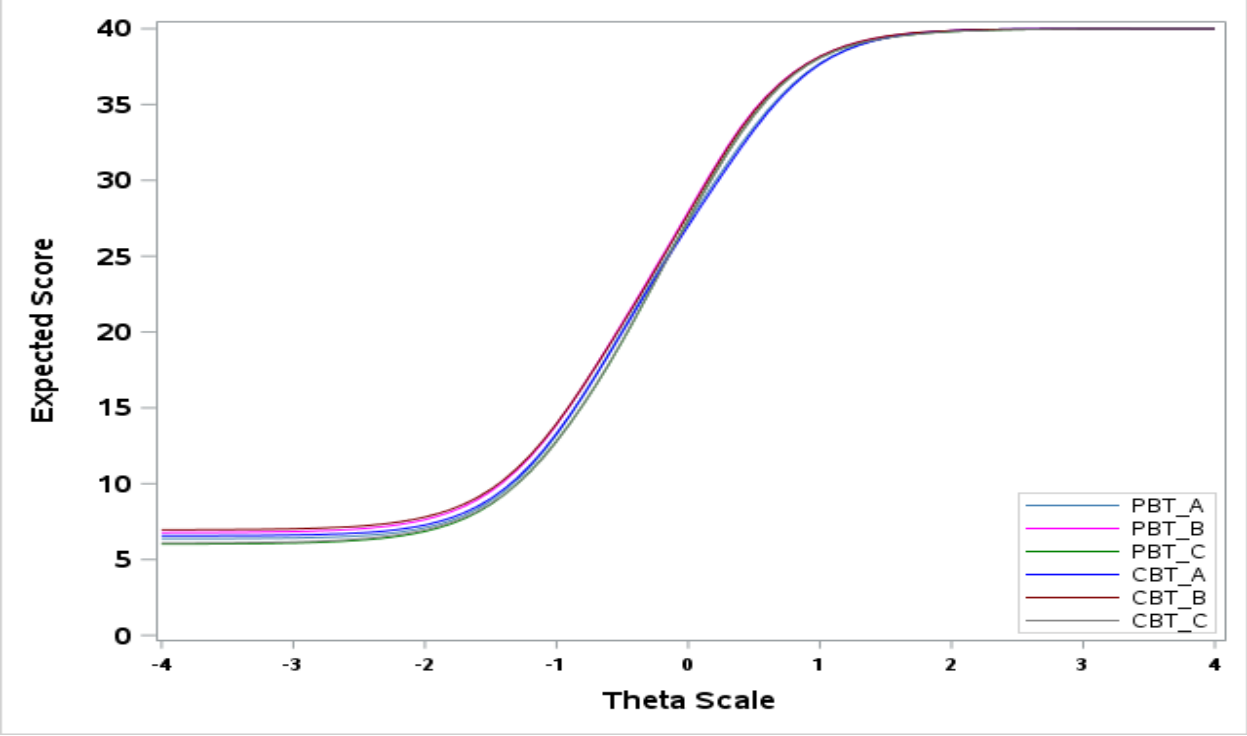


Figure 6.3 Grade 5 TCCs Math 2018-19 Operational Forms

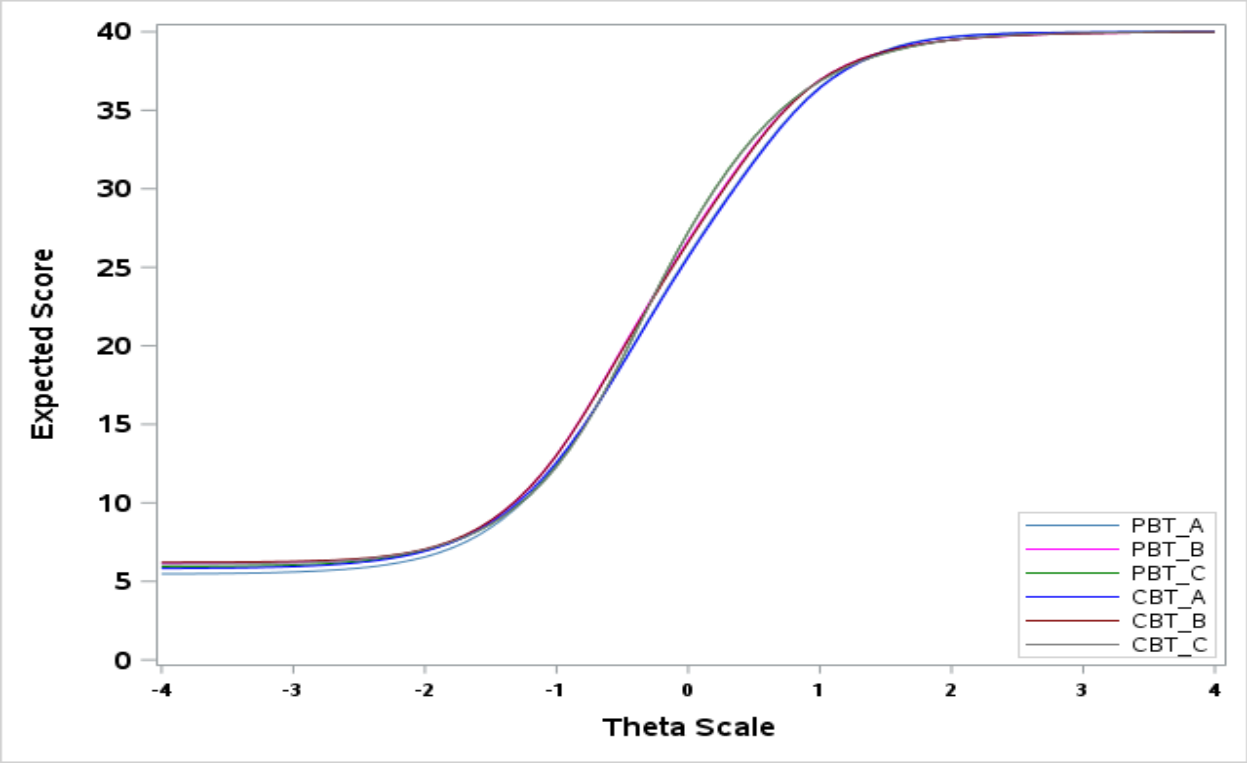



Figure 6.4 Grade 6 TCCs Math 2018-19 Operational Forms

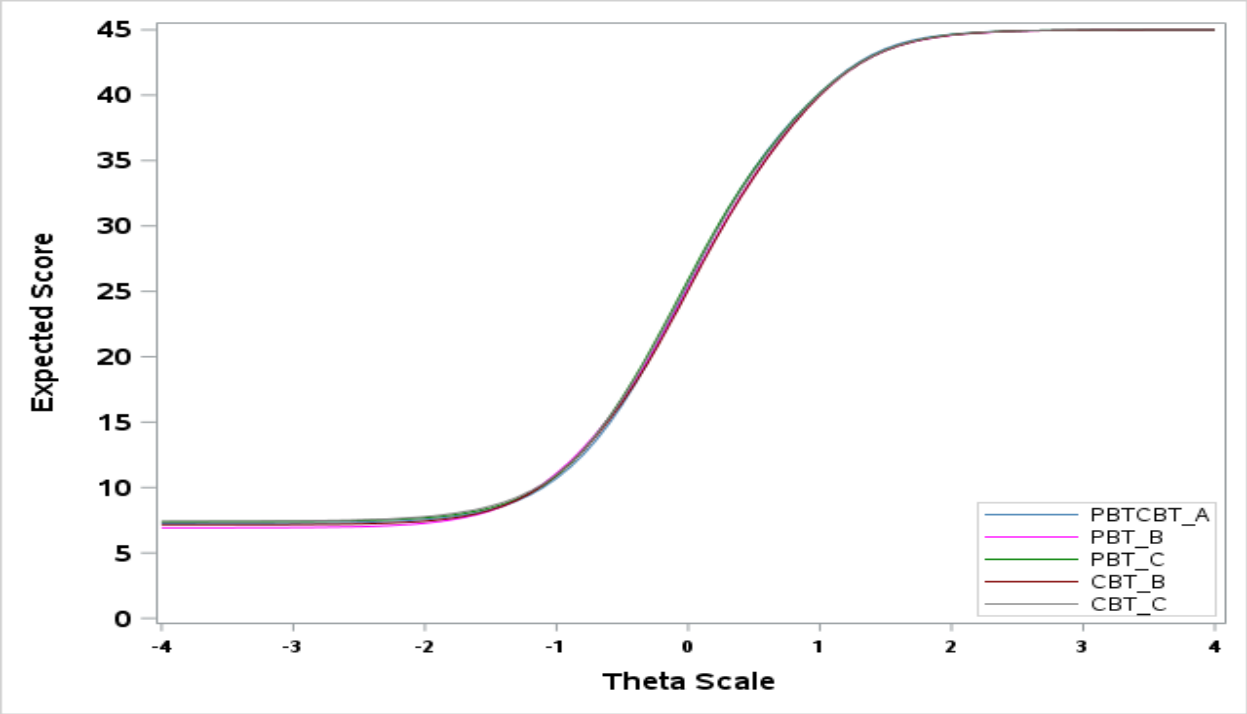


Figure 6.5 Grade 7 TCCs Math 2018-19 Operational Forms

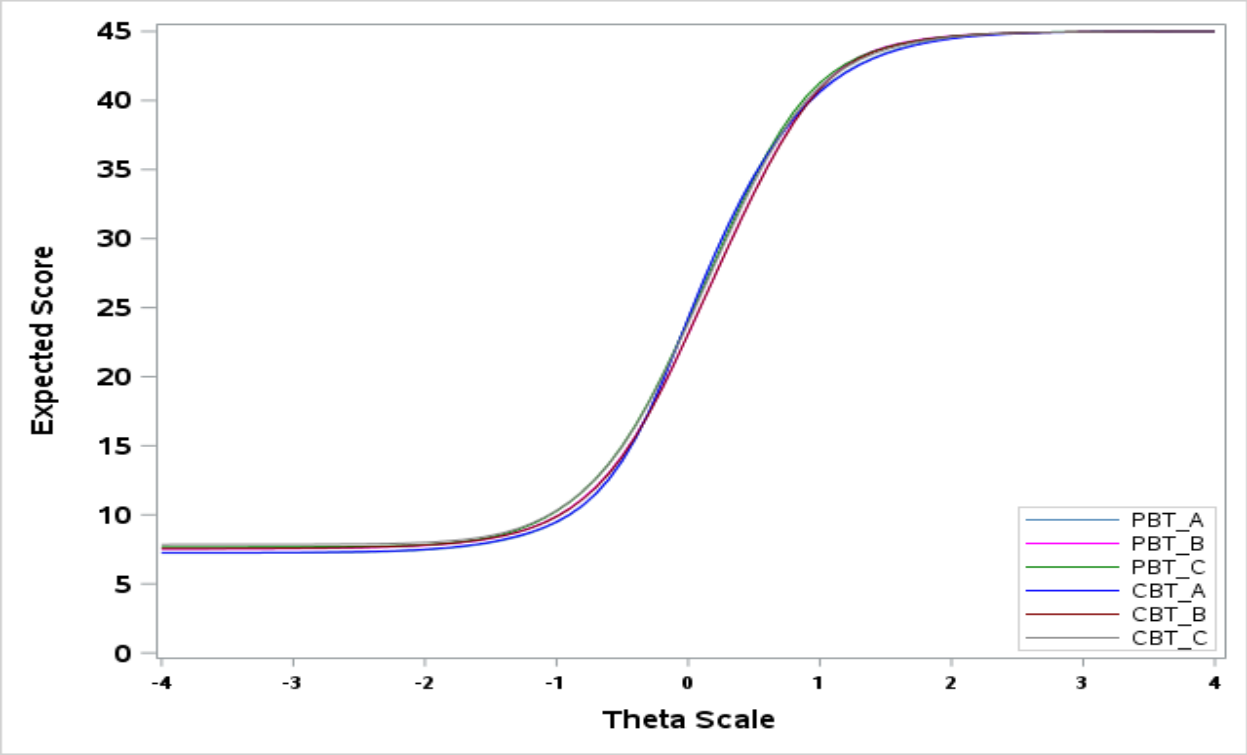



Figure 6.6 Grade 8 TCCs Math 2018-19 Operational Forms

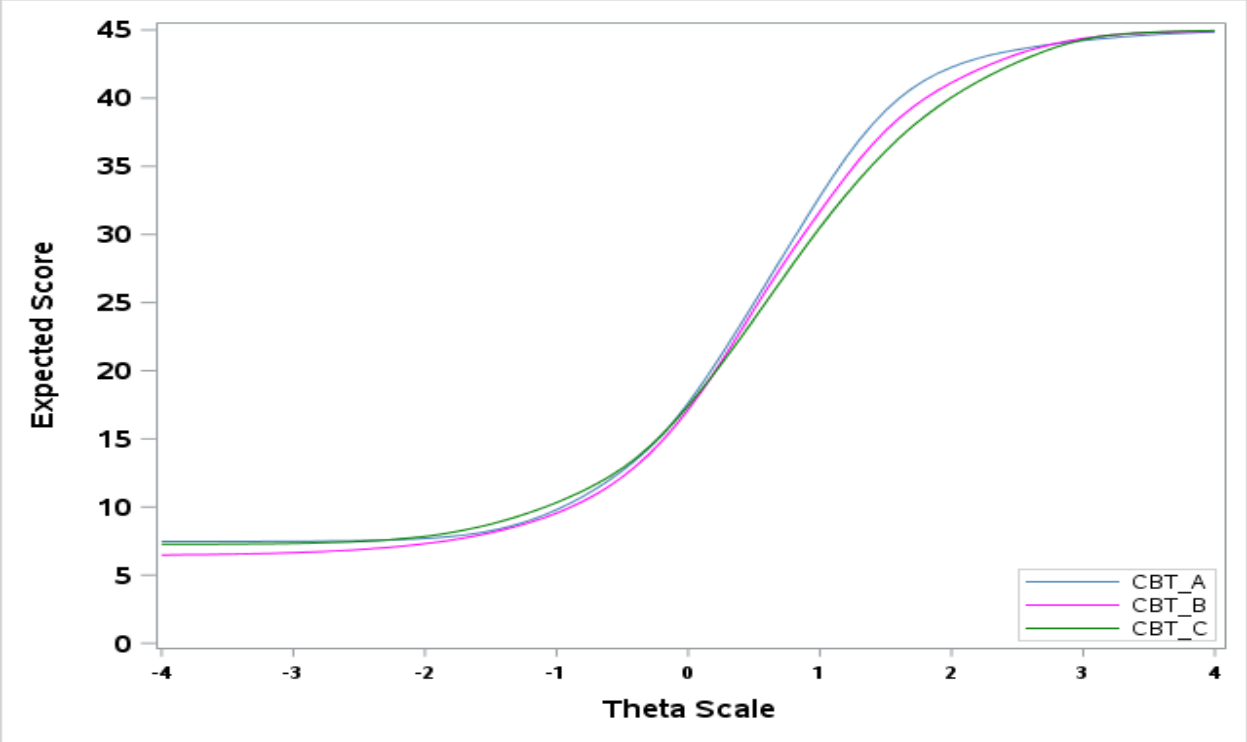


Figure 6.7 NC Math 1 TCCs 2018-19 Operational Forms

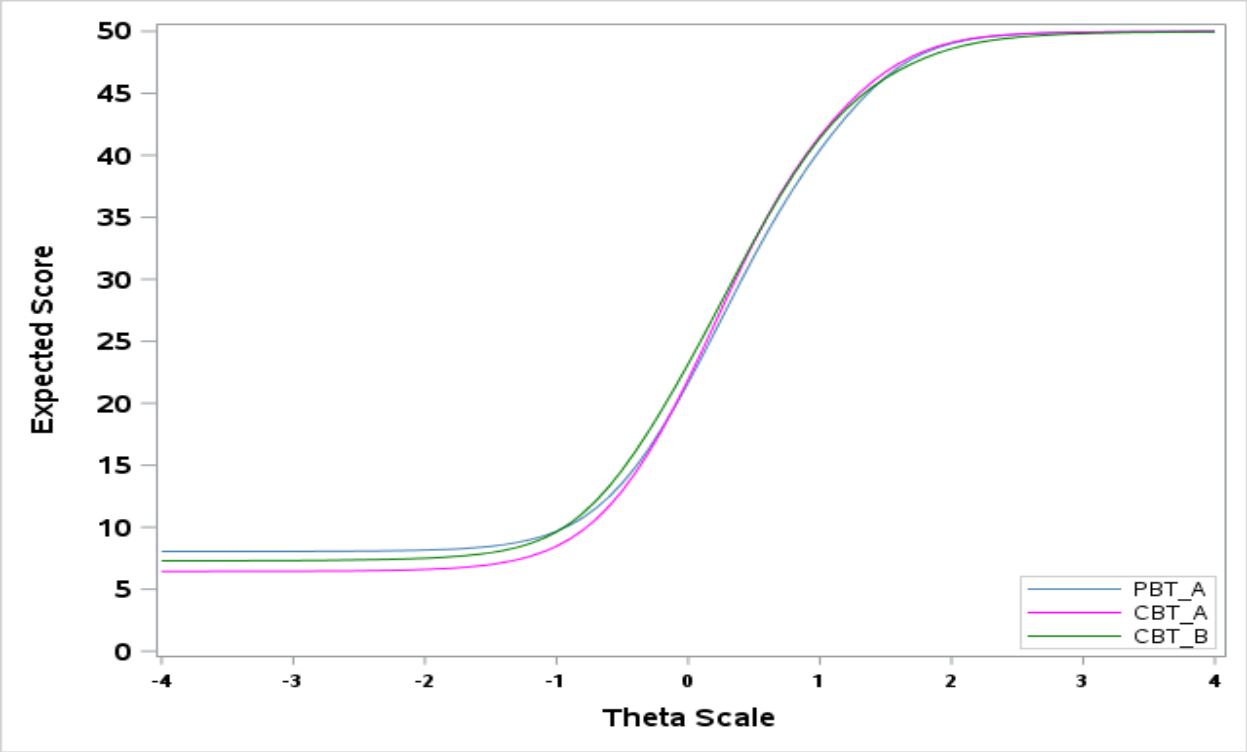
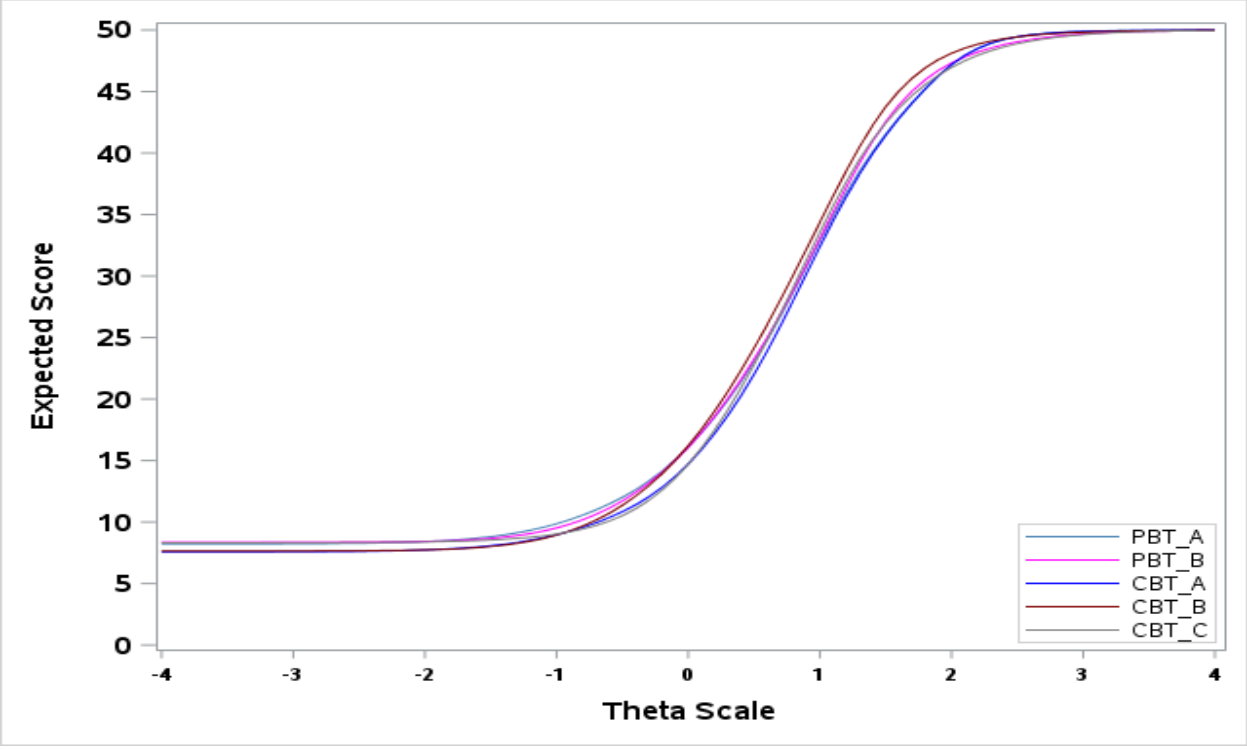



Figure 6.8 NC Math 3 TCCs 2018-19 Operational Forms



**Test Information Functions (TIF) and Conditional Standard Error of Measurements (CSEM)**

Figure 6.9 Math Grade 3 TIFs and CSEM 2018-19 Operational Forms

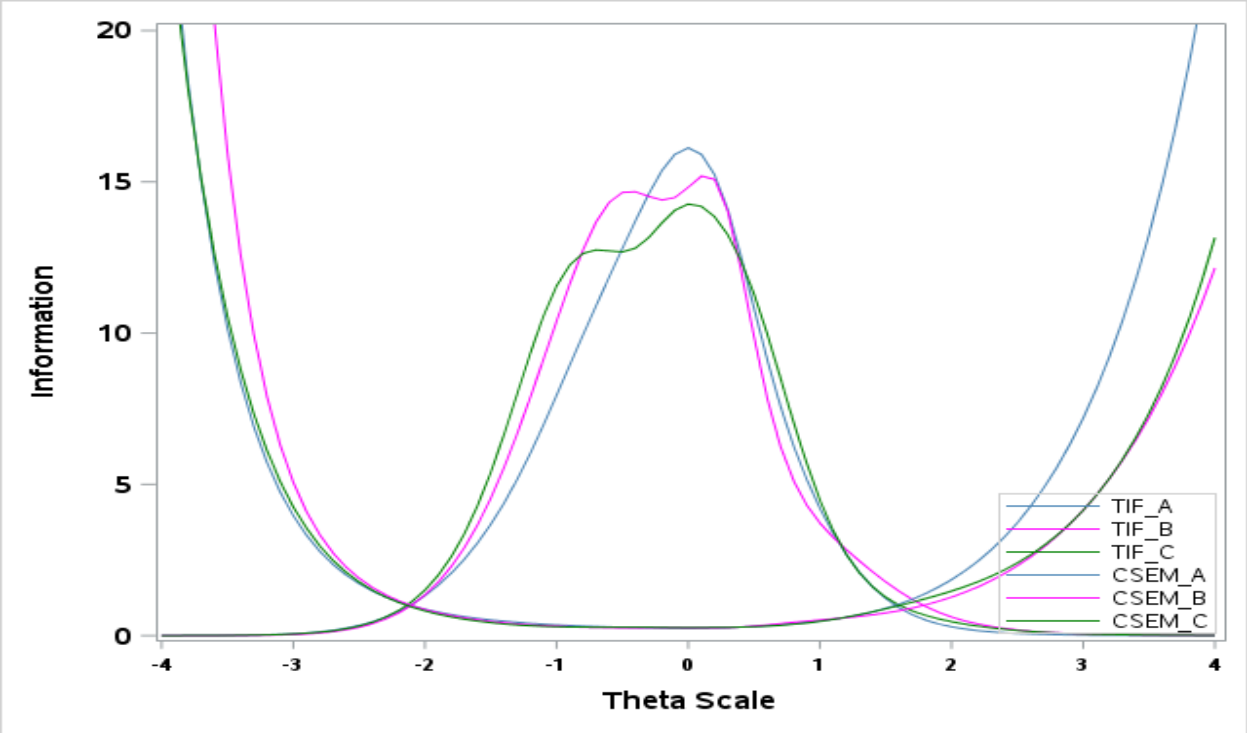



Figure 6.10 Math Grade 4 TIFs and CSEM 2018-19 Operational Forms

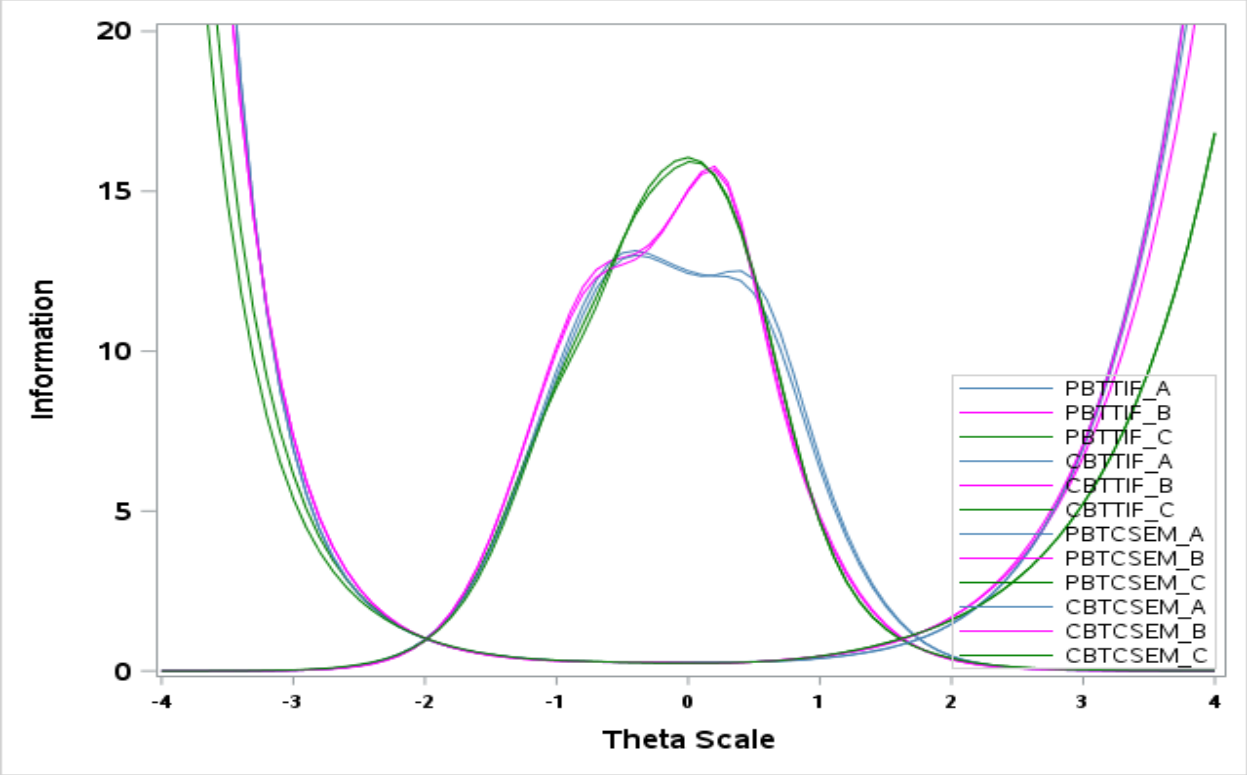


Figure 6.11 Math Grade 5 TIFs and CSEM 2018-19 Operational Forms

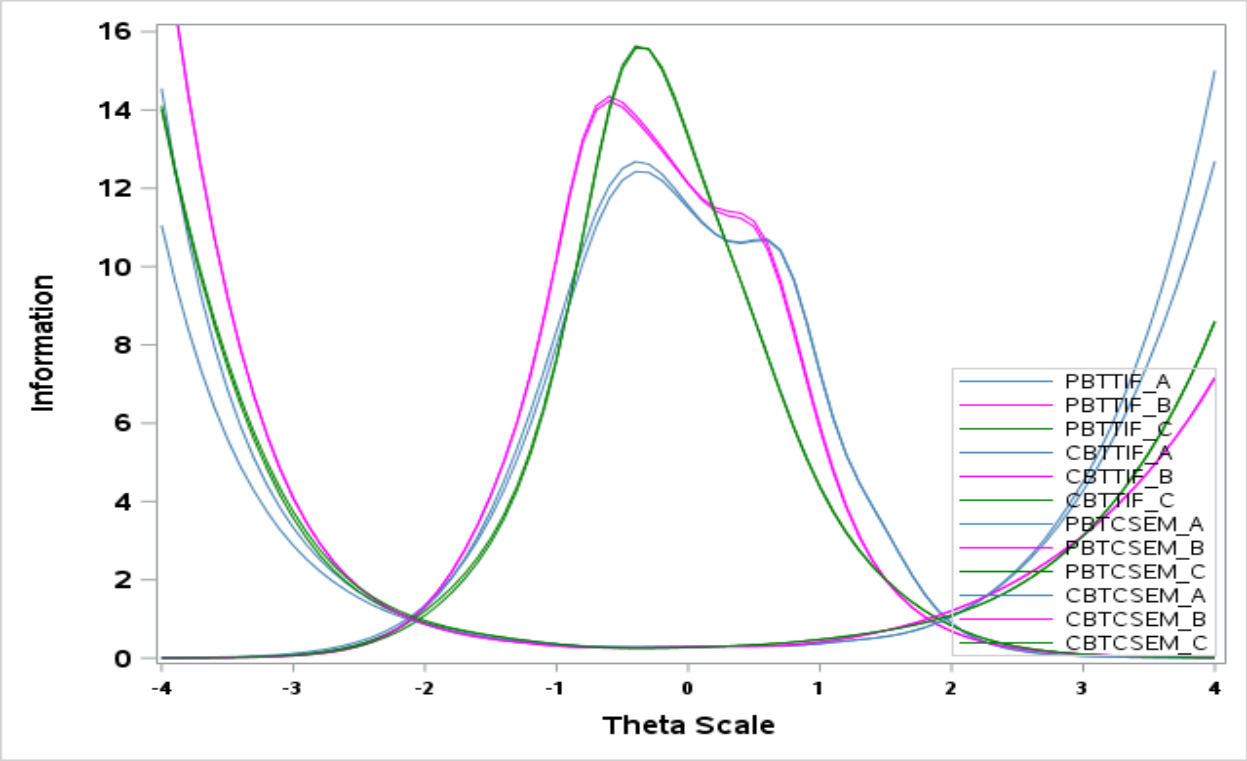



Figure 6.12 Math Grade 6 TIFs and CSEM 2018-19 Operational Forms

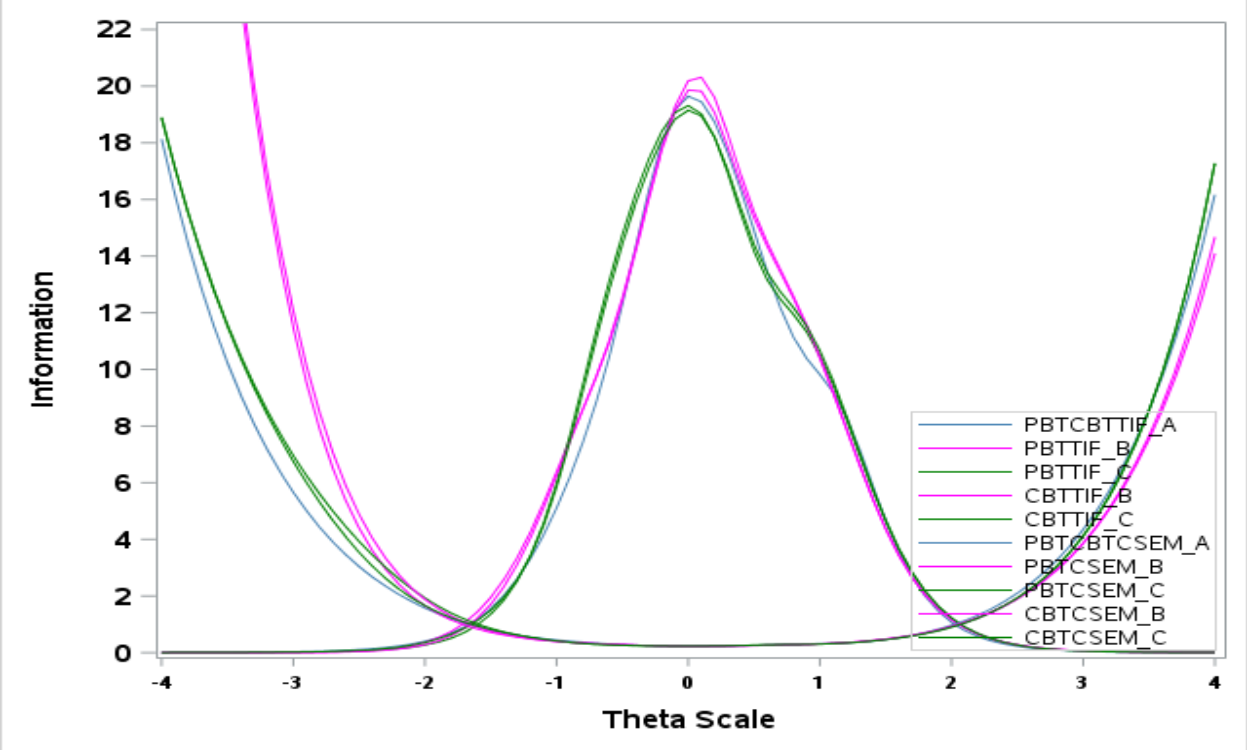


Figure 6.13 Math Grade 7 TIFs and CSEM 2018-19 Operational Forms

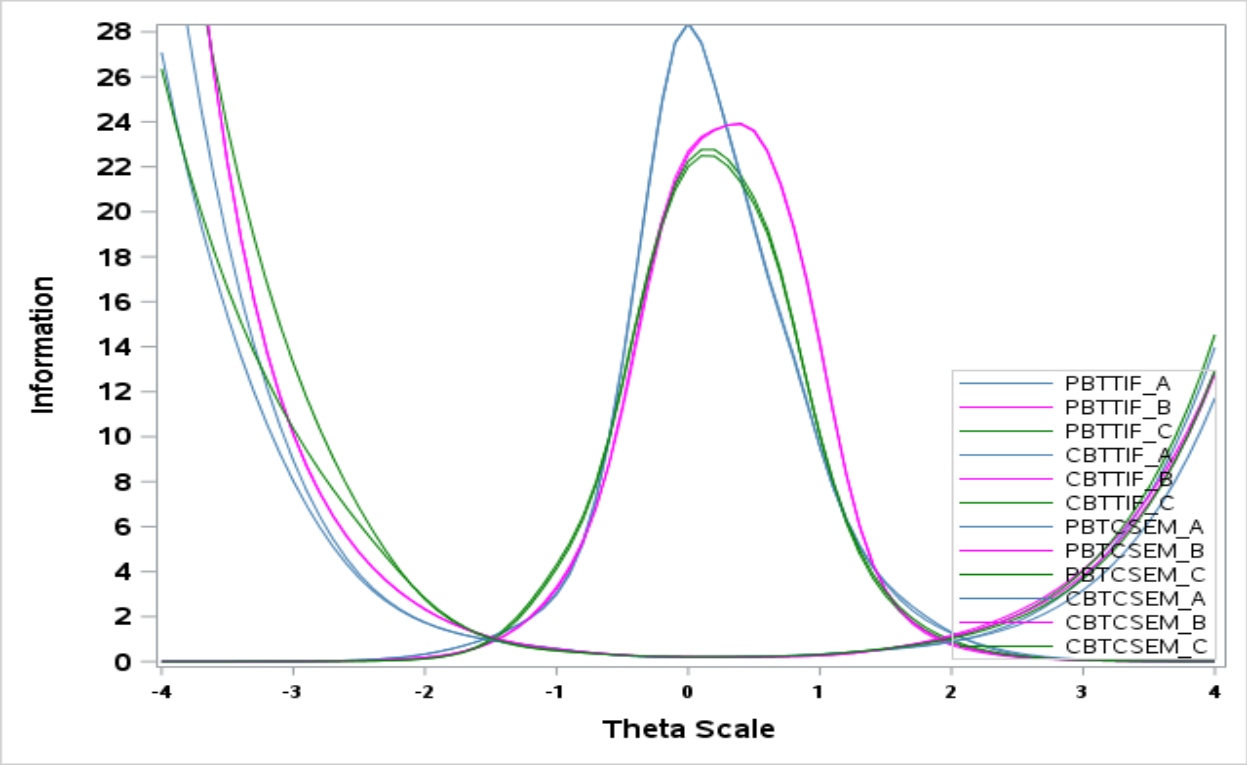



Figure 6.14 Math Grade 8 TIFs and CSEM 2018-19 Operational Forms

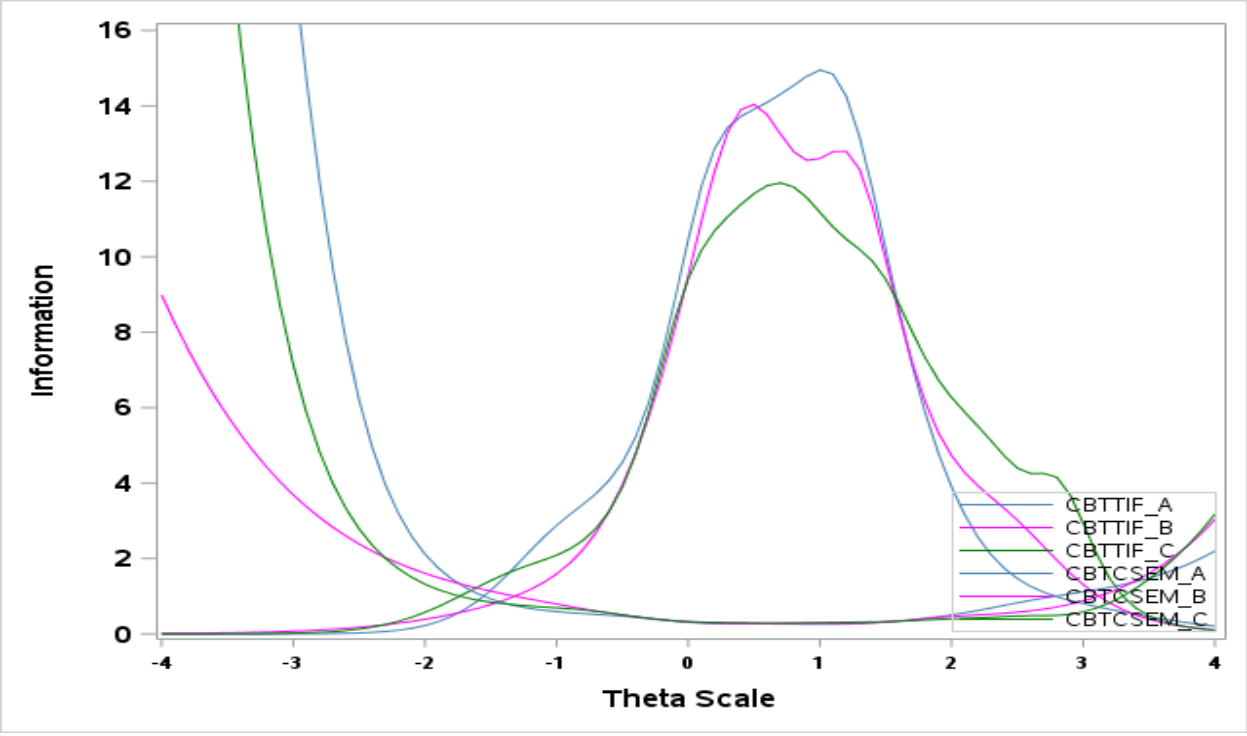


Figure 6.15 NC Math 1 TIFs and CSEM 2018-19 Operational Forms

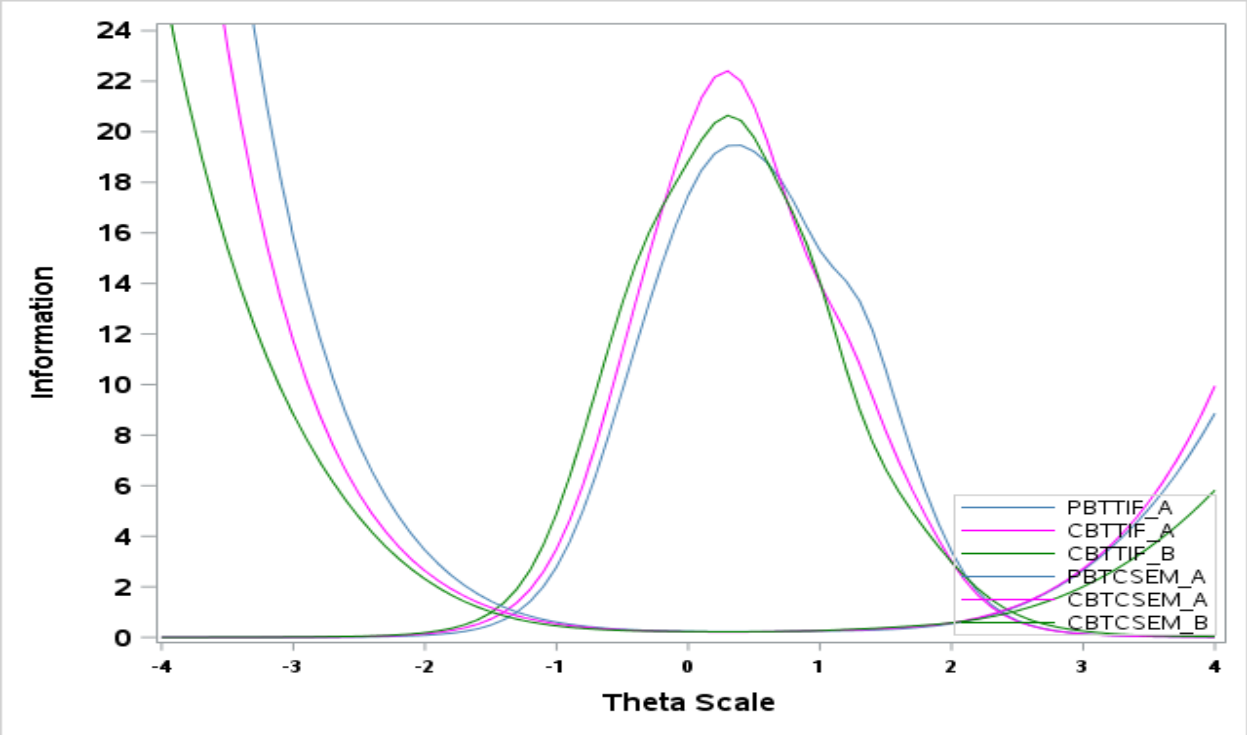
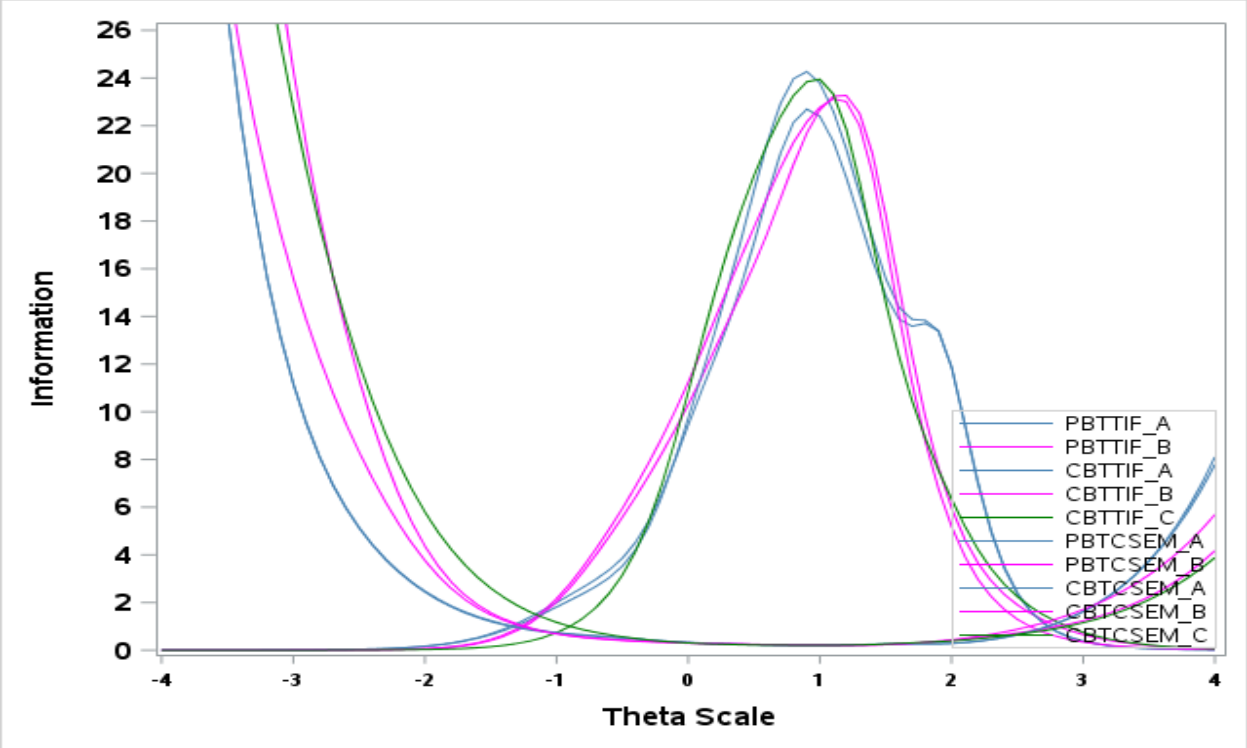



Figure 6.16 NC Math 3 TIFs and CSEM 2018-19 Operational Forms



# Appendix 7

## Appendix 7-A

## Math Standard Setting 2019 Technical Report

https://www.dpi.nc.gov/districts-schools/testing-and-school-accountability/technical-information-state-tests#standard-setting-resources-and-reports

## Appendix 7-B

## North Carolina Standard Setting Review Report (2019-07-17 FINAL)

https://www.dpi.nc.gov/districts-schools/testing-and-school-accountability/technical-information-state-tests#standard-setting-resources-and-reports

# Appendix 8

## Appendix 8-A

## Math 2018-19 Scale Score for Regular Students by Subgroups

*Table 1. 2018-19 Mathematics Scale Scores by Subgroups, Grades 3-5*

| Grade | Type | Category | N | Statistics | | Range | | Percentile | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Mean | SD | Min | Max | 25th | Median | 75th |
| | | Regular | 102313 | 550 | 10 | 523 | 570 | 543 | 550 | 557 |
| | SWD | Students with Disability | 13746 | 540 | 9 | 523 | 570 | 533 | 538 | 546 |
| | | Not Economically Disadvantaged | 59204 | 552 | 10 | 523 | 570 | 545 | 552 | 558 |
| | EDS | Economically Disadvantaged | 56855 | 545 | 9 | 523 | 570 | 538 | 545 | 552 |
| | | Regular | 101438 | 549 | 10 | 523 | 570 | 542 | 549 | 557 |
| | | Other | 2547 | 548 | 12 | 523 | 570 | 538 | 549 | 558 |
| | Els | English Language Learner | 12074 | 544 | 9 | 524 | 570 | 537 | 544 | 551 |
| EOG 3 | | All | 116059 | 548 | 10 | 523 | 570 | 541 | 549 | 556 |
| | | Regular | 105376 | 550 | 10 | 525 | 570 | 543 | 550 | 556 |
| | SWD | Students with Disability | 14944 | 540 | 9 | 525 | 570 | 533 | 538 | 545 |
| | | Not Economically Disadvantaged | 61348 | 552 | 10 | 525 | 570 | 545 | 552 | 559 |
| | EDS | Economically Disadvantaged | 58972 | 545 | 9 | 525 | 570 | 537 | 545 | 552 |
| | | Regular | 104308 | 549 | 10 | 525 | 570 | 541 | 549 | 556 |
| | | Other | 3669 | 550 | 11 | 525 | 570 | 542 | 551 | 559 |
| | ELs | English Language Learner | 12343 | 544 | 9 | 525 | 570 | 537 | 544 | 551 |
| EOG 4 | | All | 120320 | 548 | 10 | 525 | 570 | 541 | 549 | 556 |
| | | Regular | 107045 | 549 | 10 | 524 | 570 | 542 | 550 | 556 |
| | SWD | Students with Disability | 14890 | 538 | 8 | 524 | 570 | 532 | 536 | 543 |
| | | Not Economically Disadvantaged | 62764 | 551 | 10 | 524 | 570 | 545 | 552 | 558 |
| | EDS | Economically Disadvantaged | 59171 | 545 | 9 | 524 | 570 | 537 | 544 | 551 |
| | | Regular | 105942 | 549 | 10 | 524 | 570 | 541 | 549 | 556 |
| | | Other | 4935 | 550 | 10 | 524 | 570 | 543 | 551 | 557 |
| | ELs | English Language Learner | 11058 | 543 | 8 | 524 | 570 | 536 | 542 | 548 |
| EOG 5 | | All | 121935 | 548 | 10 | 524 | 570 | 540 | 548 | 556 |

*Table 2. 2018-19 Mathematics Scale Scores by Subgroups, Grade 6-8*

| Grade | Type | Category | N | Statistics | | Range | | Percentile | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Mean | SD | Min | Max | 25th | Median | 75th |
| | | Regular | 107076 | 549 | 10 | 527 | 573 | 542 | 550 | 556 |
| | SWD | Students with Disability | 14537 | 539 | 7 | 527 | 573 | 533 | 537 | 543 |
| | | Not Economically Disadvantaged | 62211 | 552 | 10 | 527 | 573 | 545 | 552 | 559 |
| | EDS | Economically Disadvantaged | 59402 | 545 | 9 | 527 | 573 | 537 | 544 | 551 |
| | | Regular | 105630 | 549 | 10 | 527 | 573 | 541 | 549 | 556 |
| | | Other | 10118 | 548 | 9 | 527 | 573 | 541 | 548 | 554 |
| | ELs | English Language Learner | 5865 | 540 | 7 | 528 | 573 | 535 | 538 | 544 |
| EOG 6 | | All | 121613 | 548 | 10 | 527 | 573 | 540 | 548 | 555 |
| | | Regular | 104244 | 549 | 9 | 528 | 573 | 542 | 549 | 556 |
| | SWD | Students with Disability | 14227 | 539 | 6 | 528 | 573 | 535 | 537 | 542 |
| | | Not Economically Disadvantaged | 62503 | 551 | 10 | 528 | 573 | 544 | 552 | 558 |
| | EDS | Economically Disadvantaged | 55968 | 544 | 8 | 529 | 573 | 537 | 543 | 551 |
| | | Regular | 103229 | 549 | 10 | 528 | 573 | 541 | 549 | 556 |
| | | Other | 10801 | 547 | 9 | 529 | 573 | 539 | 546 | 553 |
| | ELs | English Language Learner | 4441 | 540 | 7 | 529 | 573 | 535 | 538 | 544 |
| EOG 7 | | All | 118471 | 548 | 10 | 528 | 573 | 540 | 548 | 555 |
| | | Regular | 68294 | 540 | 9 | 517 | 570 | 532 | 540 | 547 |
| | SWD | Students with Disability | 12603 | 532 | 7 | 517 | 565 | 527 | 530 | 536 |
| | | Not Economically Disadvantaged | 36985 | 541 | 10 | 517 | 570 | 534 | 541 | 548 |
| | EDS | Economically Disadvantaged | 43912 | 537 | 9 | 518 | 570 | 529 | 536 | 543 |
| | | Regular | 72483 | 539 | 9 | 517 | 570 | 531 | 538 | 546 |
| | | Other | 4520 | 537 | 9 | 518 | 570 | 529 | 536 | 543 |
| | ELs | English Language Learner | 3894 | 533 | 8 | 519 | 565 | 527 | 531 | 538 |
| EOG 8 | | All | 80897 | 539 | 9 | 517 | 570 | 531 | 538 | 545 |

*Table 3. 2018-19 Mathematics Scale Scores by Subgroups, NC Math 1 and NC Math 3*

| Grade | Type | Category | N | Statistics | | Range | | Percentile | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Mean | SD | Min | Max | 25th | Median | 75th |
| NC Math 1 | | Regular | 104,350 | 550 | 9 | 528 | 575 | 543 | 550 | 557 |
| | SWD | Students with Disability | 13,646 | 540 | 7 | 528 | 575 | 535 | 538 | 543 |
| | EDS | Not Economically Disadvantaged | 65,824 | 552 | 10 | 528 | 575 | 545 | 553 | 559 |
| | | Economically Disadvantaged | 52,172 | 545 | 9 | 528 | 575 | 538 | 545 | 552 |
| | ELs | Regular | 109,357 | 550 | 10 | 528 | 575 | 542 | 550 | 556 |
| | | Other | 3,676 | 547 | 10 | 528 | 575 | 538 | 546 | 554 |
| | | English Language Learner | 4,963 | 541 | 7 | 529 | 573 | 535 | 539 | 545 |
| | | All | 117,996 | 549 | 10 | 528 | 575 | 541 | 549 | 556 |
| NC Math 3 | | Regular | 99,912 | 550 | 9 | 530 | 575 | 542 | 550 | 556 |
| | SWD | Students with Disability | 7,665 | 543 | 6 | 532 | 575 | 538 | 541 | 546 |
| | EDS | Not Economically Disadvantaged | 65,717 | 552 | 9 | 531 | 575 | 544 | 551 | 558 |
| | | Economically Disadvantaged | 41,860 | 546 | 8 | 530 | 575 | 540 | 545 | 551 |
| | ELs | Regular | 101,758 | 550 | 9 | 530 | 575 | 542 | 549 | 556 |
| | | Other | 1,794 | 549 | 9 | 531 | 575 | 541 | 548 | 555 |
| | | English Language Learner | 4,025 | 542 | 6 | 532 | 571 | 538 | 541 | 546 |
| | | All | 107,577 | 549 | 9 | 530 | 575 | 541 | 549 | 556 |

## Appendix 8-B

## Achievement Level Ranges and Descriptors

EOG: https://files.nc.gov/dpi/documents/files/achievement-level-ranges-and-alds-gen-math-eog.pdf

EOC: https://files.nc.gov/dpi/documents/files/achievement-level-ranges-and-alds-gen-math-eoc-081019_0.pdf

## Appendix 8-C

## Math 2018-19 Proficiency Classifications for Regular Students by Subgroups

*Table 1. 2018-19 Mathematics Proficiency Classifications by Subgroups, Grades 3-5*

| Grade | Type | Category | N | Level 2 and Below | Level 3 | Level 4 | Level 5 |
|---|---|---|---|---|---|---|---|
| 3 | | Regular | 102,313 | 21 | 33 | 15 | 31 |
| | SWD | Students with Disability | 13,746 | 13 | 13 | 3 | 71 |
| | EDS | Not Economically Disadvantaged | 59,204 | 19 | 37 | 21 | 23 |
| | | Economically Disadvantaged | 56,855 | 21 | 24 | 6 | 49 |
| | ELs | Regular | 101,438 | 20 | 32 | 15 | 33 |
| | | Other | 2,547 | 14 | 28 | 19 | 40 |
| | | English Language Learner | 12,074 | 21 | 22 | 5 | 52 |
| | | All | 116,059 | 20 | 31 | 14 | 36 |
| 4 | | Regular | 105,376 | 19 | 28 | 16 | 37 |
| | SWD | Students with Disability | 14,944 | 10 | 8 | 3 | 78 |
| | EDS | Not Economically Disadvantaged | 61,348 | 17 | 31 | 22 | 29 |
| | | Economically Disadvantaged | 58,972 | 18 | 19 | 6 | 57 |
| | ELs | Regular | 104,308 | 18 | 26 | 16 | 41 |
| | | Other | 3,669 | 15 | 28 | 21 | 35 |
| | | English Language Learner | 12,343 | 19 | 17 | 4 | 59 |
| | | All | 120,320 | 18 | 25 | 15 | 43 |
| 5 | | Regular | 107,045 | 19 | 34 | 13 | 34 |
| | SWD | Students with Disability | 14,890 | 10 | 8 | 1 | 80 |
| | EDS | Not Economically Disadvantaged | 62,764 | 17 | 38 | 18 | 26 |
| | | Economically Disadvantaged | 59,171 | 19 | 23 | 4 | 53 |
| | ELs | Regular | 105,942 | 18 | 32 | 12 | 37 |
| | | Other | 4,935 | 17 | 37 | 14 | 32 |
| | | English Language Learner | 11,058 | 19 | 16 | 2 | 63 |
| | | All | 121,935 | 18 | 31 | 11 | 40 |

Note: Level 2 and Below-Not Proficient, not CCR, Level 3- Sufficient Understanding, Not CCR, Level 4-Thorough Understanding, CCR, Level 5-Comprehensive Understanding, CCR

*Table 2. 2018-19 Mathematics Proficiency Classifications by Subgroups, Grades 6-8*

| Grade | Type | Category | N | Level 2 and Below | Level 3 | Level 4 | Level 5 |
|---|---|---|---|---|---|---|---|
| 6 | SWD | Regular | 107,076 | 18 | 33 | 14 | 35 |
| | | Students with Disability | 14,537 | 9 | 7 | 1 | 82 |
| | EDS | Not Economically Disadvantaged | 62,211 | 16 | 37 | 20 | 27 |
| | | Economically Disadvantaged | 59,402 | 18 | 22 | 4 | 56 |
| | ELs | Regular | 105,630 | 17 | 31 | 13 | 39 |
| | | Other | 10,118 | 20 | 30 | 9 | 41 |
| | | English Language Learner | 5,865 | 12 | 8 | 1 | 79 |
| | | All | 121,613 | 17 | 30 | 12 | 41 |
| 7 | SWD | Regular | 104,244 | 15 | 35 | 15 | 36 |
| | | Students with Disability | 14,227 | 8 | 7 | 1 | 84 |
| | EDS | Not Economically Disadvantaged | 62,503 | 13 | 39 | 21 | 27 |
| | | Economically Disadvantaged | 55,968 | 15 | 23 | 5 | 57 |
| | ELs | Regular | 103,229 | 14 | 33 | 14 | 39 |
| | | Other | 10,801 | 16 | 30 | 8 | 46 |
| | | English Language Learner | 4,441 | 10 | 10 | 2 | 79 |
| | | All | 118,471 | 14 | 31 | 13 | 42 |
| 8 | SWD | Regular | 68,294 | 18 | 16 | 7 | 60 |
| | | Students with Disability | 12,603 | 6 | 3 | 1 | 90 |
| | EDS | Not Economically Disadvantaged | 36,985 | 19 | 18 | 8 | 55 |
| | | Economically Disadvantaged | 43,912 | 14 | 10 | 3 | 73 |
| | ELs | Regular | 72,483 | 17 | 15 | 6 | 63 |
| | | Other | 4,520 | 14 | 11 | 4 | 72 |
| | | English Language Learner | 3,894 | 7 | 5 | 1 | 86 |
| | | All | 80,897 | 16 | 14 | 6 | 64 |

Note: Level 2 and Below-Not Proficient, not CCR, Level 3- Sufficient Understanding, Not CCR, Level 4-Thorough Understanding, CCR, Level 5-Comprehensive Understanding, CCR

*Table 3. 2018-19 Mathematics Proficiency Classifications by Subgroups, NC Math 1 and NC Math 3*

| Grade | Type | Category | N | Level 2 and Below | Level 3 | Level 4 | Level 5 |
|---|---|---|---|---|---|---|---|
| NC Math 1 | SWD | Regular | 104,350 | 28 | 25 | 10 | 38 |
| | | Students with Disability | 13,646 | 10 | 3 | 1 | 86 |
| | EDS | Not Economically Disadvantaged | 65,824 | 27 | 29 | 13 | 30 |
| | | Economically Disadvantaged | 52,172 | 24 | 14 | 3 | 60 |
| | ELs | Regular | 109,357 | 26 | 23 | 9 | 41 |
| | | Other | 3,676 | 22 | 16 | 8 | 54 |
| | | English Language Learner | 4,963 | 13 | 4 | 1 | 82 |
| | | All | 117,996 | 26 | 22 | 9 | 43 |
| NC Math 3 | SWD | Regular | 99,912 | 22 | 19 | 10 | 50 |
| | | Students with Disability | 7,665 | 9 | 3 | 1 | 87 |
| | EDS | Not Economically Disadvantaged | 65,717 | 22 | 22 | 13 | 43 |
| | | Economically Disadvantaged | 41,860 | 18 | 11 | 3 | 68 |
| | ELs | Regular | 101,758 | 21 | 18 | 9 | 51 |
| | | Other | 1,794 | 21 | 14 | 9 | 56 |
| | | English Language Learner | 4,025 | 9 | 3 | 0 | 88 |
| | | All | 107,577 | 21 | 18 | 9 | 53 |

Note: Level 2 and Below-Not Proficient, not CCR, Level 3- Sufficient Understanding, Not CCR, Level 4-Thorough Understanding, CCR, Level 5-Comprehensive Understanding, CCR

## Appendix 8-D

**Interpretive Guide to the Score Reports for the North Carolina End-of Grade Assessments, 2018–19**
https://files.nc.gov/dpi/documents/files/1819eogwsguidefinal.pdf

# Appendix 9

# Appendix 9-A

## Two Factors Exploratory Factor Analysis with Simple Structure Math 2018-19

Grade 3

| Order | Form A/M | | Form B/N | | Form C/O | |
|---|---|---|---|---|---|---|
| | Factor | | Factor | | Factor | |
| | 1 | 2 | 1 | 2 | 1 | 2 |
| 1 | 0.44 | -0.48 | 0.66 | -0.17 | 0.74 | 0.23 |
| 2 | 0.57 | -0.20 | 0.49 | -0.21 | 0.56 | -0.51 |
| 3 | 0.58 | 0.50 | 0.77 | -0.31 | 0.57 | 0.39 |
| 4 | 0.68 | 0.00 | 0.70 | 0.07 | 0.66 | -0.19 |
| 5 | 0.26 | 0.56 | 0.69 | -0.40 | 0.42 | -0.37 |
| 6 | 0.55 | -0.41 | 0.15 | 0.05 | 0.42 | 0.59 |
| 7 | -0.11 | 0.40 | 0.28 | 0.12 | 0.60 | -0.52 |
| 8 | 0.64 | 0.56 | 0.30 | 0.38 | 0.88 | -0.19 |
| 9 | 0.57 | -0.31 | 0.71 | -0.45 | 0.71 | 0.55 |
| 10 | 0.02 | -0.10 | 0.74 | -0.07 | 0.28 | -0.55 |
| 11 | 0.27 | -0.34 | 0.59 | -0.16 | 0.80 | 0.40 |
| 12 | -0.04 | 0.46 | 0.64 | -0.45 | 0.84 | -0.07 |
| 13 | 0.67 | -0.13 | 0.42 | 0.11 | 0.64 | -0.16 |
| 14 | 0.45 | -0.31 | 0.57 | 0.26 | -0.47 | 0.16 |
| 15 | 0.66 | -0.12 | 0.04 | 0.34 | 0.80 | -0.11 |
| 16 | 0.27 | -0.32 | 0.23 | 0.28 | 0.68 | 0.13 |
| 17 | 0.58 | 0.35 | 0.63 | 0.14 | 0.67 | -0.23 |
| 18 | 0.03 | 0.22 | 0.71 | -0.13 | 0.34 | 0.18 |
| 19 | 0.29 | -0.15 | 0.02 | 0.11 | 0.07 | 0.19 |
| 20 | 0.79 | 0.40 | 0.61 | 0.66 | 0.81 | 0.44 |
| 21 | 0.27 | -0.21 | 0.64 | -0.11 | 0.59 | -0.27 |
| 22 | 0.57 | 0.36 | 0.75 | 0.52 | 0.82 | 0.03 |
| 23 | 0.40 | -0.42 | 0.84 | 0.35 | 0.55 | -0.36 |
| 24 | 0.47 | -0.01 | -0.39 | 0.33 | 0.11 | 0.41 |
| 25 | 0.59 | -0.13 | 0.71 | -0.26 | 0.85 | -0.22 |
| 26 | 0.43 | 0.26 | 0.59 | -0.10 | 0.88 | -0.25 |
| 27 | 0.49 | 0.50 | 0.87 | 0.32 | 0.38 | 0.54 |
| 28 | 0.53 | 0.04 | 0.13 | 0.17 | 0.54 | -0.14 |
| 29 | 0.58 | -0.42 | 0.66 | -0.20 | 0.69 | 0.34 |
| 30 | 0.48 | 0.58 | 0.63 | 0.66 | 0.15 | 0.33 |
| 31 | 0.50 | 0.44 | 0.85 | -0.11 | 0.51 | 0.27 |
| 32 | 0.62 | -0.13 | -0.13 | 0.24 | 0.49 | 0.54 |
| 33 | 0.39 | 0.21 | 0.52 | 0.29 | 0.45 | 0.17 |
| 34 | 0.76 | -0.34 | 0.41 | -0.18 | 0.40 | 0.50 |
| 35 | 0.55 | -0.24 | 0.78 | -0.23 | 0.81 | -0.03 |
| 36 | 0.56 | -0.29 | 0.72 | 0.53 | 0.51 | 0.12 |

| Order | Form A/M | | Form B/N | | Form C/O | |
|---|---|---|---|---|---|---|
| | Factor | | Factor | | Factor | |
| | 1 | 2 | 1 | 2 | 1 | 2 |
| 37 | 0.42 | 0.62 | 0.49 | 0.13 | 0.30 | 0.20 |
| 38 | 0.05 | -0.27 | 0.51 | 0.03 | 0.56 | -0.16 |
| 39 | 0.60 | -0.28 | 0.71 | -0.30 | 0.70 | -0.48 |
| 40 | 0.39 | -0.17 | 0.67 | -0.34 | 0.59 | -0.42 |

Grade 4

| Order | Form A | | Form B | | Form C | | Form M | | Form N | | Form O | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Factor | | Factor | | Factor | | Factor | | Factor | | Factor | |
| | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| 1 | 0.74 | -0.45 | 0.47 | -0.25 | 0.66 | -0.33 | 0.77 | -0.41 | 0.45 | -0.28 | 0.62 | -0.41 |
| 2 | 0.15 | -0.27 | 0.88 | -0.12 | 0.53 | 0.35 | 0.24 | -0.23 | 0.87 | -0.22 | 0.61 | 0.35 |
| 3 | 0.78 | 0.10 | 0.69 | 0.02 | 0.77 | -0.37 | 0.82 | 0.05 | 0.64 | 0.05 | 0.73 | -0.46 |
| 4 | 0.72 | -0.32 | 0.08 | -0.21 | 0.58 | -0.53 | 0.73 | -0.34 | -0.12 | -0.07 | 0.56 | -0.53 |
| 5 | 0.46 | 0.32 | 0.76 | -0.31 | 0.83 | 0.07 | 0.47 | 0.47 | 0.73 | -0.41 | 0.86 | 0.01 |
| 6 | 0.81 | -0.35 | -0.02 | 0.48 | 0.70 | -0.53 | 0.81 | -0.40 | 0.00 | 0.54 | 0.71 | -0.55 |
| 8 | 0.29 | 0.38 | 0.65 | -0.37 | 0.39 | 0.28 | 0.24 | 0.43 | 0.61 | -0.42 | 0.39 | 0.16 |
| 9 | 0.60 | 0.21 | 0.80 | -0.43 | 0.80 | -0.02 | 0.56 | 0.28 | 0.77 | -0.47 | 0.80 | -0.06 |
| 10 | -0.20 | 0.52 | -0.16 | 0.42 | 0.18 | -0.33 | -0.29 | 0.46 | -0.08 | 0.49 | 0.23 | -0.37 |
| 11 | 0.53 | 0.54 | 0.70 | -0.43 | 0.04 | 0.46 | 0.61 | 0.47 | 0.63 | -0.49 | 0.28 | 0.45 |
| 12 | 0.79 | -0.09 | 0.74 | 0.28 | 0.55 | 0.54 | 0.77 | -0.04 | 0.66 | 0.39 | 0.64 | 0.41 |
| 13 | 0.77 | -0.38 | 0.78 | -0.35 | 0.70 | -0.58 | 0.78 | -0.37 | 0.73 | -0.44 | 0.65 | -0.62 |
| 15 | 0.76 | -0.33 | 0.52 | 0.31 | 0.52 | -0.50 | 0.77 | -0.36 | 0.48 | 0.38 | 0.52 | -0.45 |
| 16 | 0.58 | 0.35 | -0.17 | 0.08 | 0.76 | 0.14 | 0.57 | 0.34 | -0.15 | 0.19 | 0.72 | 0.14 |
| 17 | 0.20 | 0.40 | 0.77 | -0.02 | 0.51 | -0.34 | 0.30 | 0.40 | 0.75 | 0.00 | 0.43 | -0.41 |
| 18 | 0.58 | 0.05 | 0.43 | 0.49 | 0.16 | 0.27 | 0.66 | 0.00 | 0.45 | 0.51 | 0.19 | 0.29 |
| 19 | 0.32 | -0.10 | 0.77 | 0.48 | 0.58 | 0.45 | 0.33 | -0.10 | 0.78 | 0.47 | 0.64 | 0.47 |
| 20 | 0.63 | 0.15 | 0.46 | 0.02 | 0.65 | 0.17 | 0.58 | 0.09 | 0.42 | 0.05 | 0.68 | 0.14 |
| 22 | 0.67 | 0.45 | 0.64 | 0.00 | 0.72 | -0.01 | 0.68 | 0.44 | 0.56 | 0.13 | 0.66 | 0.06 |
| 23 | 0.54 | -0.12 | 0.59 | 0.02 | 0.71 | 0.49 | 0.61 | -0.13 | 0.60 | 0.05 | 0.75 | 0.42 |
| 24 | 0.53 | 0.29 | -0.22 | 0.17 | 0.28 | 0.11 | 0.51 | 0.38 | -0.12 | 0.16 | 0.18 | 0.06 |
| 25 | 0.66 | 0.31 | 0.70 | 0.16 | 0.30 | 0.27 | 0.65 | 0.33 | 0.76 | 0.15 | 0.35 | 0.30 |
| 26 | 0.84 | -0.09 | 0.21 | 0.17 | 0.46 | 0.23 | 0.85 | -0.07 | 0.13 | 0.26 | 0.57 | 0.18 |
| 28 | -0.12 | -0.03 | 0.72 | 0.37 | 0.63 | 0.17 | -0.16 | 0.02 | 0.72 | 0.33 | 0.67 | 0.00 |
| 29 | -0.33 | 0.34 | 0.59 | -0.29 | -0.25 | 0.13 | -0.35 | 0.28 | 0.56 | -0.27 | -0.27 | 0.23 |
| 30 | 0.13 | 0.12 | 0.43 | 0.33 | 0.70 | 0.19 | -0.11 | 0.07 | 0.37 | 0.34 | 0.71 | 0.28 |
| 31 | 0.39 | 0.48 | 0.47 | 0.51 | 0.18 | -0.01 | 0.34 | 0.45 | 0.40 | 0.60 | 0.11 | 0.03 |
| 32 | -0.01 | 0.40 | 0.80 | -0.11 | 0.56 | 0.03 | 0.04 | 0.37 | 0.78 | -0.17 | 0.53 | -0.01 |

| Order | Form A | | Form B | | Form C | | Form M | | Form N | | Form O | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Factor | | Factor | | Factor | | Factor | | Factor | | Factor | |
| | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| 33 | 0.03 | 0.49 | 0.62 | 0.28 | -0.16 | 0.12 | 0.09 | 0.42 | 0.64 | 0.31 | -0.10 | 0.27 |
| 34 | 0.18 | 0.75 | 0.82 | -0.02 | 0.46 | 0.48 | 0.19 | 0.74 | 0.84 | 0.00 | 0.56 | 0.39 |
| 36 | 0.13 | 0.21 | 0.65 | 0.28 | 0.80 | -0.04 | 0.13 | 0.17 | 0.67 | 0.32 | 0.81 | -0.03 |
| 37 | 0.51 | 0.25 | 0.86 | 0.00 | 0.65 | 0.50 | 0.50 | 0.27 | 0.85 | -0.06 | 0.70 | 0.50 |
| 38 | 0.56 | 0.37 | 0.76 | 0.00 | 0.65 | 0.17 | 0.53 | 0.42 | 0.74 | 0.10 | 0.66 | 0.17 |
| 39 | -0.20 | -0.12 | 0.91 | 0.02 | 0.75 | -0.08 | -0.25 | -0.13 | 0.91 | -0.02 | 0.77 | -0.06 |
| 40 | 0.67 | 0.14 | 0.50 | 0.41 | 0.25 | 0.09 | 0.64 | 0.17 | 0.50 | 0.38 | 0.27 | 0.15 |
| 41 | -0.13 | 0.54 | 0.76 | -0.44 | 0.83 | -0.04 | -0.07 | 0.57 | 0.72 | -0.47 | 0.83 | -0.05 |
| 42 | 0.74 | -0.37 | 0.12 | 0.17 | 0.46 | -0.07 | 0.75 | -0.39 | 0.05 | 0.24 | 0.42 | -0.18 |
| 44 | 0.77 | -0.33 | 0.64 | 0.15 | 0.64 | -0.09 | 0.81 | -0.32 | 0.62 | 0.20 | 0.59 | 0.09 |
| 45 | 0.53 | 0.26 | 0.76 | -0.14 | 0.40 | 0.18 | 0.48 | 0.24 | 0.75 | -0.19 | 0.36 | 0.21 |
| 46 | 0.81 | -0.37 | 0.57 | -0.03 | 0.67 | -0.56 | 0.82 | -0.37 | 0.59 | 0.05 | 0.69 | -0.52 |

Grade 5

| Order | Form A | | Form B | | Form C | | Form M | | Form N | | Form O | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Factor | | Factor | | Factor | | Factor | | Factor | | Factor | |
| | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| 1 | 0.65 | -0.01 | 0.83 | -0.05 | 0.70 | 0.01 | 0.64 | -0.02 | 0.79 | -0.14 | 0.71 | -0.03 |
| 2 | 0.65 | 0.56 | 0.66 | 0.34 | 0.65 | 0.51 | 0.70 | 0.53 | 0.62 | 0.39 | 0.74 | 0.42 |
| 3 | 0.56 | -0.41 | 0.77 | -0.29 | 0.61 | -0.08 | 0.47 | -0.45 | 0.74 | -0.31 | 0.58 | 0.10 |
| 5 | 0.58 | 0.17 | 0.94 | -0.04 | 0.61 | -0.19 | 0.61 | 0.09 | 0.92 | -0.06 | 0.58 | 0.01 |
| 6 | 0.34 | 0.52 | 0.88 | -0.04 | 0.42 | -0.24 | 0.38 | 0.50 | 0.89 | -0.05 | 0.35 | -0.35 |
| 7 | 0.12 | -0.16 | -0.08 | -0.11 | 0.71 | 0.24 | 0.04 | -0.09 | -0.12 | -0.11 | 0.72 | 0.32 |
| 8 | 0.55 | -0.04 | 0.62 | 0.08 | 0.53 | 0.03 | 0.59 | -0.04 | 0.58 | 0.06 | 0.52 | -0.06 |
| 10 | 0.73 | -0.24 | 0.09 | 0.58 | 0.41 | 0.04 | 0.67 | -0.26 | 0.19 | 0.45 | 0.31 | 0.12 |
| 11 | 0.16 | 0.72 | 0.20 | 0.69 | 0.49 | -0.02 | 0.22 | 0.73 | 0.18 | 0.76 | 0.48 | -0.09 |
| 12 | 0.65 | 0.08 | 0.40 | 0.11 | 0.75 | -0.32 | 0.63 | 0.11 | 0.37 | 0.11 | 0.75 | -0.35 |
| 13 | 0.28 | 0.63 | 0.80 | -0.22 | 0.48 | -0.32 | 0.39 | 0.59 | 0.77 | -0.28 | 0.46 | -0.43 |
| 15 | 0.85 | 0.14 | 0.90 | -0.15 | 0.76 | -0.13 | 0.88 | 0.05 | 0.85 | -0.22 | 0.79 | -0.03 |
| 16 | 0.81 | 0.12 | 0.85 | -0.28 | 0.77 | -0.07 | 0.83 | 0.04 | 0.82 | -0.29 | 0.74 | 0.01 |
| 17 | 0.86 | -0.07 | 0.64 | 0.61 | 0.06 | 0.51 | 0.86 | -0.13 | 0.62 | 0.63 | 0.01 | 0.56 |
| 18 | 0.48 | 0.26 | 0.95 | -0.07 | 0.74 | -0.20 | 0.49 | 0.28 | 0.95 | -0.10 | 0.75 | -0.28 |
| 19 | 0.79 | -0.08 | 0.84 | -0.25 | 0.03 | 0.44 | 0.78 | -0.14 | 0.79 | -0.31 | -0.02 | 0.35 |
| 20 | -0.17 | 0.22 | 0.68 | 0.00 | 0.36 | 0.47 | -0.05 | 0.14 | 0.68 | 0.04 | 0.37 | 0.51 |
| 22 | 0.70 | 0.04 | 0.86 | 0.07 | 0.80 | -0.18 | 0.73 | 0.02 | 0.79 | 0.08 | 0.82 | -0.19 |
| 23 | 0.29 | -0.26 | 0.03 | 0.12 | 0.57 | 0.01 | 0.33 | -0.20 | 0.01 | 0.16 | 0.57 | -0.01 |
| 24 | 0.66 | -0.21 | 0.04 | -0.16 | 0.64 | 0.06 | 0.63 | -0.23 | 0.07 | -0.11 | 0.62 | 0.05 |

| Order | Form A | | Form B | | Form C | | Form M | | Form N | | Form O | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Factor | | Factor | | Factor | | Factor | | Factor | | Factor | |
| | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| 25 | -0.44 | 0.16 | 0.38 | 0.29 | 0.34 | -0.09 | -0.42 | 0.30 | 0.44 | 0.29 | 0.38 | -0.02 |
| 26 | 0.80 | -0.27 | 0.57 | 0.30 | 0.49 | -0.13 | 0.74 | -0.30 | 0.57 | 0.34 | 0.59 | -0.13 |
| 28 | 0.75 | -0.08 | 0.93 | 0.12 | 0.64 | -0.06 | 0.74 | -0.09 | 0.94 | 0.11 | 0.68 | 0.05 |
| 29 | 0.32 | 0.12 | 0.66 | -0.04 | 0.66 | 0.31 | 0.23 | 0.20 | 0.63 | 0.00 | 0.65 | 0.30 |
| 30 | 0.15 | 0.33 | 0.41 | -0.07 | 0.07 | 0.16 | 0.20 | 0.32 | 0.32 | 0.08 | 0.08 | 0.09 |
| 31 | 0.33 | -0.15 | 0.72 | -0.06 | -0.03 | 0.18 | 0.25 | -0.08 | 0.69 | -0.09 | 0.02 | 0.01 |
| 32 | 0.42 | 0.32 | -0.16 | 0.12 | 0.83 | -0.20 | 0.49 | 0.29 | -0.10 | 0.21 | 0.83 | -0.22 |
| 33 | 0.37 | -0.05 | 0.56 | -0.23 | 0.13 | 0.33 | 0.27 | -0.16 | 0.51 | -0.22 | 0.12 | 0.24 |
| 35 | 0.81 | -0.03 | 0.74 | -0.20 | 0.55 | 0.20 | 0.81 | -0.10 | 0.71 | -0.16 | 0.55 | 0.18 |
| 36 | 0.63 | -0.15 | 0.33 | -0.04 | -0.25 | -0.29 | 0.56 | -0.16 | 0.37 | -0.05 | -0.21 | -0.32 |
| 37 | 0.34 | -0.13 | 0.45 | -0.09 | 0.28 | -0.18 | 0.33 | 0.03 | 0.44 | -0.07 | 0.25 | -0.24 |
| 38 | 0.06 | 0.42 | 0.11 | 0.26 | 0.66 | -0.16 | 0.26 | 0.42 | 0.16 | 0.21 | 0.67 | -0.24 |
| 39 | 0.75 | -0.03 | 0.01 | 0.47 | 0.53 | -0.10 | 0.79 | -0.02 | 0.08 | 0.49 | 0.52 | -0.17 |
| 41 | 0.70 | 0.21 | 0.38 | 0.45 | 0.49 | 0.65 | 0.76 | 0.19 | 0.44 | 0.40 | 0.56 | 0.63 |
| 42 | 0.03 | 0.30 | 0.74 | 0.08 | -0.31 | -0.05 | 0.10 | 0.38 | 0.74 | 0.04 | -0.23 | 0.03 |
| 43 | 0.72 | 0.04 | 0.77 | -0.10 | 0.75 | -0.16 | 0.72 | 0.02 | 0.72 | -0.12 | 0.76 | -0.09 |
| 44 | 0.86 | -0.08 | 0.76 | 0.12 | 0.73 | 0.17 | 0.88 | -0.02 | 0.73 | 0.14 | 0.76 | 0.17 |
| 46 | 0.72 | -0.08 | 0.79 | -0.08 | 0.51 | 0.00 | 0.70 | -0.11 | 0.74 | -0.09 | 0.56 | -0.06 |
| 47 | 0.77 | -0.25 | 0.75 | 0.30 | 0.08 | 0.75 | 0.67 | -0.30 | 0.74 | 0.35 | 0.10 | 0.70 |
| 48 | 0.81 | -0.05 | -0.23 | 0.78 | 0.84 | -0.02 | 0.76 | -0.05 | -0.24 | 0.80 | 0.87 | -0.12 |

Grade 6

| Order | Form A | | Form B | | Form C | | Form M | | Form N | | Form O | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Factor | | Factor | | Factor | | Factor | | Factor | | Factor | |
| | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| 1 | 0.66 | -0.31 | 0.77 | 0.17 | 0.82 | -0.06 | 0.63 | -0.31 | 0.79 | 0.16 | 0.83 | -0.05 |
| 2 | 0.48 | 0.19 | 0.64 | 0.30 | 0.66 | 0.21 | 0.51 | 0.28 | 0.63 | 0.25 | 0.67 | 0.25 |
| 3 | 0.58 | 0.04 | 0.63 | -0.25 | 0.39 | -0.37 | 0.60 | 0.14 | 0.65 | -0.27 | 0.38 | -0.34 |
| 5 | 0.54 | 0.12 | 0.49 | -0.03 | -0.08 | 0.29 | 0.59 | 0.25 | 0.51 | -0.04 | -0.10 | 0.27 |
| 6 | 0.61 | -0.49 | 0.10 | -0.13 | -0.17 | 0.34 | 0.58 | -0.39 | 0.00 | -0.15 | -0.03 | 0.30 |
| 7 | 0.02 | 0.07 | 0.55 | -0.31 | 0.81 | 0.05 | -0.02 | 0.24 | 0.43 | -0.26 | 0.82 | 0.07 |
| 8 | 0.81 | -0.01 | -0.10 | 0.14 | 0.11 | 0.22 | 0.84 | -0.04 | -0.07 | 0.09 | 0.18 | 0.20 |
| 10 | 0.58 | 0.01 | -0.02 | 0.46 | 0.32 | 0.15 | 0.54 | 0.12 | -0.02 | 0.34 | 0.28 | 0.18 |
| 11 | 0.46 | 0.03 | 0.60 | 0.23 | 0.73 | -0.12 | 0.43 | 0.12 | 0.64 | 0.13 | 0.73 | -0.20 |
| 12 | 0.51 | 0.07 | 0.22 | 0.11 | 0.39 | 0.12 | 0.58 | 0.17 | 0.21 | 0.10 | 0.40 | 0.11 |
| 13 | 0.53 | -0.48 | 0.67 | -0.31 | 0.23 | -0.22 | 0.54 | -0.35 | 0.67 | -0.34 | 0.16 | -0.16 |
| 14 | 0.73 | -0.29 | 0.82 | -0.02 | 0.85 | -0.02 | 0.73 | -0.28 | 0.77 | -0.05 | 0.85 | -0.01 |

| Order | Form A | | Form B | | Form C | | Form M | | Form N | | Form O | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Factor | | Factor | | Factor | | Factor | | Factor | | Factor | |
| | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| 16 | 0.28 | -0.20 | 0.64 | 0.05 | 0.76 | 0.30 | 0.40 | 0.01 | 0.65 | 0.03 | 0.81 | 0.22 |
| 17 | 0.53 | -0.47 | -0.13 | 0.36 | 0.72 | -0.33 | 0.51 | -0.42 | -0.12 | 0.35 | 0.64 | -0.37 |
| 18 | 0.34 | 0.11 | 0.14 | 0.31 | 0.70 | 0.02 | 0.40 | 0.18 | 0.10 | 0.33 | 0.71 | -0.04 |
| 19 | 0.80 | 0.01 | 0.65 | 0.35 | 0.76 | 0.15 | 0.83 | 0.05 | 0.67 | 0.33 | 0.72 | 0.15 |
| 20 | 0.36 | 0.17 | -0.15 | 0.18 | 0.65 | 0.06 | 0.42 | 0.23 | -0.03 | 0.10 | 0.61 | 0.11 |
| 22 | 0.52 | 0.40 | 0.77 | -0.21 | 0.48 | 0.27 | 0.50 | 0.50 | 0.76 | -0.21 | 0.54 | 0.19 |
| 23 | 0.47 | -0.12 | 0.05 | -0.21 | -0.15 | 0.12 | 0.45 | -0.04 | -0.07 | -0.13 | -0.05 | 0.21 |
| 24 | -0.20 | 0.00 | 0.62 | 0.32 | 0.69 | -0.02 | -0.22 | 0.10 | 0.65 | 0.36 | 0.69 | 0.01 |
| 25 | 0.75 | -0.15 | 0.68 | 0.18 | 0.82 | 0.11 | 0.78 | -0.18 | 0.61 | 0.29 | 0.81 | 0.13 |
| 26 | -0.22 | 0.39 | 0.27 | 0.64 | -0.18 | 0.38 | -0.26 | 0.33 | 0.35 | 0.59 | -0.03 | 0.36 |
| 28 | 0.71 | -0.05 | 0.82 | -0.34 | 0.85 | 0.01 | 0.73 | -0.07 | 0.80 | -0.33 | 0.85 | -0.01 |
| 29 | 0.28 | 0.08 | 0.75 | -0.19 | 0.12 | 0.24 | 0.34 | -0.14 | 0.73 | -0.19 | 0.08 | 0.23 |
| 30 | 0.60 | 0.17 | 0.16 | 0.31 | 0.80 | -0.01 | 0.58 | 0.04 | 0.31 | 0.30 | 0.78 | -0.06 |
| 31 | -0.13 | -0.07 | 0.70 | 0.22 | 0.78 | 0.35 | -0.03 | -0.17 | 0.67 | 0.25 | 0.78 | 0.32 |
| 32 | -0.20 | 0.44 | 0.80 | -0.20 | 0.48 | -0.56 | -0.16 | 0.47 | 0.78 | -0.19 | 0.46 | -0.57 |
| 33 | 0.61 | -0.13 | 0.71 | 0.07 | 0.45 | -0.39 | 0.62 | -0.24 | 0.68 | 0.17 | 0.33 | -0.48 |
| 35 | 0.78 | 0.03 | 0.71 | -0.33 | 0.89 | 0.08 | 0.77 | 0.11 | 0.71 | -0.34 | 0.90 | 0.10 |
| 36 | 0.54 | 0.31 | 0.76 | 0.13 | 0.67 | 0.06 | 0.51 | 0.20 | 0.73 | 0.11 | 0.67 | 0.14 |
| 37 | 0.76 | 0.20 | 0.83 | -0.05 | 0.11 | -0.22 | 0.73 | 0.11 | 0.81 | -0.14 | 0.09 | -0.26 |
| 38 | 0.61 | 0.17 | 0.72 | -0.35 | 0.61 | 0.23 | 0.65 | 0.02 | 0.70 | -0.34 | 0.60 | 0.18 |
| 39 | 0.66 | 0.23 | 0.54 | -0.13 | -0.08 | 0.46 | 0.66 | 0.12 | 0.53 | 0.02 | -0.07 | 0.44 |
| 41 | 0.70 | 0.07 | 0.83 | 0.16 | 0.85 | 0.02 | 0.71 | 0.09 | 0.83 | 0.09 | 0.87 | 0.02 |
| 42 | 0.57 | 0.40 | 0.61 | 0.15 | 0.41 | 0.45 | 0.56 | 0.43 | 0.54 | 0.14 | 0.43 | 0.36 |
| 43 | 0.78 | 0.01 | 0.73 | 0.05 | 0.69 | -0.20 | 0.82 | 0.02 | 0.77 | -0.08 | 0.60 | -0.22 |
| 44 | 0.39 | 0.19 | 0.86 | -0.21 | 0.77 | -0.16 | 0.43 | 0.29 | 0.86 | -0.20 | 0.76 | -0.21 |
| 46 | 0.57 | 0.27 | 0.79 | 0.27 | -0.09 | -0.08 | 0.61 | 0.09 | 0.80 | 0.19 | -0.12 | -0.03 |
| 47 | 0.64 | 0.16 | 0.57 | 0.42 | 0.60 | -0.03 | 0.74 | 0.02 | 0.57 | 0.47 | 0.56 | -0.08 |
| 48 | 0.50 | 0.15 | 0.80 | -0.01 | 0.67 | -0.28 | 0.50 | 0.17 | 0.76 | 0.04 | 0.60 | -0.36 |
| 49 | 0.73 | -0.33 | 0.58 | -0.17 | 0.74 | -0.11 | 0.73 | -0.39 | 0.58 | -0.18 | 0.74 | -0.15 |
| 50 | 0.55 | -0.32 | 0.78 | -0.10 | 0.80 | -0.08 | 0.48 | -0.38 | 0.79 | -0.18 | 0.81 | -0.07 |
| 51 | 0.85 | 0.01 | 0.59 | 0.28 | 0.81 | 0.25 | 0.89 | 0.02 | 0.59 | 0.29 | 0.81 | 0.29 |
| 52 | 0.82 | -0.16 | 0.83 | -0.08 | 0.50 | -0.23 | 0.85 | -0.19 | 0.81 | -0.15 | 0.45 | -0.22 |
| 53 | 0.73 | 0.40 | 0.77 | -0.07 | 0.87 | 0.14 | 0.77 | 0.31 | 0.79 | 0.03 | 0.86 | 0.16 |

Grade 7

| Order | Form A | | Form B | | Form C | | Form M | | Form N | | Form O | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Factor | | Factor | | Factor | | Factor | | Factor | | Factor | |
| | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| 1 | 0.86 | -0.07 | 0.94 | -0.17 | 0.84 | -0.06 | 0.89 | -0.02 | 0.93 | -0.13 | 0.85 | -0.12 |
| 2 | 0.84 | -0.19 | 0.91 | 0.31 | 0.69 | -0.03 | 0.84 | -0.12 | 0.92 | 0.20 | 0.77 | 0.16 |
| 3 | 0.79 | 0.44 | 0.65 | -0.05 | 0.41 | 0.46 | 0.79 | 0.41 | 0.64 | 0.18 | 0.47 | 0.35 |
| 5 | 0.62 | -0.11 | 0.24 | 0.29 | 0.44 | 0.11 | 0.66 | -0.06 | 0.35 | 0.20 | 0.54 | 0.10 |
| 6 | 0.41 | 0.64 | 0.75 | -0.31 | 0.60 | 0.23 | 0.51 | 0.64 | 0.73 | -0.33 | 0.70 | 0.07 |
| 7 | 0.68 | 0.50 | 0.69 | -0.03 | 0.59 | 0.07 | 0.73 | 0.46 | 0.72 | 0.29 | 0.67 | 0.15 |
| 8 | 0.21 | -0.20 | 0.39 | -0.11 | 0.30 | 0.29 | 0.25 | -0.21 | 0.38 | -0.09 | 0.41 | 0.36 |
| 10 | 0.87 | 0.14 | 0.70 | -0.07 | 0.39 | 0.28 | 0.89 | 0.09 | 0.65 | 0.26 | 0.51 | 0.11 |
| 11 | 0.81 | -0.07 | 0.49 | -0.43 | 0.35 | 0.50 | 0.84 | -0.12 | 0.46 | -0.47 | 0.38 | 0.46 |
| 12 | 0.60 | 0.51 | 0.39 | 0.21 | 0.46 | -0.14 | 0.61 | 0.54 | 0.49 | 0.38 | 0.48 | -0.13 |
| 13 | 0.90 | -0.01 | 0.39 | -0.12 | 0.46 | -0.11 | 0.92 | -0.05 | 0.44 | -0.32 | 0.52 | -0.17 |
| 14 | 0.82 | -0.13 | 0.91 | -0.03 | 0.85 | -0.19 | 0.86 | -0.09 | 0.91 | 0.05 | 0.86 | -0.05 |
| 16 | 0.68 | -0.32 | 0.60 | 0.10 | 0.82 | -0.20 | 0.73 | -0.30 | 0.68 | 0.20 | 0.83 | -0.28 |
| 17 | 0.53 | -0.31 | 0.80 | -0.21 | 0.80 | -0.32 | 0.48 | -0.44 | 0.83 | -0.22 | 0.78 | -0.28 |
| 18 | 0.67 | 0.43 | 0.71 | -0.20 | 0.76 | -0.34 | 0.67 | 0.47 | 0.66 | -0.19 | 0.83 | -0.25 |
| 19 | 0.46 | -0.23 | 0.84 | -0.11 | 0.00 | -0.10 | 0.49 | -0.18 | 0.87 | -0.17 | 0.08 | 0.25 |
| 20 | 0.41 | 0.11 | 0.58 | 0.46 | 0.40 | 0.07 | 0.42 | 0.00 | 0.67 | 0.16 | 0.24 | -0.10 |
| 22 | 0.89 | -0.09 | 0.31 | 0.58 | 0.33 | -0.03 | 0.91 | -0.11 | 0.39 | 0.36 | 0.35 | 0.18 |
| 23 | 0.46 | 0.02 | 0.36 | 0.20 | -0.03 | 0.12 | 0.53 | -0.06 | 0.38 | 0.23 | 0.03 | 0.35 |
| 24 | 0.54 | -0.33 | 0.81 | -0.10 | 0.41 | 0.18 | 0.53 | -0.34 | 0.81 | -0.19 | 0.52 | 0.01 |
| 25 | 0.76 | -0.13 | 0.63 | -0.14 | 0.81 | -0.01 | 0.77 | -0.10 | 0.68 | -0.12 | 0.85 | -0.09 |
| 26 | 0.23 | 0.12 | -0.08 | 0.19 | 0.75 | -0.04 | 0.38 | 0.12 | -0.15 | 0.21 | 0.74 | -0.10 |
| 28 | 0.87 | -0.27 | 0.93 | -0.08 | 0.87 | 0.12 | 0.89 | -0.24 | 0.92 | -0.07 | 0.89 | 0.03 |
| 29 | 0.09 | -0.09 | 0.61 | 0.38 | 0.29 | -0.20 | 0.23 | -0.16 | 0.71 | 0.06 | 0.25 | -0.16 |
| 30 | -0.11 | -0.15 | 0.70 | 0.32 | -0.06 | -0.52 | -0.18 | -0.09 | 0.71 | 0.20 | -0.14 | -0.20 |
| 31 | 0.79 | -0.25 | 0.09 | 0.04 | 0.57 | -0.21 | 0.80 | -0.24 | 0.05 | -0.16 | 0.58 | -0.10 |
| 32 | 0.81 | -0.10 | 0.74 | 0.26 | 0.75 | 0.14 | 0.86 | -0.14 | 0.79 | 0.16 | 0.84 | 0.15 |
| 33 | 0.70 | -0.14 | 0.75 | -0.09 | 0.28 | 0.61 | 0.75 | -0.06 | 0.74 | -0.09 | 0.42 | 0.60 |
| 35 | 0.15 | 0.53 | 0.77 | 0.18 | 0.14 | -0.06 | 0.25 | 0.49 | 0.83 | 0.10 | 0.14 | 0.33 |
| 36 | 0.56 | -0.03 | 0.54 | -0.04 | 0.70 | -0.17 | 0.56 | 0.10 | 0.44 | 0.32 | 0.66 | -0.28 |
| 37 | 0.83 | 0.00 | 0.56 | 0.37 | 0.79 | -0.31 | 0.82 | -0.06 | 0.65 | 0.24 | 0.76 | -0.38 |
| 38 | 0.61 | 0.12 | 0.42 | -0.11 | 0.73 | -0.17 | 0.59 | 0.13 | 0.38 | -0.28 | 0.76 | -0.12 |
| 39 | 0.49 | 0.48 | 0.84 | 0.02 | 0.63 | 0.47 | 0.60 | 0.45 | 0.87 | -0.08 | 0.71 | 0.27 |
| 41 | 0.92 | -0.12 | 0.75 | -0.06 | 0.93 | 0.05 | 0.93 | -0.16 | 0.76 | 0.16 | 0.93 | 0.05 |
| 42 | 0.57 | 0.41 | 0.69 | -0.13 | 0.62 | 0.18 | 0.68 | 0.32 | 0.70 | -0.23 | 0.54 | 0.00 |
| 43 | 0.43 | 0.16 | 0.57 | -0.02 | 0.69 | 0.11 | 0.50 | 0.08 | 0.70 | -0.08 | 0.79 | -0.11 |

| Order | Form A | | Form B | | Form C | | Form M | | Form N | | Form O | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Factor | | Factor | | Factor | | Factor | | Factor | | Factor | |
| | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| 44 | 0.83 | 0.00 | 0.12 | -0.39 | 0.78 | -0.01 | 0.87 | 0.02 | 0.14 | -0.35 | 0.71 | -0.16 |
| 46 | 0.17 | 0.09 | 0.81 | -0.30 | 0.49 | 0.27 | 0.20 | 0.16 | 0.80 | -0.31 | 0.55 | 0.16 |
| 47 | 0.88 | -0.05 | 0.81 | -0.20 | 0.69 | -0.22 | 0.90 | -0.04 | 0.81 | -0.12 | 0.70 | -0.01 |
| 48 | -0.38 | -0.20 | 0.78 | 0.26 | 0.58 | -0.07 | -0.46 | -0.03 | 0.78 | 0.26 | 0.61 | -0.04 |
| 49 | 0.85 | -0.24 | 0.89 | 0.14 | 0.71 | 0.41 | 0.87 | -0.24 | 0.90 | 0.07 | 0.79 | 0.35 |
| 50 | 0.75 | -0.14 | 0.79 | -0.06 | 0.44 | 0.09 | 0.77 | -0.11 | 0.78 | 0.21 | 0.57 | 0.41 |
| 51 | 0.83 | 0.01 | 0.76 | -0.22 | 0.84 | -0.20 | 0.84 | 0.04 | 0.75 | -0.22 | 0.86 | -0.07 |
| 52 | 0.64 | 0.03 | 0.90 | 0.05 | 0.58 | -0.17 | 0.54 | 0.03 | 0.89 | -0.11 | 0.46 | -0.30 |
| 53 | 0.86 | -0.14 | 0.84 | 0.04 | 0.81 | -0.07 | 0.87 | -0.22 | 0.84 | 0.00 | 0.84 | -0.09 |

Grade 8 and NC Math 1

| Order | Grade 8 Math | | | | | | NC Math I | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Form M | | Form N | | Form O | | Form A | | Form M | | Form N | |
| | Factor | | Factor | | Factor | | Factor | | Factor | | Factor | |
| | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| **1** | 0.39 | 0.26 | 0.86 | -0.07 | 0.80 | -0.08 | 0.78 | 0.31 | 0.71 | 0.34 | 0.57 | 0.13 |
| **2** | 0.70 | 0.40 | 0.83 | 0.02 | 0.79 | 0.30 | 0.77 | -0.03 | 0.77 | -0.24 | 0.86 | -0.12 |
| **3** | 0.72 | -0.36 | 0.48 | -0.14 | 0.69 | -0.46 | 0.61 | -0.09 | 0.52 | 0.48 | 0.33 | 0.53 |
| **5** | 0.32 | -0.08 | 0.24 | -0.06 | 0.47 | -0.20 | 0.82 | -0.09 | 0.76 | -0.04 | 0.73 | -0.28 |
| **6** | 0.14 | 0.12 | 0.44 | -0.40 | 0.78 | -0.17 | 0.92 | -0.09 | 0.62 | -0.07 | 0.80 | -0.18 |
| **7** | 0.53 | -0.24 | -0.16 | -0.04 | -0.01 | 0.19 | 0.89 | -0.10 | 0.90 | -0.20 | 0.89 | 0.28 |
| **8** | 0.31 | 0.19 | 0.70 | -0.12 | -0.24 | 0.21 | 0.86 | -0.08 | 0.74 | -0.16 | 0.80 | 0.08 |
| **10** | 0.82 | 0.04 | 0.34 | 0.12 | 0.44 | 0.18 | 0.90 | 0.00 | 0.72 | 0.28 | 0.82 | 0.28 |
| **11** | -0.29 | -0.06 | -0.12 | -0.48 | 0.66 | -0.13 | 0.95 | -0.17 | 0.35 | -0.32 | 0.87 | 0.13 |
| **12** | 0.53 | 0.35 | 0.33 | -0.47 | 0.51 | -0.22 | 0.91 | -0.11 | 0.33 | -0.46 | 0.73 | -0.14 |
| **13** | 0.44 | -0.41 | 0.27 | -0.02 | 0.64 | -0.16 | 0.89 | 0.16 | 0.94 | -0.18 | 0.89 | -0.14 |
| **14** | 0.83 | 0.03 | 0.82 | -0.09 | 0.77 | 0.18 | 0.52 | 0.05 | 0.60 | 0.14 | 0.10 | 0.37 |
| **16** | 0.35 | -0.03 | 0.65 | -0.24 | 0.72 | -0.16 | 0.87 | -0.16 | 0.82 | 0.01 | 0.75 | -0.29 |
| **17** | 0.26 | 0.12 | 0.52 | -0.19 | 0.20 | 0.32 | 0.87 | -0.18 | 0.43 | 0.26 | 0.55 | -0.17 |
| **18** | 0.50 | -0.10 | 0.75 | 0.07 | 0.65 | -0.15 | 0.82 | 0.37 | 0.81 | -0.20 | 0.09 | -0.06 |
| **19** | 0.63 | -0.02 | 0.57 | -0.17 | 0.59 | 0.14 | 0.59 | 0.47 | 0.84 | 0.19 | 0.85 | -0.02 |
| **20** | 0.52 | 0.44 | 0.41 | -0.10 | 0.69 | 0.17 | 0.38 | -0.22 | 0.83 | -0.25 | 0.91 | -0.26 |
| **22** | 0.65 | -0.21 | 0.42 | 0.52 | 0.82 | -0.04 | 0.47 | -0.38 | 0.82 | 0.25 | 0.61 | 0.28 |
| **23** | 0.23 | 0.64 | 0.20 | 0.64 | 0.49 | 0.15 | 0.71 | -0.22 | 0.09 | 0.35 | 0.17 | 0.39 |
| **24** | 0.50 | -0.18 | 0.63 | -0.26 | 0.80 | 0.01 | 0.61 | -0.13 | 0.81 | 0.01 | 0.72 | 0.16 |
| **25** | 0.76 | -0.05 | 0.45 | -0.13 | 0.66 | -0.22 | 0.75 | -0.14 | 0.56 | -0.21 | 0.22 | -0.45 |

| Order | Grade 8 Math | | | | | | NC Math I | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Form M | | Form N | | Form O | | Form A | | Form M | | Form N | |
| | Factor | | Factor | | Factor | | Factor | | Factor | | Factor | |
| | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| **26** | 0.52 | -0.36 | 0.81 | 0.18 | 0.59 | 0.15 | -0.08 | 0.36 | 0.74 | -0.17 | 0.70 | 0.00 |
| **28** | 0.76 | -0.20 | 0.51 | 0.22 | 0.89 | 0.05 | 0.82 | -0.22 | 0.78 | -0.16 | 0.92 | -0.14 |
| **29** | 0.66 | 0.40 | 0.65 | 0.46 | 0.80 | -0.08 | 0.05 | 0.40 | 0.38 | 0.35 | -0.17 | 0.00 |
| **30** | 0.77 | 0.07 | 0.64 | 0.00 | 0.70 | -0.21 | 0.51 | 0.10 | 0.72 | 0.08 | 0.09 | 0.42 |
| **31** | 0.38 | 0.09 | 0.38 | 0.16 | -0.42 | -0.15 | 0.80 | 0.00 | 0.67 | 0.09 | 0.87 | 0.01 |
| **32** | 0.73 | -0.37 | -0.23 | 0.24 | 0.59 | 0.30 | 0.63 | -0.10 | 0.72 | 0.10 | 0.60 | -0.15 |
| **33** | 0.66 | -0.32 | 0.28 | 0.60 | 0.34 | 0.67 | 0.67 | 0.01 | 0.80 | -0.19 | 0.73 | 0.33 |
| **35** | 0.71 | -0.05 | -0.09 | 0.29 | 0.12 | 0.05 | 0.53 | -0.08 | 0.75 | 0.32 | 0.52 | 0.36 |
| **36** | 0.78 | -0.02 | 0.44 | -0.27 | 0.87 | 0.01 | 0.45 | 0.54 | 0.39 | 0.29 | 0.67 | -0.05 |
| **37** | 0.24 | 0.11 | 0.50 | 0.35 | 0.17 | 0.01 | 0.75 | -0.09 | 0.05 | 0.12 | 0.67 | 0.03 |
| **38** | 0.63 | -0.26 | 0.73 | 0.23 | 0.57 | 0.14 | 0.53 | 0.04 | 0.24 | -0.20 | 0.35 | 0.05 |
| **39** | 0.65 | 0.17 | 0.19 | 0.44 | 0.69 | -0.47 | 0.41 | 0.40 | 0.85 | 0.03 | 0.78 | 0.07 |
| **41** | 0.71 | 0.02 | 0.79 | 0.07 | 0.79 | 0.10 | 0.87 | -0.13 | 0.80 | 0.38 | 0.84 | -0.02 |
| **42** | 0.21 | 0.59 | 0.57 | -0.10 | 0.64 | -0.36 | 0.56 | 0.07 | 0.88 | -0.02 | 0.87 | -0.12 |
| **43** | -0.18 | 0.34 | 0.47 | -0.23 | 0.00 | 0.35 | 0.01 | 0.09 | 0.88 | 0.04 | 0.91 | 0.22 |
| **44** | 0.21 | 0.26 | 0.78 | -0.15 | 0.87 | 0.01 | 0.05 | 0.36 | -0.35 | 0.40 | 0.10 | 0.46 |
| **46** | 0.54 | -0.14 | 0.73 | -0.20 | 0.33 | 0.57 | -0.35 | 0.29 | 0.75 | -0.34 | 0.91 | -0.20 |
| **47** | 0.79 | -0.04 | 0.25 | 0.00 | 0.45 | 0.27 | 0.79 | -0.24 | 0.44 | 0.35 | 0.47 | 0.01 |
| **48** | 0.43 | 0.05 | 0.86 | 0.13 | 0.58 | -0.18 | 0.44 | 0.30 | 0.69 | -0.13 | 0.84 | 0.19 |
| **49** | 0.83 | 0.26 | 0.78 | -0.25 | 0.89 | -0.03 | 0.29 | -0.13 | 0.75 | -0.01 | 0.34 | 0.52 |
| **50** | 0.70 | -0.06 | 0.67 | 0.24 | 0.57 | 0.50 | 0.40 | 0.25 | 0.57 | -0.10 | 0.89 | -0.09 |
| **51** | 0.87 | -0.01 | 0.39 | -0.56 | 0.72 | -0.05 | 0.64 | 0.01 | 0.71 | -0.09 | -0.04 | 0.31 |
| **52** | 0.76 | 0.23 | 0.86 | -0.02 | 0.84 | 0.11 | 0.25 | 0.36 | 0.34 | 0.35 | 0.34 | 0.39 |
| **53** | 0.82 | -0.04 | 0.69 | 0.34 | 0.78 | 0.05 | 0.56 | 0.49 | 0.80 | -0.24 | 0.78 | -0.15 |
| **54** | | | | | | | -0.10 | 0.29 | 0.05 | 0.31 | 0.50 | -0.03 |
| **55** | | | | | | | 0.29 | 0.02 | 0.79 | -0.17 | 0.78 | -0.14 |
| **56** | | | | | | | 0.79 | 0.04 | 0.89 | 0.01 | 0.92 | -0.29 |
| **57** | | | | | | | 0.08 | 0.27 | 0.82 | 0.04 | 0.92 | -0.17 |
| **58** | | | | | | | 0.64 | 0.30 | 0.52 | 0.35 | 0.72 | 0.10 |

NC Math 3

| Item Order | Form A | | Form B | | Form M | | Form N | | Form O | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Factor | | Factor | | Factor | | Factor | | Factor | |
| | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| 1 | 0.81 | -0.32 | 0.81 | -0.42 | 0.86 | -0.13 | 0.83 | -0.29 | 0.65 | -0.08 |
| 2 | 0.94 | 0.07 | 0.88 | -0.24 | 0.86 | 0.13 | 0.87 | -0.31 | 0.53 | 0.08 |
| 3 | 0.89 | -0.07 | 0.30 | 0.15 | 0.82 | 0.06 | 0.36 | -0.04 | 0.93 | -0.02 |
| 4 | 0.70 | -0.35 | 0.11 | -0.27 | 0.43 | 0.01 | 0.71 | -0.01 | 0.82 | 0.25 |
| 5 | 0.53 | 0.21 | 0.77 | -0.07 | 0.49 | 0.63 | 0.63 | 0.24 | 0.51 | 0.21 |
| 6 | 0.66 | -0.17 | 0.43 | 0.24 | 0.95 | 0.00 | 0.83 | 0.26 | 0.43 | 0.44 |
| 7 | 0.86 | -0.11 | 0.47 | 0.06 | 0.89 | 0.31 | 0.87 | 0.09 | 0.87 | -0.21 |
| 8 | 0.69 | -0.02 | 0.45 | -0.07 | 0.81 | -0.21 | 0.76 | 0.26 | 0.90 | -0.18 |
| 9 | 0.44 | -0.08 | 0.84 | -0.03 | 0.28 | -0.12 | 0.20 | 0.54 | 0.92 | -0.20 |
| 10 | -0.07 | 0.40 | 0.85 | -0.20 | -0.10 | -0.03 | 0.32 | 0.21 | 0.93 | 0.01 |
| 11 | 0.59 | -0.06 | 0.92 | -0.10 | 0.73 | -0.26 | 0.63 | 0.47 | 0.62 | 0.13 |
| 12 | 0.83 | 0.02 | 0.84 | -0.37 | 0.89 | 0.03 | 0.80 | -0.15 | 0.80 | -0.25 |
| 13 | 0.84 | -0.19 | 0.93 | -0.06 | 0.86 | -0.23 | 0.56 | 0.43 | 0.36 | -0.19 |
| 14 | 0.92 | 0.12 | 0.01 | 0.26 | 0.81 | 0.22 | 0.42 | 0.07 | 0.59 | 0.28 |
| 15 | 0.91 | 0.06 | 0.74 | 0.07 | 0.44 | -0.25 | -0.07 | -0.10 | 0.50 | 0.21 |
| 16 | 0.68 | 0.13 | 0.80 | -0.06 | 0.46 | -0.36 | 0.92 | 0.03 | 0.66 | -0.04 |
| 17 | 0.72 | 0.26 | 0.16 | 0.35 | 0.79 | -0.21 | 0.88 | 0.07 | 0.56 | 0.33 |
| 18 | 0.29 | 0.61 | 0.56 | 0.44 | 0.14 | -0.03 | 0.50 | 0.08 | 0.40 | -0.15 |
| 19 | 0.33 | 0.48 | 0.34 | 0.05 | 0.86 | -0.26 | 0.53 | 0.09 | 0.69 | -0.13 |
| 20 | 0.87 | -0.17 | 0.65 | 0.11 | 0.95 | 0.10 | 0.89 | -0.33 | 0.66 | -0.37 |
| 21 | -0.17 | 0.48 | 0.63 | -0.06 | 0.43 | 0.05 | 0.94 | -0.14 | 0.86 | -0.10 |
| 22 | 0.12 | 0.31 | 0.85 | 0.16 | 0.86 | -0.14 | 0.88 | -0.20 | 0.83 | 0.18 |
| 23 | 0.85 | -0.08 | 0.77 | -0.32 | 0.76 | -0.02 | 0.80 | -0.10 | 0.43 | 0.05 |
| 24 | 0.62 | -0.27 | 0.85 | 0.13 | 0.75 | -0.05 | 0.71 | -0.05 | 0.20 | 0.00 |
| 25 | 0.85 | -0.06 | 0.89 | 0.24 | 0.89 | -0.02 | 0.13 | -0.08 | 0.75 | -0.19 |
| 26 | 0.94 | 0.04 | 0.70 | 0.28 | 0.73 | 0.19 | 0.63 | -0.02 | 0.63 | 0.26 |
| 27 | 0.88 | 0.30 | 0.74 | -0.32 | 0.95 | -0.04 | 0.87 | -0.15 | 0.80 | -0.25 |
| 28 | 0.93 | 0.11 | 0.44 | -0.28 | 0.96 | 0.03 | 0.47 | 0.22 | 0.53 | 0.13 |
| 29 | 0.54 | -0.23 | 0.45 | 0.25 | 0.74 | 0.42 | 0.03 | 0.09 | 0.72 | 0.02 |
| 30 | 0.72 | 0.12 | 0.71 | 0.15 | 0.71 | 0.53 | 0.93 | -0.06 | 0.79 | -0.37 |
| 31 | 0.75 | -0.21 | 0.83 | -0.14 | 0.51 | -0.03 | 0.90 | -0.08 | 0.61 | 0.18 |
| 32 | 0.39 | 0.09 | 0.73 | 0.24 | 0.85 | -0.05 | 0.80 | 0.11 | 0.10 | 0.48 |
| 33 | -0.12 | 0.20 | 0.32 | 0.46 | 0.71 | 0.20 | 0.01 | 0.54 | 0.34 | 0.20 |
| 34 | 0.58 | -0.35 | 0.61 | -0.38 | 0.93 | -0.14 | 0.00 | 0.12 | 0.87 | -0.26 |
| 35 | -0.04 | 0.46 | 0.43 | 0.21 | -0.04 | 0.14 | 0.84 | 0.11 | 0.82 | -0.04 |
| 36 | 0.75 | -0.01 | -0.26 | -0.05 | 0.29 | -0.14 | 0.91 | 0.26 | 0.47 | 0.47 |

| Item Order | Form A | | Form B | | Form M | | Form N | | Form O | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Factor | | Factor | | Factor | | Factor | | Factor | |
| | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| 37 | 0.57 | 0.39 | 0.34 | 0.62 | 0.93 | 0.12 | 0.26 | 0.33 | 0.58 | 0.42 |
| 38 | 0.45 | 0.55 | 0.61 | 0.37 | 0.34 | -0.15 | 0.24 | -0.07 | 0.58 | 0.53 |
| 39 | 0.35 | -0.08 | 0.31 | 0.33 | 0.50 | 0.23 | 0.81 | -0.01 | 0.15 | 0.13 |
| 40 | 0.72 | 0.19 | 0.41 | 0.30 | 0.10 | -0.05 | 0.69 | -0.23 | -0.35 | -0.07 |
| 41 | 0.45 | 0.14 | 0.24 | 0.54 | 0.06 | 0.03 | 0.39 | 0.44 | 0.37 | 0.26 |
| 42 | 0.45 | -0.06 | 0.25 | 0.40 | -0.32 | 0.18 | 0.76 | -0.02 | 0.55 | -0.28 |
| 43 | 0.49 | 0.09 | 0.35 | 0.32 | 0.84 | -0.24 | 0.02 | 0.54 | 0.56 | 0.40 |
| 44 | 0.43 | 0.35 | 0.16 | 0.18 | 0.62 | -0.38 | 0.77 | -0.08 | 0.34 | -0.12 |
| 45 | 0.40 | 0.21 | 0.92 | -0.21 | 0.89 | -0.02 | 0.89 | -0.15 | 0.59 | -0.06 |
| 46 | 0.11 | 0.24 | 0.25 | 0.45 | 0.52 | -0.38 | 0.78 | -0.09 | 0.63 | 0.06 |
| 47 | 0.77 | 0.15 | 0.83 | -0.32 | 0.80 | 0.24 | 0.06 | 0.41 | 0.92 | -0.06 |
| 48 | 0.32 | -0.13 | 0.92 | -0.08 | 0.59 | 0.06 | 0.74 | 0.02 | 0.48 | -0.17 |
| 49 | 0.59 | -0.43 | 0.91 | -0.18 | 0.88 | -0.06 | 0.87 | -0.07 | 0.77 | -0.33 |
| 50 | 0.93 | -0.18 | 0.93 | 0.00 | 0.94 | 0.13 | 0.85 | -0.27 | 0.16 | -0.01 |

## Appendix 9-B

## North Carolina Quantile Linking Report by MetaMetrics

https://www.dpi.nc.gov/districts-schools/testing-and-school-accountability/technical-information-state-tests#white-papers-and-technical-resources