The North Carolina Department of Public Instruction
Grades 3–8 Reading End-of-Grade (EOG)
High School English II End-of-Course (EOC)
Edition 5 Technical Report
2020–21

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

Prepared by:

North Carolina Department of Public Instruction
Accountability Division
2022
Tammy Howard, Ph.D., Director of Accountability
Kinge Mbella, Ph.D., Lead Psychometrician
Thakur Karkee, Ph. D., Psychometrician
Maxey-Moore, Section Chief, Test Development

Table of Contents

Table of Tables

Table of Figures

CHAPTER 1 INTRODUCTION

The intent of this technical report is to provide comprehensive and detailed description of steps implemented towards the development, analysis, and reporting of test scores from the North Carolina Annual Testing Program (NCATP). Technical evidence presented throughout this report also serve as primary sources of validity to support intended test score uses and interpretation. The validity evidence is documented in terms of processes used in review, revision, and implementation of new content standards; development of test specifications and items; field-test and item analysis; bias and sensitivity review; test development; scoring and scale development; and standard setting.

The first part of this report presents a brief overview of the revision and eventual adoption of new grades 3–8 Reading and high school English II content standards which are bases for the development of new assessments. The remaining sections describe a brief history of the NCATP 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 evidence for the Edition 5 grades 3–8 End-of-Grade (EOG) Reading and high school End-of-Course (EOC) English II summative assessments.

The North Carolina Department of Public Instruction (NCDPI) recommends interpreting 2020–21 summary results cautiously as circumstances of the school year were affected by COVID-19 pandemic. First, COVID-19 related disruptions to normal learning and school environments lead to varied instructional practices across public schools systems in the state ranging from several models of in-person, virtual, to hybrid instruction. Second, in the 2020–21 school year the United States Department of Education and State of North Carolina waived accountability, which implied the high stakes consequences usually attributed to test scores did not apply in 2020–21. Finally, the accountability waiver also applied to the 95% participation requirement. Even though participation rate for state assessments in 2020–21 were close to the expected 95%, participation rates across districts and subgroups varied and there is no direct evidence that the missingness was random. As a result of these circumstances, caution is advised when attempting to compare student performance from 2020–21 with other years.

1.1 Purpose and Background of the North Carolina Annual Testing Program

The General Assembly GCS 115C-174.10T specified the purpose of the NCATP 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 that *"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 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*).

Historically, EOG assessments were designed and administered for the first time in 1994 to measure the NCSBE-adopted content standards 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 in 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 the English Language Arts/Reading (ELA) 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 3–8 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 2017, under the leadership from the NCSBE, the *Common Core State Standards* for reading were replaced by new NC*SCoS*. To maintain strong content alignment and validity evidence of uses and score interpretations, EOG and EOC assessments were redesigned. As a result, item development, field-test, and form development process with new items aligned to the new

NC*SCoS* was planned in 2018–19 administration with subsequent operational administration in 2019–20. The operational forms for English II were administered in Fall 2019. However, due to the COVID-19 pandemic Spring 2020 assessments were waived. The operational administration resumed in 2020–21. Subsequently, a standard setting meeting and approval of the standards from the NCSBE occurred in the summer of 2021.

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 grades 3–8 Reading and EOC English II assessments. 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 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 a comprehensive content standard review process in 2015. The findings and recommendations of ASRC's reviews are documented in the commission's report (*Appendix 1-B*). In early 2016, Division of Academic Standards 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 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 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 Division of Academic Standards developed and implemented a plan of action and timeline in 2016 to review and revise the English II and Reading content standards. During the review phase, Division of Academic Standards worked with the ASRC as facilitator to help with research and provided guidance on state and federal policy requirements. The role of the Division of Academic Standards 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's progress throughout the process.

Following the review, Division of Academic Standards adopted a 6–step iterative process summarized below to revise and draft new English II and Reading 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 PSU in the areas of instruction and assessment, Division of Academic Standards 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.

Table 1.1 outlines 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 English II and Reading from 2016 through 2018. 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 English II and Reading standards. The data review and writing committee consisted of educators including Institution of Higher Education (IHE), teachers, coaches, administrators, as well as business partners.

Table 1.1 English II and Reading Standards Review, Revision and Implementation

Date	Actions	Descriptions
June 2015	Educator English II and Reading Survey	The commission surveyed approximately 5,000 educators. Similarly, 283 educators provided feedback. The commission also conducted 36 professional development (PD) training sessions and four webinars. Feedback from the survey and PD training collected information regarding English II and Reading resources and needs.
December 2015	The North Carolina Academic Standards Review Commission Reporting Findings and Recommendations	The commission reviewed other states' standards and identified research-based practices. Some of the findings included: - The English II and Reading standards were poorly distributed across grades. - Need of developmentally appropriate practices. - Absence of comprehensive writing instruction. - Suggestion for teaching rich historical literature.
June 2016	English II and Reading Data review committee reviewed responses	This group reviewed data from surveys, focus groups, and ASRC report to determine patterns and concerns.
July 2016	English II and Reading Data Review Committee findings are compiled	Findings were compiled and shared with leadership and NCSBE.
July–November 2016	English II and Reading Writing Teams meet	The Writing Teams act on the findings of the data review committee by working virtually and face-to-face.
December 2016	Share English II and Reading drafts	The team shared drafts of English II and Reading standards with NCSBE and leadership.

Date	Actions	Descriptions
January 2017	Release drafts for public comment	The team shared drafts with the public for comments.
February 2017	Writing Teams review feedback	The team sorted and reviewed comments.
March–April 2017	Present comments and English II and Reading drafts to NCSBE	Presented comments and draft of English II and Reading standards to NCSBE for discussion and actions.
May 2017	Create English II and Reading professional development and resources	Created professional development and resources to support the revisions and needs as reflected by the surveys and comments.
June–August 2017	Regional PD sessions	Conducted regional professional development.
August 2017	Implement Standards	Districts implemented new standards and continued support to schools.
August 2017–May 2018	Standards implementation preparation	• Professional development. • Develop resources and revision of all support materials. • Test specification workshop.
June–August 2018	Summer professional development delivery	PD webinars were conducted.
2018–19	New standards implemented	Items based on new standards developed and field tested by embedding in the operational forms.
2019–20	English II operational administration.	English II operational forms aligned to the new NC*SCoS* administered in Fall 2019.
2020–21	Standard setting and score reporting	• Standard setting conducted in July of 2021 for the *Edition 5* EOG Reading and EOC English II. Raw-to-scale tables developed based on item parameters from 2018–19 field test administration for grades 3–8 Reading and Fall 2019 item parameters for English II. • Score reported on new achievement level scale.

The attributes described above are a part of validity evidence to show that North Carolina English II and Reading 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 evidence of EOG and EOC assessment score uses and interpretation, North Carolina redesigned and administered new assessments that are aligned to the new NC*SCoS* adopted for English II and Reading.

1.3 Overview of the North Carolina Annual Testing Program

The NCDPI designs, develops and administers high-quality statewide reading 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* (*Appendix 1-C*) 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 evidence 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 specification meetings, test blueprints, cognitive complexity, item format and mode of test administration. An overview of the item development process, which includes item writer training, item writing, and reviews, is also documented. The final section describes field-test plans to replenish item pools 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 screening out items with less-than-optimal characteristics. Final sections describe summary of item analysis and calibration of item responses for the purpose of estimating item parameters and 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 evidence 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 procedure used for scoring and scale development 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 2021 after the first operational administration of EOGs and EOCs. Item parameters from the 2018–19 pre-pandemic field-test administration were used for the standard setting. 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 EOG grades 3–8 Reading and EOC

English II 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 2020–21 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 evidence collected in support of the interpretation of EOG and EOC test scores. The first two sections in this chapter present validity evidence 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 evidence 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 1.2 lists glossary of abbreviations used throughout this document.

Table 1.2 Glossary of Abbreviations

Abbreviations	Full Form
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

Abbreviations	Full Form
EDS	Economically Disadvantaged Students
EL	English Learner
ELA	English Language Arts/Reading
EOC	End-of-Course
EOG	End-of-Grade
FERPA	Family Educational Rights and Privacy Act
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
NC	North Carolina
NCDPI	North Carolina Department of Public Instruction
NCLB	No Child Left Behind
NCSBE	North Carolina State Board of Education
NC*SCoS*	North Carolina Standard Course of Study
NCATP	North Carolina Annual Testing Program
NCSU-TOPS	North Carolina State University-Technical Outreach for Public Schools
NCTAC	North Carolina Technical Advisory Committee
OTISS	Online Testing Irregularity Submission System
PBT	Paper-Based Test
PCA	Principle Component Analysis
PII	Personally Identifiable Information
RAC	Regional Accountability Coordinator
SE	Standard Error
SR	String Replace
TCC	Test Characteristic Curve
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* EOG Reading and EOC English II 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 are standards-based assessments that serve summative purposes. These assessments were redesigned to align with new grades 3–8 Reading and English II content standards adopted in 2017 to ensure adequate validity evidence 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 demand, and psychometric specifications.

The NCDPI recruited North Carolina educators from across the state and conducted an on-site test specification workshop in February 2018. Participants invited to this meeting represented North Carolina educators 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 reading assessments.

2.1.1 Content Blueprint

The main goal of the test specification workshop facilitated by the NCDPI Test Development staff was to get participants to recommend content blueprints for *Edition 5*. The workshops were held by grade spans: grades 3–5, 6–8, and 9–12. During these interactive 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 English II and Reading 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 workshop, the NCDPI team members from Test Development and Division of Academic Standards reviewed the recommended blueprints to ensure adequate across-grades articulation. The final recommendations shown in *Table 2.1* were then adopted as *Edition 5* content blueprints for EOG Reading and EOC English II assessments.

Table 2.1 Grades 3–8 Reading and English II Test Blueprints

Domain	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	English II
Reading for Literature	38–42%	38–42%	38–42%	36–41%	36–41%	36–41%	35–39%
Reading for Informational Text	46–50%	46–50%	46–50%	43–47%	43–47%	43–47%	42–46%
Language	13–15%	13–15%	13–15%	11–16%	11–16%	11–16%	9–13%
Total	100%	100%	100%	100%	100%	100%	100%

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 Division of Academic Standards staff received training on Webb's DOK classifications in 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 reading content standards. At the conclusion of the meeting, the NCDPI's Test Development and Division of Academic Standards 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 tests are shown in *Table 2.2*.

Table 2.2 Proposed DOKs (%) Across Grades/Courses

Grade	Number of Items	Category (%)		
		DOK 1	DOK 2	DOK 3
3	40	20–40	60–80	
4	40	12–25	50–75	5–10
5	40	—	75–90	10–25
6	44	—	60–82	18–40
7	44	—	60–82	18–40
8	44	—	60–82	18–40
English II	51	—	60–75	25–40

2.1.3 Item Format

For EOC English II, three main item types are used for the computer-based fixed forms: four-foil multiple-choice, two types of technology-enhanced (TE) items (computer-based form only), and a short constructed-response (CR). The two types of TE items included Text Identify (TI) and String Replace (SR). A TI item consisted of a stem and multiple options. Students are instructed to read the stem, then identify the correct text provided by clicking on all correct options. A SR item consisted of a short text that has one word highlighted ("hot text") and a list of four possible

replacement words. The task is to select a response option by clicking or hovering using the mouse pointer over any choice from the list provided. This action replaces the "hot-text" in the reading selection. For grades 3–8 EOG Reading, all items are four-foil multiple-choice in both paper/pencil and computer-based forms. A usability study of the TE item types was conducted in *Edition 4* of the test and the item types are a continuation to *Edition 5*.

The EOC English II CR items are short written responses that typically about a couple of sentences to a paragraph. These short response items are scored on a 3-point scale of 0–2 points each. The majority of students participating in online administration have a 1,000-character limit for their responses. Students participating in paper accommodation administration are given a text box with lines in the answer sheet to write their responses. Students must not write beyond the end of the line or in the margins. Words written in the margins or unlined areas of the answer sheet are not scored. Students are instructed to not add additional lines to the answer sheet. Words written on extra lines are not scored. Scoring rubric is limited to the specific criteria as stated in the item. Students are not penalized for grammar or rewarded for providing additional information.

North Carolina has a long tradition of instant score reporting upon completion of test. This with the pre-equating model used for scoring are important consideration in determining new item types to use for state EOG and EOC assessments.

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 allowing them time to acquire 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 their mode of administration. Beginning from 2018–19 administration, the NCDPI requires EOC English II assessment to be administered in CBT mode with the PBT mode available to only students and schools with documented special accommodation needs.

From 2021–22 school year, EOG reading grades 6–8 and EOC English II were required to administer in CBT mode, with accommodated paper forms for students and schools who cannot access a computer. However, schools were offered flexibility for the requirement to choose the best mode that fits their students' needs as they dealt with COVID-19 related disruptions. Despite this, a record of about 84% of all EOG assessments in 2020–21 were administered in CBT across grades 3–8. From 2022–23 school year, all EOG and EOC tests are required to be administered in CBT mode.

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. Educators 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 NCATP based on the North Carolina *Standard Course of Study*.

These courses provide an overview of the test development process as well as the basic rules and structures of item formats used by the NCATP. 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 sequential, requiring the online participant to step through each module in a structured sequence. At the end of most modules, participants are required to take a short quiz before moving to the next. 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 web link below: 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 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 NCATP, visit the website: https://center.ncsu.edu/ncpd/course/view.php?id=128.

2.4 Item Development Process

The item development, field-test, and form building process for *Edition 5* began after the NCSBE adopted the new NC*SCoS*. The item development and field-test for the newly aligned content standards occurred in 2018–19 and Forms were developed in 2019–20. North Carolina assessment items are written and reviewed by trained North Carolina teachers who served as item writers. Additionally, the NCDPI's Academic Standards and Test Development content experts in partnership with content specialists at North Carolina State University Technical Outreach for Public Schools (NCSU-TOPS) review all items before they are field-tested. The NCDPI's TMSs served 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 as item writers is evidence of instructional validity pertaining to how well test items align to standards and classroom curriculum.

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 for copyright concerns and proper citation. At Step 4, each item is independently reviewed by two North Carolina teachers or educators. 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).

Steps 7–8 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 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 may be unfamiliar due to cultural or socioeconomic reasons. Review of the reading level of the item is considered along with stem and foil options for multiple-choice. Items are also reviewed to ensure the stem is a clear and complete question, the foils are straightforward, there are no repetitive words, and the grammar of the stem agrees with the foils. These reviews also include 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 students. Any items with comments that cannot be reconciled are 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 for graphic (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), Division of Academic Standards 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 checks 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 Division of Academic Standards to review the item again; the TMS moves the item back to Step 14.
- Delete the item.

Item development and review are ongoing year-round to continuously replenish the item pool. Final approved items are then embedded and field tested and must undergo a post-field-test round of statistical reviews before they are placed on operational forms.

2.5 Field-Test Plan

An embedded field-test design was adopted for the development of *Edition 5* EOC English II and EOG Reading items for the North Carolina summative assessments. The main purpose to field testing items prior to the development of new operational forms is to gather reliable item level metadata to evaluate all aspects of item statistical characteristics, accessibility, fairness, and to provide baseline statistical targets to assemble pre-equated parallel forms. With the adoption of new content standards, the use of standalone 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 standalone 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 standalone field-test by reducing the test burden on students. A standalone 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 on 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 and Item Embedding Plan

The field-test plan for the EOG and EOC *Edition 5* assessments was to create sufficient item pools aligned to new NC*SCoS* in the operational forms. Specifically, the goal was to create grade specific item bank sufficient to develop at least two new parallel operational pre-equated test forms. A matrix sampling design that included eight field-test embedding slots for grades 3–8 Reading and 15 for English II within *Edition 4* forms was used in 2018–19 to create sub-versions called "flavors" to embed new NC*SCoS* aligned field-test items.

The rationale to embed new items aligned to *Edition 5* content standards within *Edition 4* operational tests was because changes on the overall assessed content standards from *Edition 4* to *Edition 5* were minimal. *Table* 2.3 shows the embedded field-test plan used to generate item pools for the new EOG and EOC assessments aligned to the new content standards.

Table 2.3 Grades 3–8 Reading and English II Embedded Field-Test Plan, 2018–19

Grades	Base Forms	Items Base Operational Form	Flavors	FT Items/ Flavor	Total FT Items
3	2	44	32	8	512
4–5	3	44	14	8	336
6–8	3	48	14	8	336
English II	2	53	18	15	540

CHAPTER 3 ITEM ANALYSIS

This chapter summarizes procedures and criteria the NCDPI used to analyze and evaluate the statistical and psychometric characteristic of new test items. Item analysis serves as the final quantitative process for item review and to establish grade level operational item pools 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 statistics for potential differential item functioning (DIF). 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 item 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 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 defined as optimal. These items are only included in the final form assembly item 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 and bias review to ensure the content is accurate and the item is free from potential bias for all student sub-groups.
- Weak: These are items with at least one major statistical parameter being significantly outside the range to be considered as optimal 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 analyses 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 generate CTT item level summary 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-values for dichotomously scored items range 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. The 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 ≤ 0.850	Keep
$0.100 \leq$ p-value ≤ 0.149 or $0.851 \leq$ p-value ≤ 0.900	Reserve
p-value ≤ 0.099 or p-value ≥ 0.901	Weak
biserial ≥ 0.250	Keep
$0.150 \leq$ biserial ≤ 0.249	Reserve
biserial < 0.150	Weak

Grades 3–8 Reading and English II CTT Descriptive Summaries

The CTT descriptive summary from field-test in 2018–19 for EOG Reading and EOC English II 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 Reading and English II Field-Test Item Pool, Spring 2019

Grade	CTT Flag	Total Items	P–value				Biserial Correlation			
			Mean	SD	Min	Max	Mean	SD	Min	Max
3	Keep	416	0.65	0.13	0.32	0.85	0.65	0.13	0.32	0.85
	Reserve	68	0.64	0.18	0.22	0.90	0.64	0.18	0.22	0.90
	Weak	28	0.48	0.14	0.22	0.66	0.48	0.14	0.22	0.66
4	Keep	226	0.68	0.12	0.28	0.85	0.46	0.08	0.20	0.62
	Reserve	93	0.67	0.21	0.21	0.90	0.35	0.12	0.16	0.58
	Weak	31	0.63	0.19	0.33	0.93	0.25	0.13	0.05	0.50
5	Keep	234	0.67	0.13	0.31	0.85	0.47	0.08	0.21	0.66
	Reserve	60	0.66	0.19	0.23	0.90	0.37	0.10	0.13	0.57
	Weak	56	0.49	0.23	0.21	0.94	0.20	0.14	-0.03	0.53
6	Keep	237	0.62	0.14	0.25	0.84	0.44	0.09	0.19	0.62
	Reserve	57	0.60	0.17	0.25	0.89	0.35	0.11	0.15	0.59
	Weak	42	0.50	0.20	0.12	0.93	0.19	0.14	-0.17	0.46
7	Keep	197	0.68	0.12	0.29	0.85	0.47	0.08	0.21	0.63
	Reserve	69	0.68	0.18	0.18	0.90	0.36	0.11	0.12	0.57
	Weak	70	0.56	0.20	0.14	0.94	0.22	0.14	-0.19	0.49
8	Keep	189	0.64	0.14	0.28	0.85	0.42	0.08	0.19	0.57
	Reserve	88	0.56	0.23	0.20	0.90	0.29	0.11	0.11	0.53
	Weak	73	0.53	0.18	0.17	0.96	0.19	0.11	-0.03	0.44
English II	Keep	325	0.61	0.14	0.18	0.84	0.42	0.08	0.19	0.59
	Reserve	94	0.56	0.17	0.16	0.90	0.31	0.08	0.13	0.49
	Weak	84	0.47	0.17	0.12	0.91	0.18	0.10	-0.09	0.40

The initial CTT results from field-test indicated that about 81% items in grade 3, 65% in grade 4, 67% grade 5, 71% grade 6, 59% grade 7, 54% grade 8, and 65% in English II were classified as meeting the NCDPI optimal standards of "Keep". The CTT flags along with p-value and biserial ranges show the item pool had enough range of item difficulty and biserial correlation for high quality operational form assembly for three forms in grades 3, 5, 6, and English II; and two forms in grades 4, 7, and 8.

3.3 IRT-Based Item Analysis

IRT offers a more robust approach to item analysis compared to CTT. CTT uses 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).

Moreover, 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 Reading items are binary scored (only two possible outcomes: correct or incorrect) and EOC English II contains TE and CR items, 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 two IRT models included three-parameter logistic (3PL) model for multiple-choice and technology enhanced items and the Graded Response Model (GRM) for constructed response items. These models make three general assumptions:

- unidimensionality – that there is one dominant latent trait being measured by the grade level tests 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 (Birnbaum, 1968) is:

$$P_i(\theta) = c_i + \frac{1-c_i}{1+\exp\left[-Da_i(\theta-b_i)\right]} \qquad 3\text{–}1$$

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 1.702. The major difference between a 3PL model and a GRM model is that the GRM model does not directly account for a chance-score parameter. The GRM assumes that the categories to which

an individual responds can be ordered or placed on a hierarchy, for example, with probabilistic scales for summation estimates or Likert-type scales. The GRM attempts to gather more information than a scale with a dichotomous response (e.g., "yes" or "no"). In this case, the GRM model can be considered an extension of the two-parameter logistic model (2PL). The mathematical function for the GRM (Samejima, 1969) model is:

$$P_{ik}^{*}(\theta_j) = \frac{e^{D\alpha i(\theta_j - \beta_{ik})}}{1+e^{D\alpha i(\theta_j - \beta_{ik})}} \qquad 3\text{–}2$$

$$P_{ik}(\theta_j) = P_{ik}^{*}(\theta_j) - P_{ik+1}^{*}(\theta_j) \qquad 3\text{–}3$$

where k is the ordered response option; $P_{ik}(\theta_j)$ is the probability of responding with option k of item i with a latent trait level θ_j; $P^*_{ik}(\theta_j)$ is the probability of responding to option k or above of item i with a latent trait level θ_j; θ_j is the latent trait level of the participant; β_{ik} is the localization parameter of alternative k of item i; α_i is the discrimination parameter of item i; and D is the constant 1.702.

All item types from field test administration were calibrated concurrently in IRTPRO (Cai et al., 2011). 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 guessing for each item can be made conditioned on ability levels. 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 expected to be 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). For constructed response items calibrated using the GRM each item is model using three nonlinear probability functions for each of the three score scales 0–2.

For final item quality, the NCDPI uses IRT parameters flagging criteria displayed in *Table 3.3* 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.

Figure 3.1 Graphical Representation of Item Characteristic Curve or Trace Line

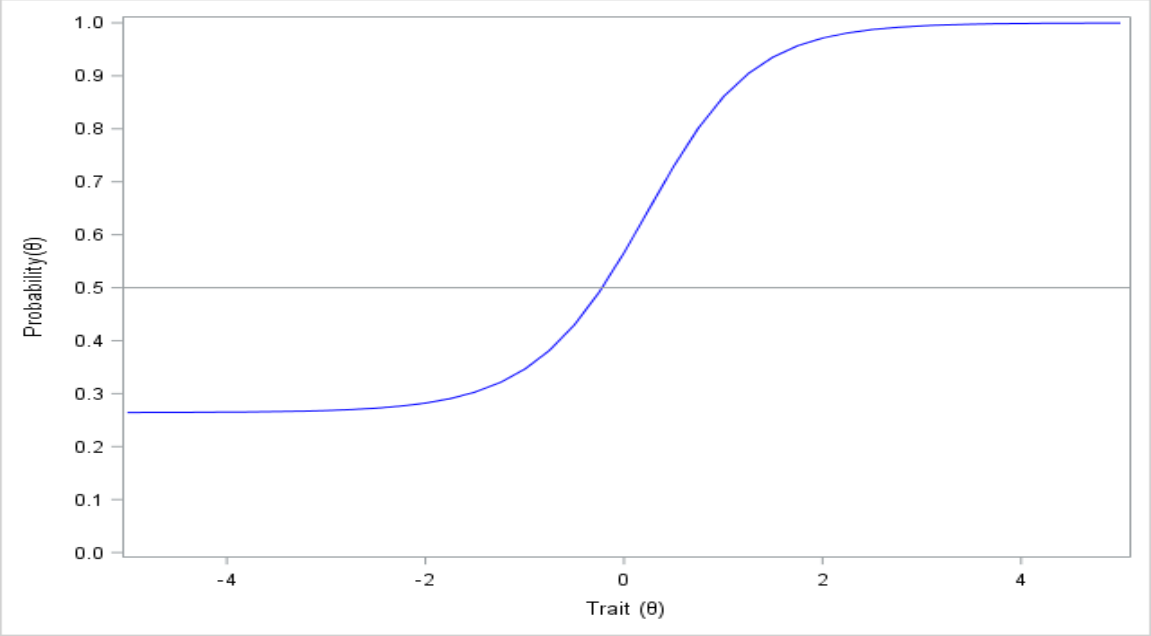


Table 3.3 IRT Parameters Threshold and Flagging Criteria

IRT Parameters Threshold	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 ≤ a ≤1.189	Reserve
a < 0.848	Weak
Asymptote or Guessing Value (c)	
≤ 0.350	Keep
0.351 ≤ c ≤ 0.450	Reserve
> 0.451	Weak

3.4 IRT Parameter Estimation

IRT parameters of the embedded field-test items are estimated by concurrent calibration of all item 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 serve two main purposes:

- Form Development: The first purpose of calibration is to develop an item bank of items with known statistical properties that are on the same latent IRT grade-level ability scale. Calibrating these items on the same IRT scale offers the NCDPI the flexibility to build multiple equivalent 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 pre-equated equivalent 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 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 each 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 field test data administered in an operational setting. 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). For 2018–19 EOG reading, the NCDPI made the decision that all grades will be calibrated using a single group design with no mode DIF.

Students response across both mode was combined and processed together. The rationale for this was supported by evidence from mode analysis performed on the previous edition of EOG reading that showed no mode DIF. IRT field-test item parameters separately calibrated across different base forms are assumed to be on a common IRT latent ability scale. The rationale is that base forms randomly spiraled and administered to representative samples of grade level population are equivalent.

Figure 3.2 Matrix Data collection For Embedded Field–Test Design

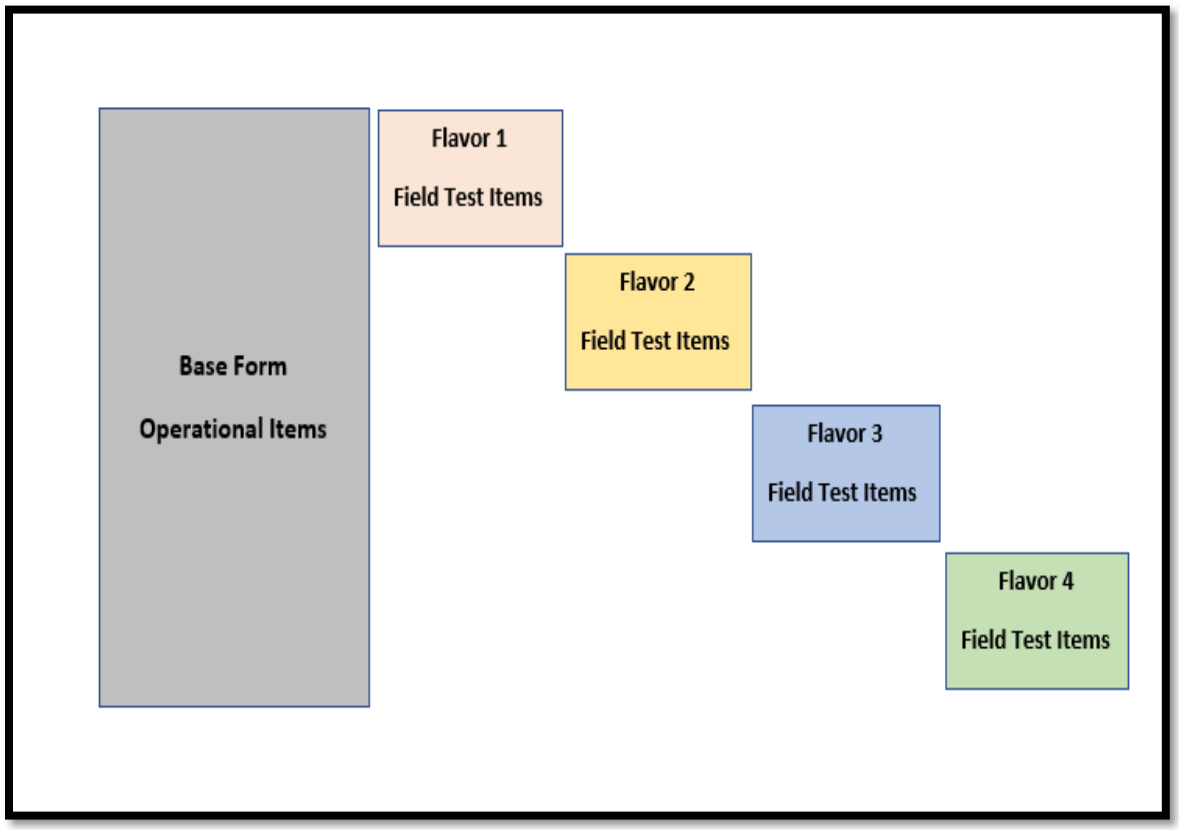


Table 3.4 and *Table* 3.5 show demographic distribution of the samples in grades 3–5 Reading, and grades 6–8 Reading and English II respectively. They show that the sample sizes, gender, and ethnic distribution across forms are very similar within each grade. In 2018–19, the grades 3–8 Reading forms were administered in both modes, and English II forms were administered in CBT mode with paper forms for accommodation only.

Table 3.4 Demographic distribution of the Field–Test Sample, Grades 3–5 Reading, 2018–19

Grade	Form	Total	Gender		Ethnicity				Other		
			F	M	B	H	W	Other	SWD	EDS	ELs
3	N	60,986	48.9	51.1	25.8	19.1	45.2	10.0	11.9	50.2	10.2
	O	60,389	48.6	51.4	25.9	19.3	45.0	9.9	12.1	50.1	10.5

	All	121,375	48.8	51.2	25.8	19.2	45.1	9.9	12.0	50.2	10.4
4	M	43,998	48.5	51.5	27.2	19.4	44.5	8.9	12.6	50.0	10.5
	N	40,736	48.8	51.3	25.4	19.0	45.7	9.9	12.6	50.1	10.1
	O	40,977	48.5	51.5	26.2	19.1	44.8	9.9	12.5	50.4	10.2
	All	125,711	48.6	51.4	26.3	19.2	45.0	9.6	12.6	50.1	10.2
5	M	42,559	48.9	51.1	26.1	19.0	45.3	9.5	12.6	49.7	9.0
	N	42,193	49.0	51.0	25.9	19.1	45.6	9.4	12.5	49.6	8.9
	O	41,801	48.8	51.2	25.8	19.3	45.6	9.4	12.1	49.3	9.2
	All	126,553	48.9	51.1	25.9	19.1	45.5	9.4	12.4	49.5	9.0

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

Table 3.5 Demographic distribution of the Field–Test Sample, Grades 6–8 Reading and English II, 2018–19

Grade /Course	Form	Total	Gender		Ethnicity				Other		
			F	M	B	H	W	Other	SWD	EDS	ELs
6	M	41,364	48.6	51.4	25.6	19.2	46.1	9.2	12.1	49.8	5.0
	N	42,915	48.7	51.3	25.5	19.5	45.6	9.4	12.1	49.9	4.7
	P	42,194	48.6	51.4	25.8	19.2	45.6	9.4	12.3	49.6	4.8
	All	126,473	48.6	51.4	25.6	19.3	45.8	9.3	12.2	49.8	4.8
7	M	41,563	48.6	51.4	26.1	18.7	46.1	9.1	12.2	48.4	3.8
	N	40,627	48.9	51.1	25.5	18.7	46.5	9.3	12.4	48.4	3.9
	O	40,875	48.9	51.1	25.3	18.8	46.8	9.2	12.3	47.8	3.7
	All	123,065	48.8	51.2	25.6	18.7	46.4	9.2	12.3	48.2	3.8

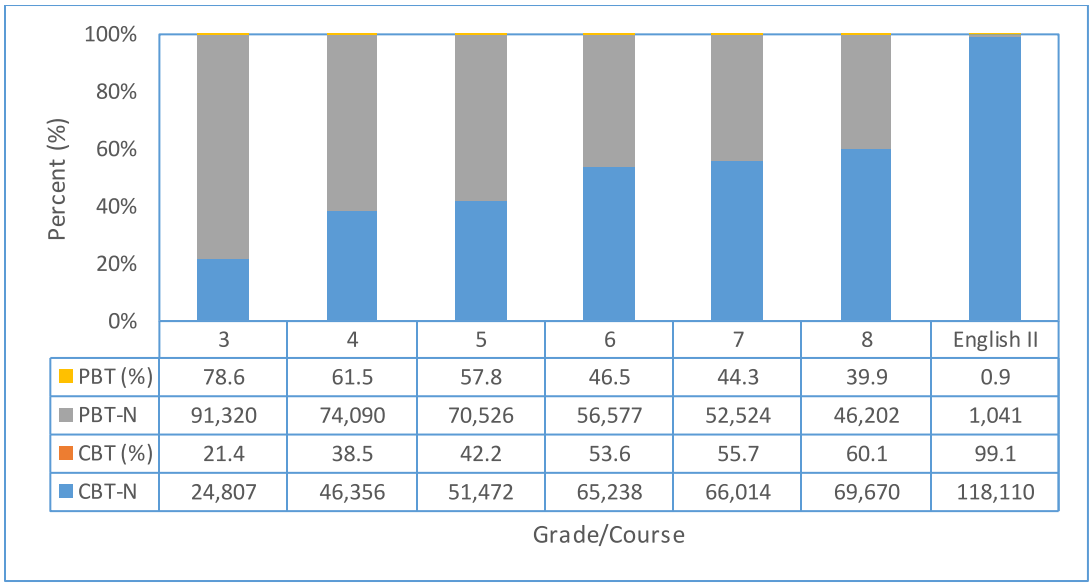
8	M	23,564	49.0	51.0	24.6	18.0	47.9	9.6	11.5	44.1	3.7
	N	23,924	48.6	51.5	24.9	18.2	47.3	9.6	11.6	44.4	3.5
	O	24,848	48.3	51.7	25.1	18.5	46.9	9.5	11.9	45.7	3.5
	All	72,336	48.6	51.4	24.8	18.2	47.3	9.6	11.7	44.7	3.6
English II	M	60,433	48.9	51.1	24.7	16.9	49.6	8.9	10.5	42.8	3.9
	O	61,110	48.8	51.2	25.1	16.8	49.4	8.6	10.7	43.4	4.0
	All	121,543	48.8	51.2	24.9	16.9	49.5	8.7	10.6	43.1	4.0

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

Figure 3.3 shows the proportion of students by mode with student count in the inset table for EOG grades 3–8 Reading and EOC English II assessments in 2018–19. Notice that the proportion of students who took the test on CBT mode increased as grade level increased. This is consistent with NCDPI's gradual transition to online mode starting with high school to middle grades and finally elementary grades. The concurrent calibration method assumed that the computer-based and paper-based duplicate test forms are equivalent and no mode DIF existed.

For English II, over 99% students completed the test in CBT mode. Therefore, only student responses from CBT mode were used in calibration for parameters estimation.

Figure 3.3 Proportion of Students by Mode, 2018–19



	3	4	5	6	7	8	English II
PBT (%)	78.6	61.5	57.8	46.5	44.3	39.9	0.9
PBT-N	91,320	74,090	70,526	56,577	52,524	46,202	1,041
CBT (%)	21.4	38.5	42.2	53.6	55.7	60.1	99.1
CBT-N	24,807	46,356	51,472	65,238	66,014	69,670	118,110

3.5 IRT Calibration Summary from 2018–19 Administration

Table 3.6 and *Table 3.7* show descriptive statistics of IRT parameters for 2018–19 embedded field-test items. The items flagged as "Keep" and "Reserve" are considered as acceptable and made up the final item pool for form assembly.

Table 3.6 Descriptive Statistics of IRT Parameters for the Grades 3–5 Reading Field-Test Items, 2018–19

Grade	Flags	N	%	Slope(a)				Threshold(b)				Asymptote(g)			
				Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
3	Keep	416	81	2.01	0.53	1.19	3.79	-0.16	0.59	-1.24	1.41	0.20	0.05	0.08	0.32
	Reserve	68	13	1.62	0.80	0.86	3.59	-0.13	1.00	-1.41	2.22	0.18	0.06	0.09	0.35
	Weak	28	5	0.94	0.91	0.36	3.67	1.23	2.03	-0.57	9.62	0.17	0.04	0.11	0.28
4	Keep	226	65	1.95	0.48	1.23	3.96	-0.32	0.58	-1.37	1.62	0.20	0.05	0.09	0.35
	Reserve	93	27	1.49	0.65	0.85	3.58	-0.31	1.16	-1.74	1.76	0.18	0.05	0.09	0.35
	Weak	31	9	1.00	0.78	0.13	2.82	0.22	2.04	-1.74	6.50	0.17	0.05	0.10	0.32
5	Keep	234	67	2.01	0.54	1.20	3.89	-0.34	0.60	-1.40	1.31	0.19	0.05	0.08	0.33
	Reserve	60	17	1.41	0.63	0.86	3.19	-0.38	0.98	-1.61	2.11	0.16	0.04	0.08	0.30
	Weak	56	16	1.42	0.97	0.09	2.93	1.10	2.30	-2.17	11.14	0.18	0.04	0.09	0.28

Table 3.7 Descriptive Statistics of IRT Parameters for the Grades 6–8 Reading and English II Field-Test Items, 2018–19

Grade	Flags	N	%	Slope(a)				Threshold(b)				Asymptote(g)			
				Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
6	Keep	237	71	1.89	0.48	1.20	3.50	-0.06	0.69	-1.50	1.72	0.20	0.05	0.08	0.33
	Reserve	57	17	1.41	0.68	0.87	3.21	0.06	0.96	-1.42	2.37	0.19	0.07	0.08	0.38
	Weak	42	13	0.81	0.56	0.22	2.80	0.71	1.73	-1.64	5.18	0.17	0.05	0.11	0.30
7	Keep	197	59	1.98	0.56	1.20	3.91	-0.32	0.59	-1.31	1.40	0.20	0.05	0.07	0.31
	Reserve	69	21	1.47	0.74	0.85	3.34	-0.41	1.03	-1.79	2.22	0.18	0.06	0.09	0.36
	Weak	70	21	0.96	0.76	0.30	3.76	0.27	1.61	-1.97	6.33	0.17	0.06	0.05	0.36
8	Keep	189	54	1.83	0.45	1.20	3.54	-0.14	0.67	-1.39	1.39	0.20	0.06	0.10	0.35
	Reserve	88	25	1.39	0.51	0.86	3.07	0.23	1.31	-1.59	2.84	0.18	0.07	0.07	0.39
	Weak	73	21	0.77	0.56	0.05	3.15	0.84	1.84	-1.94	4.44	0.18	0.06	0.07	0.34
English II	Keep	325	65	1.87	0.48	1.19	3.45	0.04	0.68	-1.36	1.58	0.22	0.07	0.08	0.43
	Reserve	94	19	1.22	0.50	0.86	3.64	0.20	0.99	-1.80	2.17	0.18	0.07	0.07	0.37
	Weak	81	16	0.98	1.16	-3.50	4.70	0.62	2.45	-10.23	8.64	0.19	0.09	0.08	0.51

3.6 Bias and Sensitivity 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 procedures 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 backgrounds 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 3.8*) is computed using the Cochran-Mantel-Haenszel (CMH) option in PROC FREQ Procedure in SAS® for j matched groups.

$$\boldsymbol{\alpha_{MH}} = \frac{\sum_j \boldsymbol{A_j D_j / T_j}}{\sum_j \boldsymbol{B_j C_j / T_j}} \qquad (3\text{–}4)$$

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 3.8 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:

$$\beta_{MH} = \log_e(\alpha_{MH}) \qquad (3\text{–}5)$$

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

$$|\mathrm{D}| = -2.35\beta_{MH} \qquad (3\text{–}6)$$

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 disfavors the focal group above or 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 3.9 shows field-test EOG and EOC item pool DIF summary by flagging classification from 2018–19 administration. The NCDPI's rule is to exclude all items from the final pool that are flagged as DIF "C". 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 3.9 Mantel-Haenszel Delta DIF Summary for the EOG Reading and EOC English II Field-Test Items, 2018–19

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	482	21	9	486	24	2	460	32	20	509	2	1	502	9	1
4	325	23	2	309	28	13	299	32	19	348	2		336	14	
5	325	23	2	321	26	3	307	33	10	343	7		344	6	
6	316	17	3	321	14	1	316	18	2	329	7		333	3	
7	297	27	12	312	22	2	299	29	8	330	6		332	4	

8	327	21	2	330	17	3	324	22	4	349	1		344	6	
English II	484	17	2	483	16	4	471	28	4	503			499	4	

A=A/A+/A-, B=B/B+/B-, C=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 for the general tests 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 Characteristic 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.

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

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 (ρ) 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 equation 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.

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

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 θ, $P_i'(\theta) = \delta P/\delta\theta$, 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.

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \quad (4–3)$$

In Phase 2 of assembly, IRT parameters and the recommended form from the macro are output into 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 Reading, and EOC English II assessments are developed to align to the newly adopted content standards. As documented in chapters 1 and 2 of this report, changes from *Edition 4* to *Edition 5* pertain to test length and content blueprint with minimal changes in the assessed standards. Therefore, 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 that 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 Reading and EOC English II on a developmental scale, the statistical targets are determined independently for each grade based on the content complexity of grade level content standards and form level statistical specifications. 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 perfectly overlay each other. The TCCs of the newly developed parallel forms across grades 3–8 Reading and English II, based on the IRT item parameters estimated from 2018–19 embedded field-test administration, are shown in *Figure* 4.1 through *Figure* 4.7. For grade 4, two selected items were marginally revised. Therefore, their parameters are taken from 2020–21 administration for TCC derivation.

The TIFs and conditional standard error of measurements (CSEMs) are shown in *Appendix 4-A*. *Figure 4.1* through *Figure 4.7* show that the TCCs for parallel forms 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. This inference is possible 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

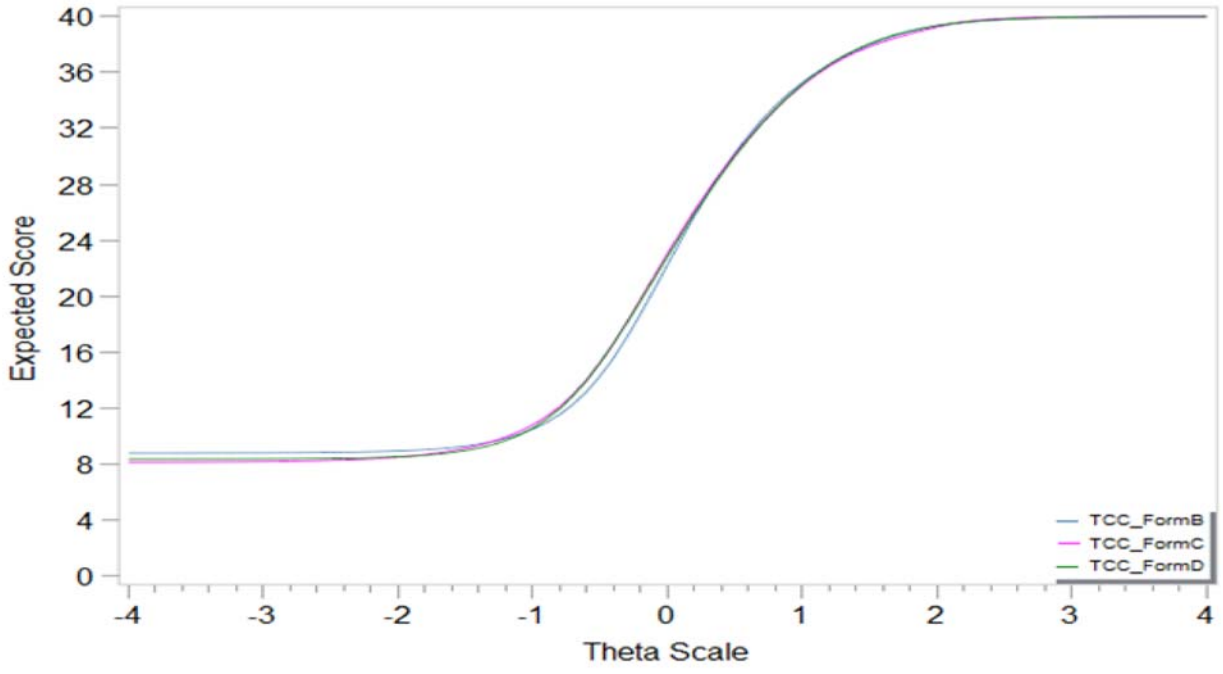


Figure 4.2 TCCs Based on Field-Test Item Parameters, Grade 4 (item parameters for two items were from 2020–21 administration)

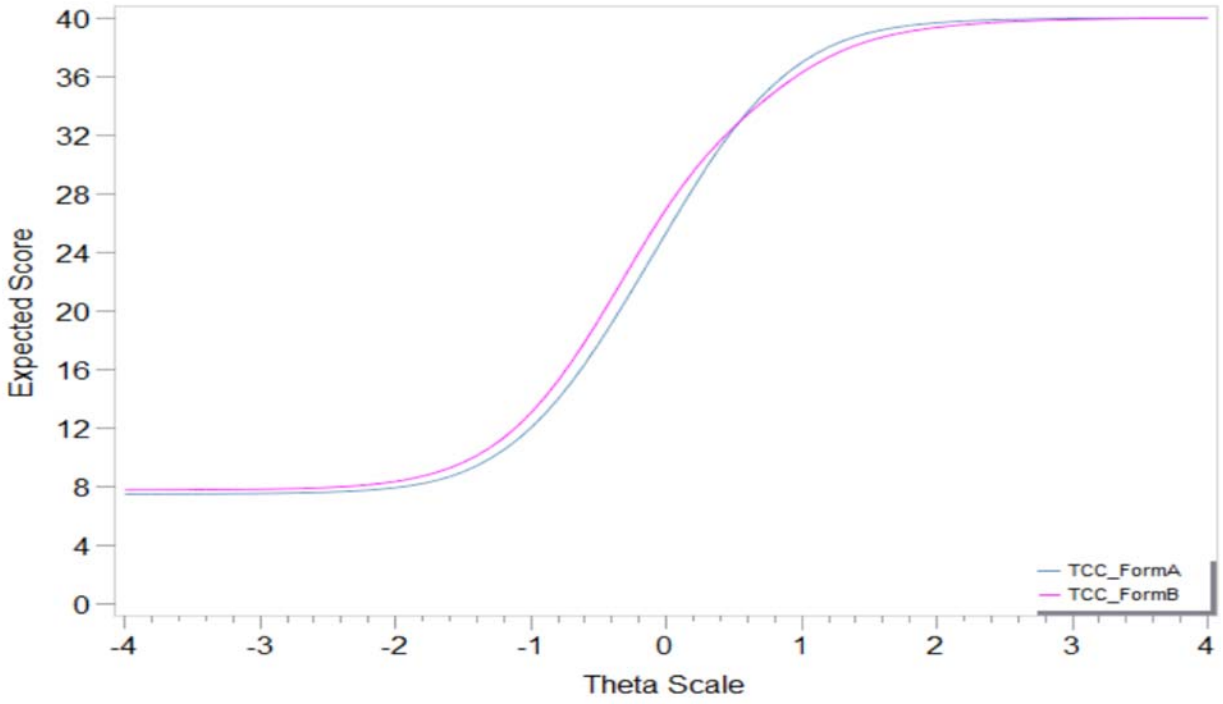


Figure 4.3 TCCs Based on Field-Test Item Parameters, Grade 5

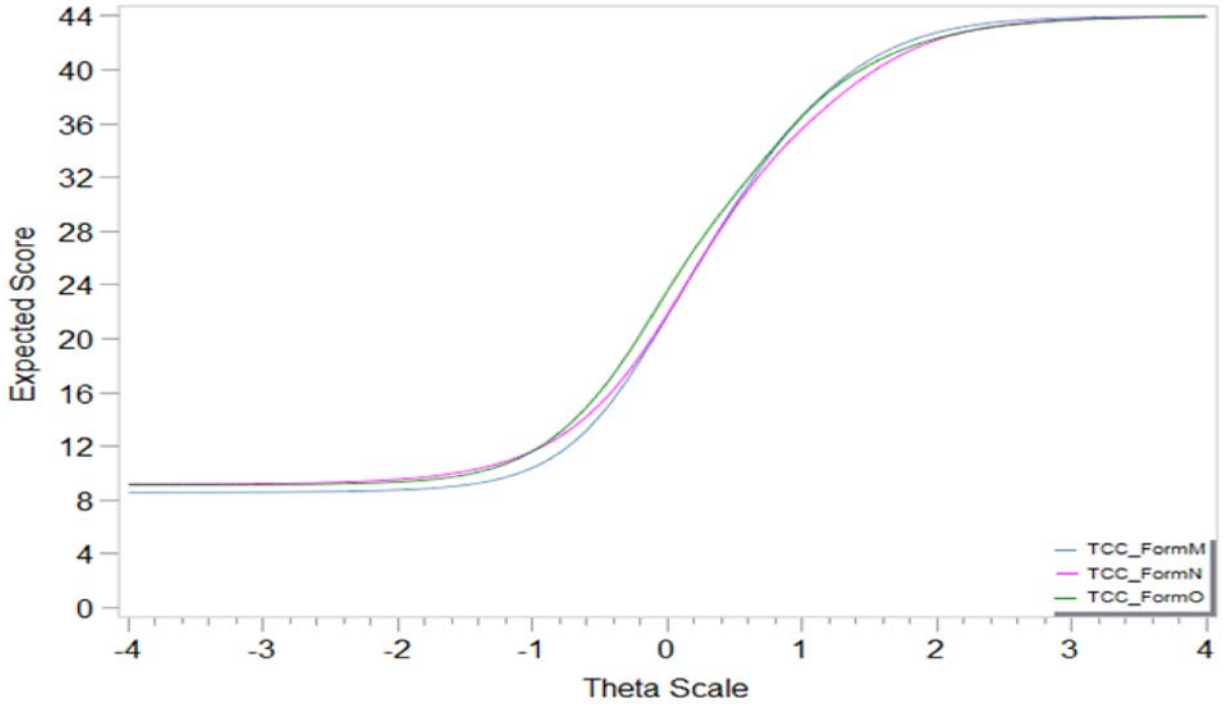


Figure 4.4 TCCs Based on Field-Test Item Parameters, Grade 6

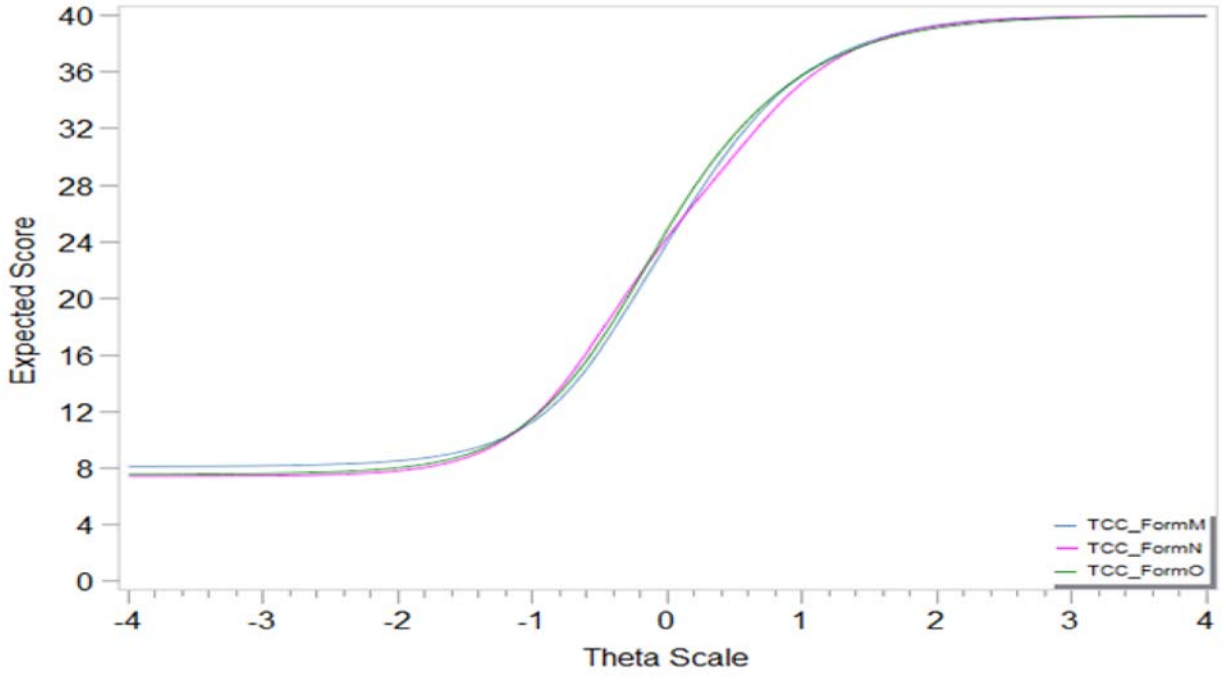


Figure 4.5 TCCs Based on Field-Test Item Parameters, Grade 7

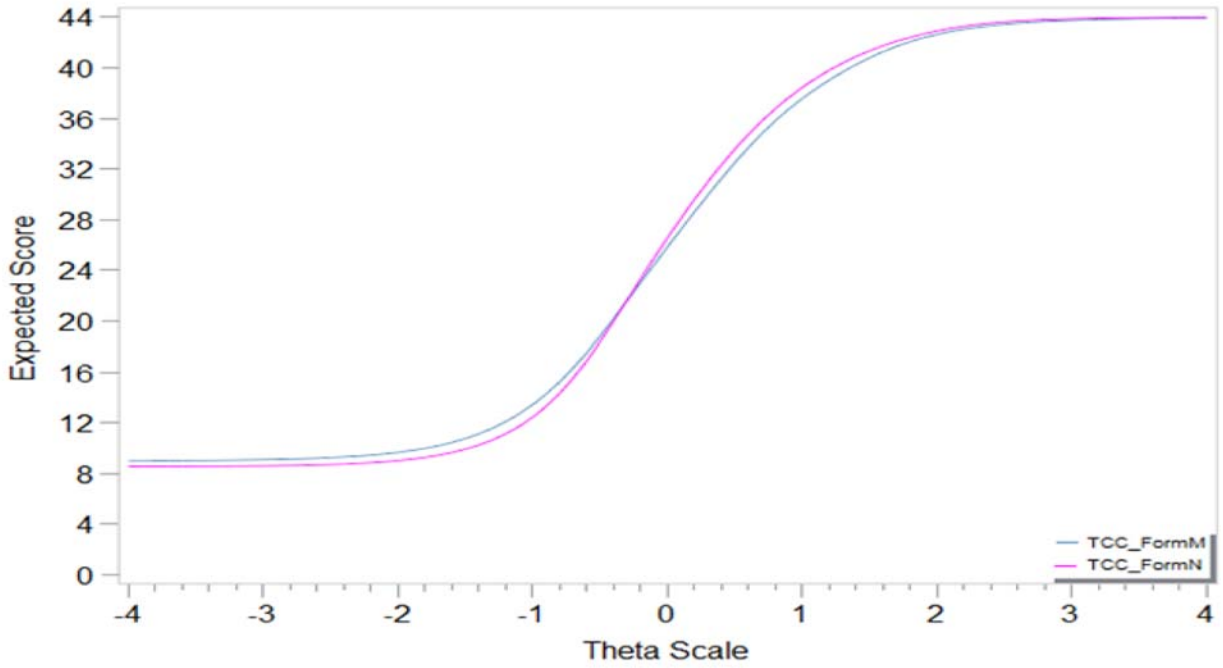


Figure 4.6 TCCs Based on Field-Test Item Parameters, Grade 8

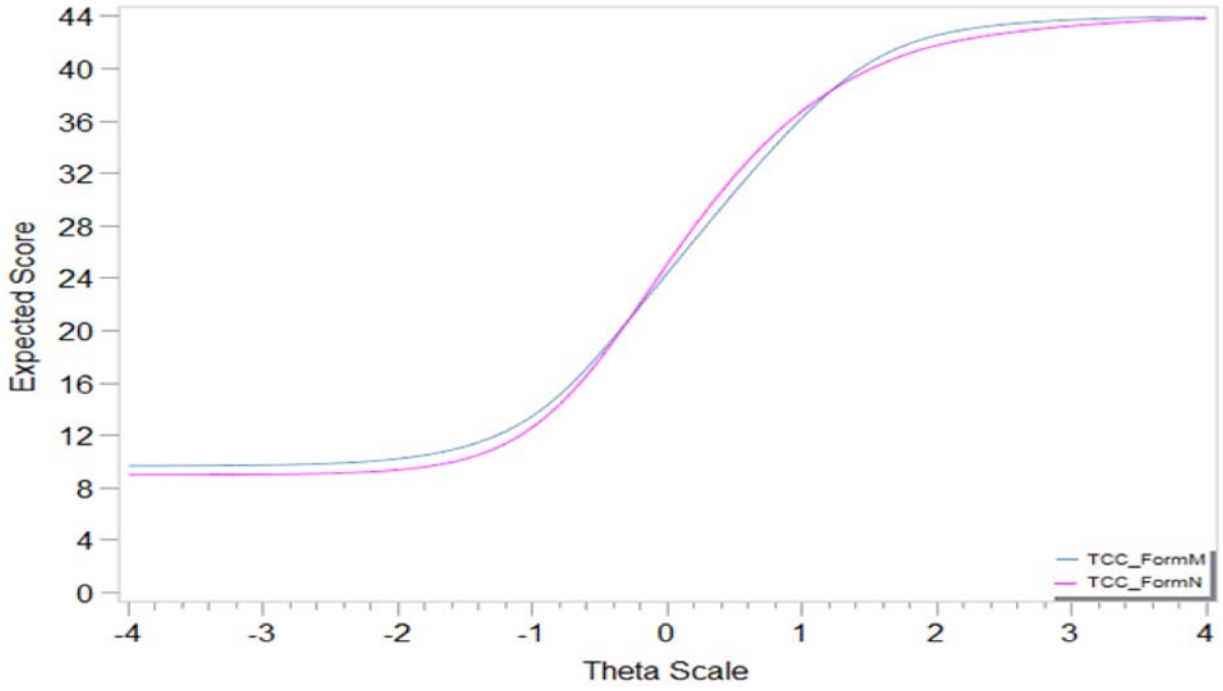
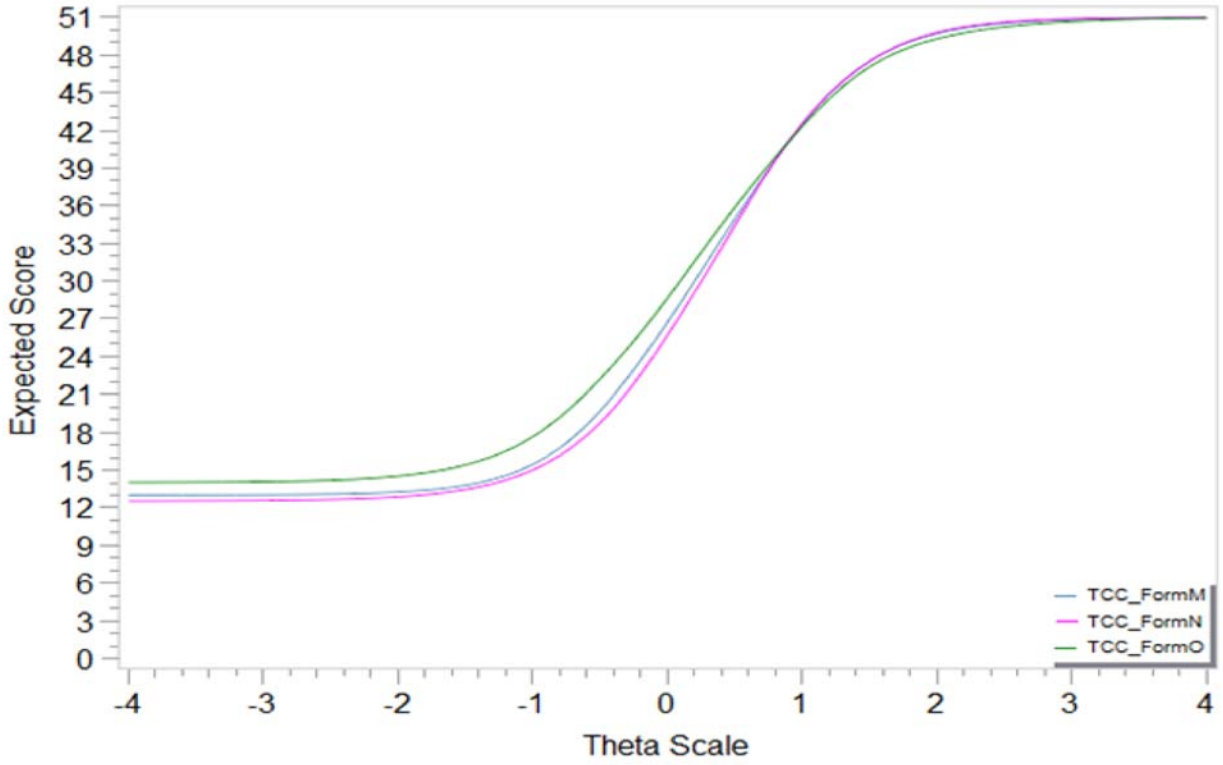


Figure 4.7 TCCs Based on Field-Test Item Parameters, English II



4.3 Form Review

After the initial assembly and statistical review (Step 1) of the form development process is complete, the forms undergo 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 Operational Base Form Review Process (*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 an 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 8 (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, 13, 15, 17, 19, and 20). 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 export the form as a document and format 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. NCSU-TOPS and 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. All forms are checked on the secure browser, the 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 places 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 2018–19, the Fairness Review panel for EOG Reading and EOC English II was made up of 11 participants representing teachers and educators. These members were selectively recruited

based on their expert knowledge of Reading content. Their demographic information is summarized in *Table 4.1*.

Table 4.1 Demographic Information of Fairness Review Panel

Category	Subcategory	N	%
Gender	Female	5	45%
	Male	6	55%
Ethnicity	African American	4	36%
	Asian	1	9%
	Caucasian	3	27%
	Hispanic	1	9%
	Native American	1	9%
	Other	1	9%
Highest Degrees Earned	BA/BS	4	36%
	MA/MS	5	45%
	Other	1	9%
	Ph. D	1	9%
Year of Experience	>20	6	55%
	10–20	3	27%
	1–10	2	18%

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 the 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 Reading and EOC English II forms. Notice that DIF flags for the forms across grades and courses were mostly category "A" and a 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* based on 2018–19 field-test 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 Grades 3–8 Reading and English II Edition 5 Test Items by DIF Types

Grade	Form	DIF Classification		Total Items
		DIF A	DIF B	
3	N	34	6	40
	O	38	2	40
	P	36	4	40
4	M	27	11	40
	N	33	7	40
5	M	34	6	40
	N	31	9	40
	O	33	7	40
6	M	38	6	44
	N	42	2	44
	O	42	2	44
7	M	36	8	44
	N	38	6	44
8	M	38	6	44
	N	41	3	44
English II	M	45	3	51
	N	42	6	51
	O	44	4	51

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 panelists, 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 NCSU-TOPS and the NCDPI 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* EOG Reading and EOC English II test forms that were built from 2018–19 embedded field-test items. All forms were built based on test specification criteria outlined in Chapter 2.

4.5.1 *Edition 5* EOG and EOC Operational Test Format

Table 4.3 and *Table 4.4* display the test format of the final assembled operational base forms in terms of item counts and item types. For grades 3–8, the paper-based, or PBT, and computer-based, or CBT forms are identical. The English II forms include technology enhanced (TE) item types and CR items. The technology enhanced item types are only placed on CBT forms. The paper-based accommodation forms only include MC and CR item types. Each MC and TE item is worth 1 point and each CR item is scored using a 0–2 scale. Examples of the TE item types can be accessed from the NCDPI website (*Appendix 4-C*).

Table 4.3 Test Format of EOG Grades 3–8 Reading

Grade	Number of Items	Form	Item Types
			MC
3	40	N	40
	40	O	40
	40	P	40
4	40	M	40
	40	N	40
5	40	M	40
	40	N	40
	40	O	40
6	44	M	44
	44	N	44
	44	O	44
7	44	M	44
	44	N	44
8	44	M	44
	44	N	44

Table 4.4 Test Format of English II

Course	Number of Items	Form	Item Types			
			MC	CR	SR	TI
English II	51	M	42 (1)	3 (2)	6 (1)	0
	51	N	43 (1)	3 (2)	3 (1)	2 (1)
	51	O	42 (1)	3 (2)	5 (1)	1 (1)

*Numbers in parenthesis are possible score points for each item type.

4.5.2 DOK Distributions

Test specification guidelines for cognitive complexity using DOK are shown in *Table 4.5* for grades 3–8 Reading and English II. The DOK specification is considered as a second order priority during form assembly and these ranges represent general expectation. Since other first order priorities such as statistical target and content specification take precedence over DOK specification a good effort is made to ensure forms are aligned to DOK specification. However, if in cases they are slightly off and all other test specifications are met, the form is not revised.

Table 4.5 Grades 3–8 Reading and English II DOK Distributions

Grade	Category	Blueprint (%)	Form M		Form N		Form O		Form P	
			N	%	N	%	N	%	N	%
3	DOK1	20–40			9	23	12	30	12	30
	DOK2	60–80			29	73	25	63	25	63
	DOK3	0			2	5	3	8	3	8
	Total				40		40		40	
4	DOK1	12–25	9	23	5	13				
	DOK2	50–75	30	75	31	78				
	DOK3	5–10	1	3	4	10				
	Total		40		40					
5	DOK2	75–90	34	85	30	75	34	85		
	DOK3	10–25	6	15	10	25	6	15		
	Total		40		40		40			
6	DOK2	60–82	29	66	27	61	31	70		
	DOK3	18–40	15	34	17	39	13	30		
	Total		44		44		44			
7	DOK2	60–82	32	73	31	70				
	DOK3	18–40	12	27	13	30				
	Total		44		44					
8	DOK2	60–82	34	77	34	77				
	DOK3	18–40	10	23	10	23				
	Total		44		44					
English II	DOK2	60–75	35 (35)	68.6 (65)	34 (34)	66.7 (63)	33 (33)	64.7 (61)		
	DOK3	25–40	16 (19)	31.3 (35)	17 (20)	33.3 (37)	18 (21)	35.3 (39)		
	Total		51 (54)		51 (54)		51 (54)			

*The DOK distributions for English II are based on score points listed in parentheses.

4.5.3 Summary CTT and IRT Statistics of Base Forms

Table 4.6 and *Table 4.7* 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 Reading and EOC English II forms. Form level statistics were based on embedded spring 2018–19 field-test data. Both CTT and IRT statistics confirmed forms within grade were built very similar to the statistical target. This evidence suggests forms within grades are statistically equivalent or parallel.

Table 4.6 Average CTT and IRT Statistics for Grades 3–5 Operational Forms Based on 2018–19 Field–Test

Grade	Form	Number of Items	CTT		IRT		
			P-value	Biserial-Corr.	Slope (a)	Threshold (b)	Asymptote (g)
3	N	40	0.58	0.42	1.98	0.20	0.22
	O	40	0.59	0.42	2.05	0.13	0.20
	P	40	0.58	0.41	1.85	0.15	0.21
4	M	40	0.63	0.43	1.82	-0.09	0.19
	N	40	0.64	0.42	1.80	-0.12	0.19
5	M	40	0.59	0.42	1.91	0.04	0.20
	N	40	0.59	0.43	1.82	0.02	0.19
	O	40	0.60	0.42	1.72	0.00	0.19

Table 4.7 Average CTT and IRT Statistics for Grades 6–8 and English II Operational Forms Based on 2018–19 Field-Test

Grade	Form	Number of Items	CTT		IRT		
			P-value	Biserial-Corr.	Slope (a)	Threshold (b)	Asymptote (g)
6	M	44	0.53	0.41	1.81	0.33	0.19
	N	44	0.54	0.38	1.80	0.37	0.21
	O	44	0.55	0.39	1.84	0.30	0.21
7	M	44	0.59	0.40	1.79	0.11	0.20
	N	44	0.59	0.41	1.62	0.07	0.19
8	M	44	0.57	0.37	1.81	0.23	0.22
	N	44	0.58	0.39	1.70	0.23	0.20
English II	M	51	0.59	0.37	1.81	0.30	0.25
	N	51	0.58	0.38	1.88	0.32	0.25
	O	51	0.60	0.36	1.77	0.28	0.27

4.6 Future Embedding Plan for Field-Test

Each grade specific operational EOG Reading form consists of eight (8) slots and EOC English II has nine (9) slots for embedding field-test items. Depending on the needed number of new operational forms for future use, appropriate embedding plans specifying number of items to be field tested are developed. The NCDPI's internal rule is to field-test three times more items than needed. For example, if there is a need to develop a new form with 40 items, at least 120 items (40×3) will be field tested in order to have quality items in the item bank.

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 Proctor's Guide can be accessed from the NCDPI website (*Appendix 5-A*).

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 testing English Learners students can be accessed from the NCDPI website shown in *Appendix 5-B*. The document for students with disabilities can be accessed from the NCDPI website shown in *Appendix 5-C*. 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 directions 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 (*Appendix 5-D*): The NCDPI publishes this document in order to maintain the integrity of the NCATP. It is essential for school personnel to develop awareness of proper testing protocol and procedures. Knowledge of testing policies and procedures helps ensure the NCATP 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 NCATP. This testing security guide may be kept in the schools.

North Carolina Test Coordinators' Policies and Procedures Handbook: The purpose of the handbook is to provide public school units (PSU) test coordinators with a reference for implementing proper test administrations for the NCATP. The handbook (*Appendix 5-E*) can be accessed from the NCDPI website. 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 NCATP 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 Administrator Training

The test administrators' training utilizes the *North Carolina Test Coordinators' Policies and Procedures Handbook* (*Appendix 5-E*) 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 NCATP 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 PSU test coordinators on the processes for proper test administration. PSU 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*.

5.3 Test Security and Administration Policies

Test security is an ongoing concern for the NCATP. 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 PSU 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, PSUs 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 authorized employees of the LEA.

At the individual school, the principal is responsible for all test materials received. As established by NCSBE 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. PSU 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 PSU 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 PSU test coordinator. Upon notification, the PSU 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 PSU test coordinator and/or the NCDPI Division of Accountability Services / NCATP.

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

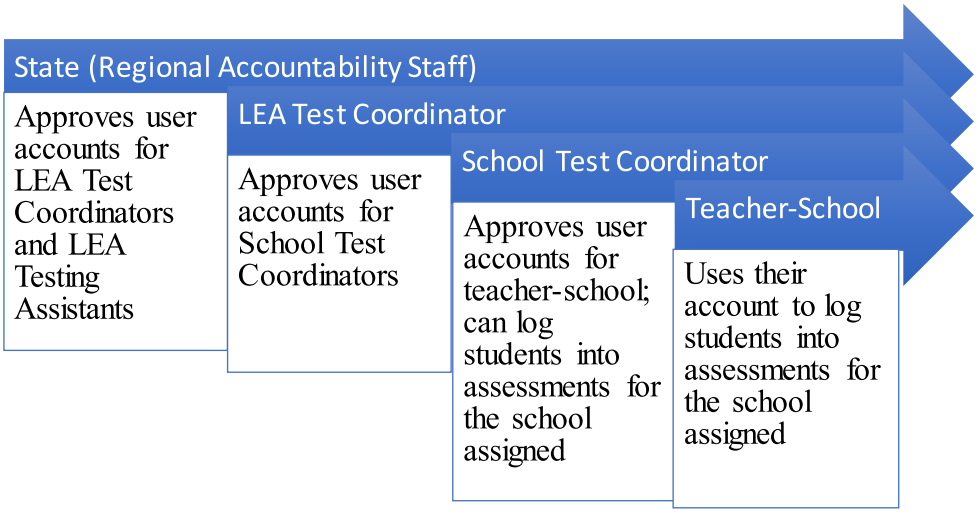
Test Material	Required Storage Time
All used answer sheets for operational tests (including scoring sheets for W-APT)	Six months after the return of students' test scores
Original responses recorded in a test book, including special print version test books (i.e., large print edition, one test item per 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 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 PSUs access to the CBT to specific testing days. The PSU 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 Chromebooks, 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.

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. English II is a semester course administered in fall and spring.

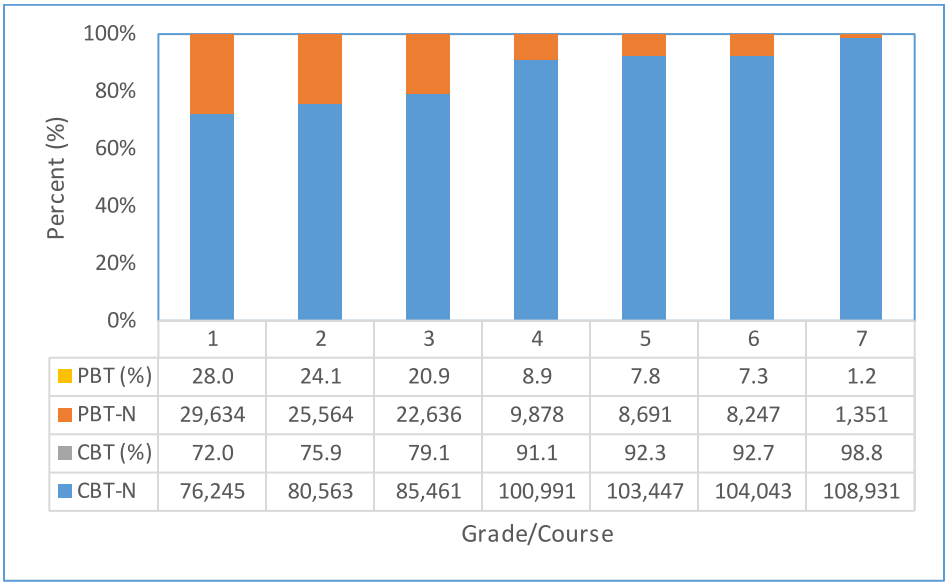
5.4.2 Modes of Test Administration

From the 2020–21 administration, grades 3–8 EOG Reading assessments are available in both PBT and CBT modes. The state's goal is to gradually transition test administrations for EOG to CBT mode as districts build their resources and technology capacity. Beginning from 2018-19 PSU are required to administer EOC English II on CBT mode and PBT are only available for students as accommodations.

Figure 5.2 shows the proportion and total number of students who took the EOG Reading and EOC English II tests by mode in the 2020–21 administration. Notice that the proportion of students who took in CBT mode increased gradually as the grade level increased from 72% in grade 3 to 98.8% in English II. The proportion of students taking CBT forms is expected to continue to grow as the state moves to require CBT administration for more grades. In 2021–22 school year, all students in grades 6–8 Reading and English II are required to take the tests in CBT mode except for those with documented accommodation needs. From 2022–23 school year, all students in grades 3–8 and English II will be required to take their EOG in CBT mode.

From 2021–22 and beyond, two operational CBT forms will be administered with two operational PBT forms allocated for accommodations and misadministration. A new embedding plan, in terms of number of flavors, for field testing items will be proposed for developing new operational forms as needed.

Figure 5.2 Number (N) and Percent (%) of Students by Mode, 2020–21



	1	2	3	4	5	6	7
PBT (%)	28.0	24.1	20.9	8.9	7.8	7.3	1.2
PBT-N	29,634	25,564	22,636	9,878	8,691	8,247	1,351
CBT (%)	72.0	75.9	79.1	91.1	92.3	92.7	98.8
CBT-N	76,245	80,563	85,461	100,991	103,447	104,043	108,931

Grade/Course

5.4.3 Testing Time Guidelines

The EOG and EOC are not speeded or power test, so the timing guidelines below are meant to guide with planning and scheduling. 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 an additional one hour if needed to complete the test. For the EOC English II, 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 2020–21 operational assessments are shown in *Table 5.2*. The table includes data for EOG and EOC CBT forms administered under regular conditions—that is, without accommodations of *Scheduled Extended Time* and *Multiple Testing Sessions*. For grades 3–8 Reading, the table shows 95% of students were able to complete their EOG session within three (3) hours (174 minutes or less). The data also shows that 50% of students were able to complete their EOG in just over 1 hour 30 minutes. For English II, data shows 95% of students were done with their test session in just over 3 hours (198 minutes) of the recommended 4 hours window. All this is evidence to show the recommended time allotted for EOG and EOC gives

students ample time to complete the assessment. For those students who are not able to complete the assessment within the recommended window, the NCDPI grants additional time.

Table 5.2 Recorded Test Duration for EOG Reading and EOC English II Forms, 2020–21

Grade	N	No. of OP+FT Items	Summary		Percentile				
			Mean	SD	25th	Median	75th	95th	99th
3	76,244	48	91.7	46.5	60	81	114	174	222
4	80,563	48	93.7	45.1	66	87	114	171	213
5	85,461	48	99.0	43.4	69	93	120	174	216
6	100,991	52	103.3	43.8	75	99	123	171	213
7	103,447	52	101.6	39.4	75	99	120	165	201
8	104,043	52	97.9	36.8	72	96	117	159	195
English II	108,931	60	128.3	42.8	99	126	153	198	240

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 integrity to ensure these scores are reliable and valid for uses.

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 standard accommodations in North Carolina Reading EOG and EOC English II assessments. Special accommodations requests are continuously reviewed by the NCDPI and approved as deemed necessary.

- Special Print Versions
    - Paper
    - Braille Edition
    - Large Print Edition
    - 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
    - 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. In addition, test administrators must be trained in the use of the specified accommodations by the PSU 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 NCSBE 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 1–12 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 (NCSBE 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, and Student Reads Test Aloud to Self.

Table 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 require that all public and charter 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 3–8

Reading 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 English II 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, PSUs 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, PSUs 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. PSUs 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 PSU 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 are 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/ NCATP 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 Med Exception Memo_CE_TH_072021 (nc.gov).

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* (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 PSU 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 an NC Education user account. The OTISS irregularity report must be submitted to the PSU 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 PSU 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/ NCATP
- Failing to follow the test schedule procedures or makeup test schedule designated by the NCDPI Division of Accountability Services/ NCATP
- 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.

PSUs must also monitor test administration procedures. According to NCSBE 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 in scores triggers further analysis to verify the sources of score gains.

Residual Analysis. At the end of every testing cycle, the NCDPI psychometric team performs a series of residual analyses at the test and item level to verify that pre-equated IRT scales and item parameters continue to maintain their same pattern and meaning. Any larger than expected drift of IRT parameter or test scale leads to further analysis to verify that differences can be explained by changes in examinee standing with respect to the constructs measured.

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 scores and reports 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. The use of pre-equated scoring model in North Carolina dates back to the 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 a decentralized scoring system based on number correct. Another practical consequence is that the NCDPI can use a short administration window (the last 5–10 days of the school year) for EOG and EOC assessments, and is still able to provide and use scores for end of year reporting.

6.2 Final Parameters for Scale Development

With any new Edition of EOG or EOC, the base year CTT and IRT parameters estimation, scale development, and standard setting are based on item parameters and scores from the first operational administration. For *Edition 5* EOG, this plan was modified to adjust for unique circumstances due to the COVID related disruption of schools in 2019–20 and the eventual suspension of all state testing. Even though *Edition 5* EOG reading assessments were operationally administered in 2020–21, after a thorough review of test data and student participation, the North Carolina Technical Advisory Committee advised 2020–21 assessment data were unreliable and recommended to not use them for scale development and standard setting. Their conclusion was supported by data from across the state which primarily showed ununiform instructional practices, varying opportunity to learn for students, and disproportionate participation rates in schools and districts across states. Also, state and local accountability were waived for 2020–21 which changed the meaning of test scores. Council of Chief State School Officers (CCSSO, 2020) in its report on "Restart & Recovery: Assessments in Spring 2021" questioned the reliability of the 2020–21 data and cautioned states against the use of data from this administration for high stakes decision making. Based on all these circumstances, North Carolina Technical Advisory (NCTA) advised the NCDPI consider using pre-pandemic field-test parameters from 2018–19 for scale development and standard setting.

For English II, new forms for *Edition 5* were administered operationally in fall 2019 (pre-pandemic) for the first time to students across the state. The sample characteristics were representative and sufficient to estimate reliable item parameters for operational administration, scale development, and standard setting.

6.3 Drift Analysis Between Field-Test and Operational Administration

At the conclusion of spring testing in 2021, the NCDPI conducted drift analysis to understand the impact on item parameters and test forms from the COVID-19 pandemic. The NCDPI compared CTT and IRT summary statistics for grades 3–8 Reading and English II, at the form level, between the 2018–19 field-test and 2020–21 operational administration shown in *Table 6.1* and *Table 6.2*. Summary results show average P-value were consistently lower (-0.02 to -0.07) across grades 3–8 from spring 2021 operational administration. For English II, the difference was minimal (0 to -0.02).

IRT parameters calibrated using field test data and again after the operational administration shown in *Table 6.1* and *Table 6.2* also show similar trends across forms for grades 3–8 where IRT threshold (b) parameters for 2020–21 are consistently larger (0.06 to 0.26) than 2018–19 indicating 2020–21 students perceived the tests as relatively difficult. For English II, the difference is reversed with b-parameter difference between 2020–21 and 2018–19 ranged from -0.02 to -0.13 across forms suggesting 2020–21 population perceived English II form as relatively easier.

The overall trend depicted by the TCCs between 2018–19 field-test and 2020–21 operational administration are shown in *Figure 6.1* through *Figure 6.7*. The TCCs based on the 2020–21 operational administration for forms (solid lines) in a given grade level, for the most part, are towards the right of 2018–19 field-test (dash lines) across ability continuum, indicating students from 2020–21 operational administration perceived the forms as difficult overall. The TCCs for English II forms are clustered between each other. These results supported the importance of using pre-pandemic data from the 2018–19 school year for grades 3–8 Reading and Fall 2019–20 for English II for scale development and standard setting.

Table 6.1 Average CTT and IRT Statistics FT vs OP Grades 3–4 Reading, 2020–21

Grade	Form	No. of Items	CTT				IRT					
			P-Value		Biserial-Corr.		Slope (a)		Threshold (b)		Asymptote (g)	
			FT	OP	FT	OP	FT	OP	FT	OP	FT	OP
3	N	40	0.58	0.52	0.42	0.45	1.98	1.97	0.20	0.41	0.22	0.20
	O	40	0.59	0.52	0.42	0.47	2.05	2.00	0.13	0.33	0.20	0.19
	P	40	0.58	0.51	0.41	0.44	1.85	1.95	0.15	0.40	0.21	0.19
4	M	40	0.63	0.58	0.43	0.47	1.82	1.74	-0.09	0.00	0.19	0.17
	N	40	0.64	0.57	0.42	0.45	1.80	1.85	-0.12	0.14	0.19	0.19
5	M	40	0.59	0.55	0.42	0.45	1.91	1.83	0.04	0.26	0.20	0.20
	N	40	0.59	0.54	0.43	0.47	1.82	1.89	0.02	0.25	0.19	0.18
	O	40	0.60	0.58	0.42	0.46	1.72	1.76	0.00	0.04	0.19	0.17

Table 6.2 Average CTT and IRT Statistics FT vs OP Grades 6–8 Reading and English II, 2020–21

Grade/ Course	Form	No. of Items	CTT				IRT					
			P-Value		Biserial-Corr.		Slope (a)		Threshold (b)		Asymptote (g)	
			FT	OP	FT	OP	FT	OP	FT	OP	FT	OP
6	M	44	0.53	0.50	0.41	0.43	1.81	1.75	0.33	0.45	0.19	0.18
	N	44	0.54	0.49	0.38	0.41	1.80	1.70	0.37	0.52	0.21	0.18
	O	44	0.55	0.51	0.39	0.42	1.84	1.74	0.30	0.44	0.21	0.19
7	M	44	0.59	0.56	0.40	0.42	1.79	1.68	0.11	0.17	0.20	0.18
	N	44	0.59	0.54	0.41	0.43	1.62	1.72	0.07	0.26	0.19	0.18
8	M	44	0.57	0.53	0.37	0.40	1.81	1.68	0.23	0.35	0.22	0.19
	N	44	0.58	0.53	0.39	0.42	1.70	1.66	0.23	0.38	0.20	0.20
English II	M	51	0.59	0.57	0.37	0.43	1.81	1.59	0.30	0.28	0.25	0.22
	N	51	0.58	0.58	0.38	0.44	1.88	1.61	0.32	0.25	0.25	0.21
	O	51	0.60	0.59	0.36	0.43	1.77	1.48	0.28	0.15	0.27	0.22

Figure 6.1 TCCs Between 2018–19 FT and 2020–21 OP Parameters, Grade 3 Reading

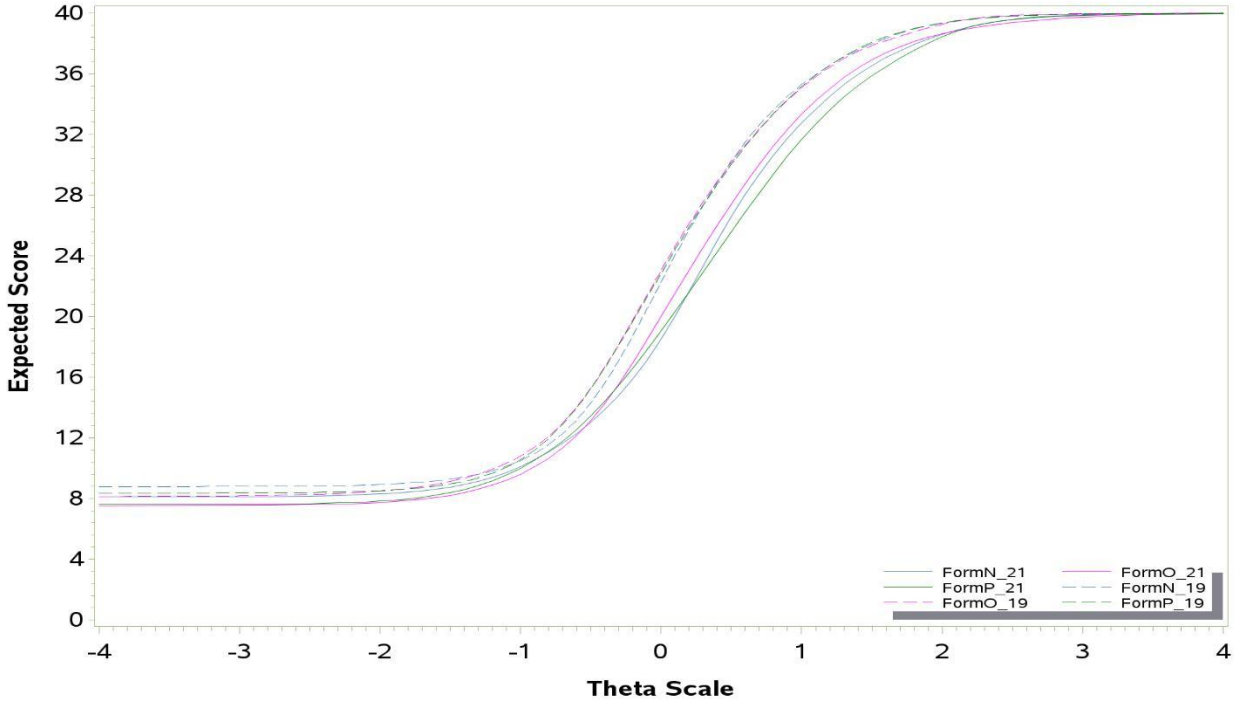


Figure 6.2 TCCs Between 2018–19 FT and 2020–21 OP Parameters, Grade 4 Reading

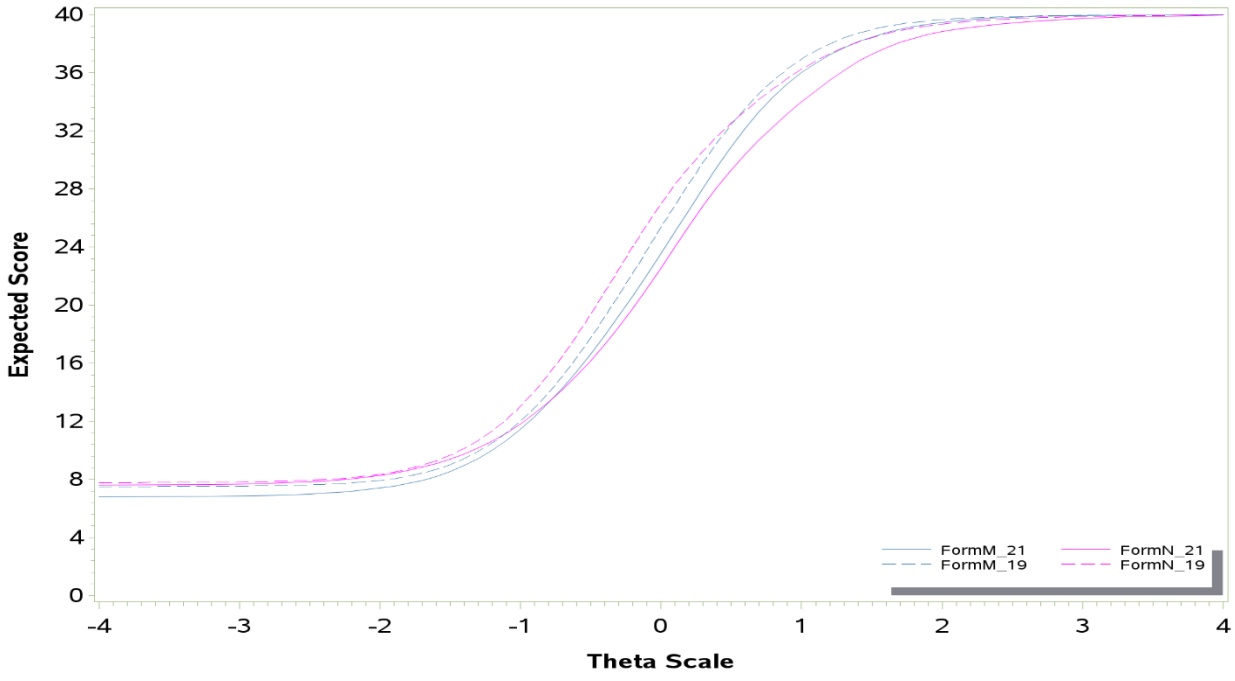


Figure 6.3 TCCs Between 2018–19 FT and 2020–21 OP Parameters, Grade 5 Reading

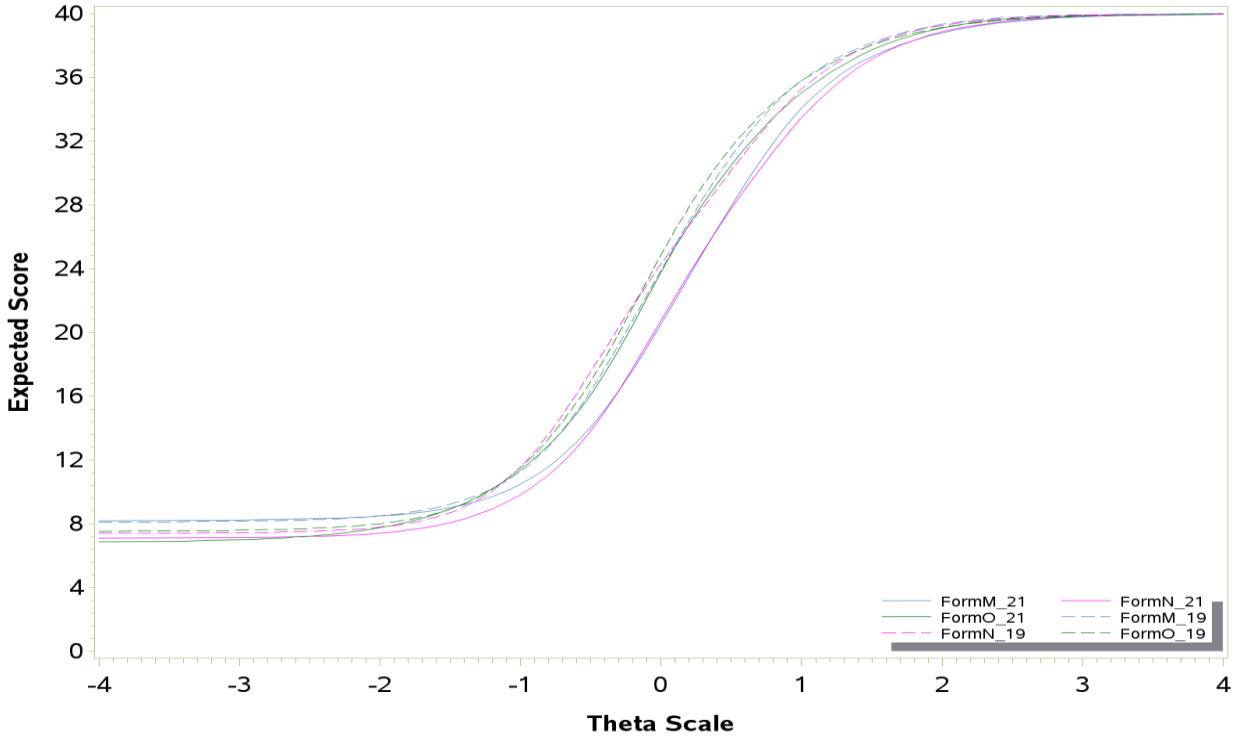


Figure 6.4 TCCs Between 2018–19 FT and 2020–21 OP Parameters, Grade 6 Reading

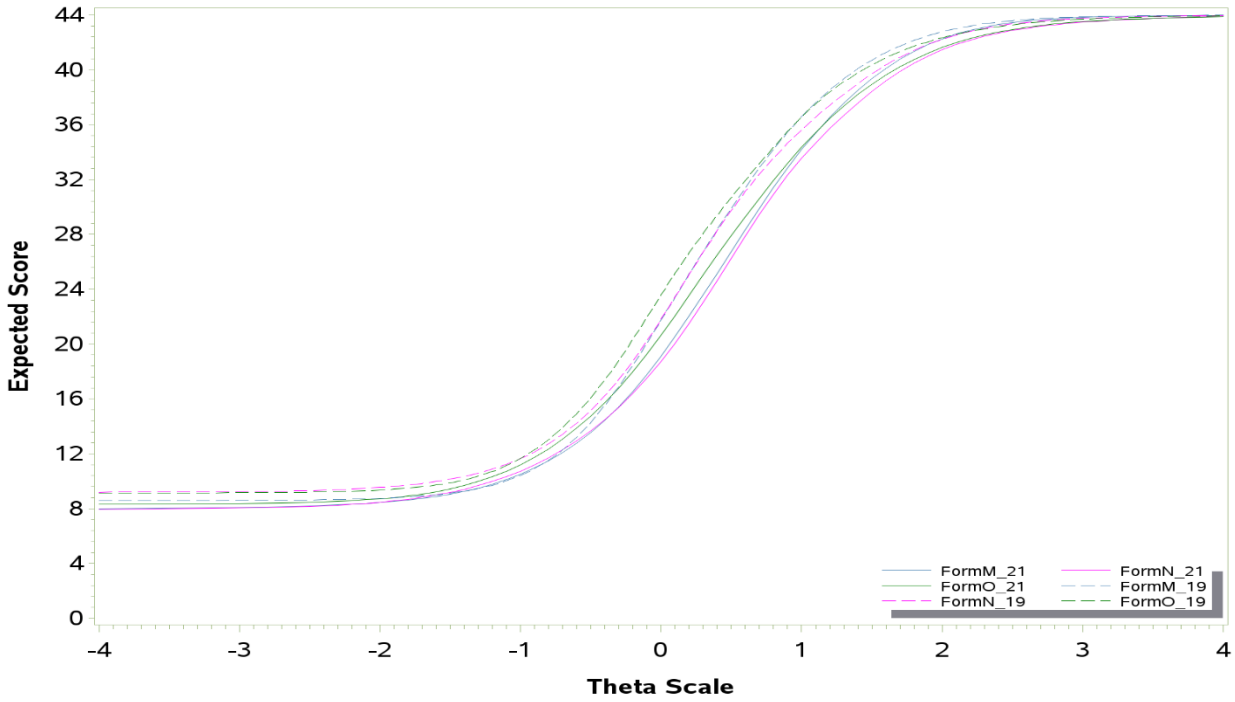


Figure 6.5 TCCs Between 2018–19 FT and 2020–21 OP Parameters, Grade 7 Reading

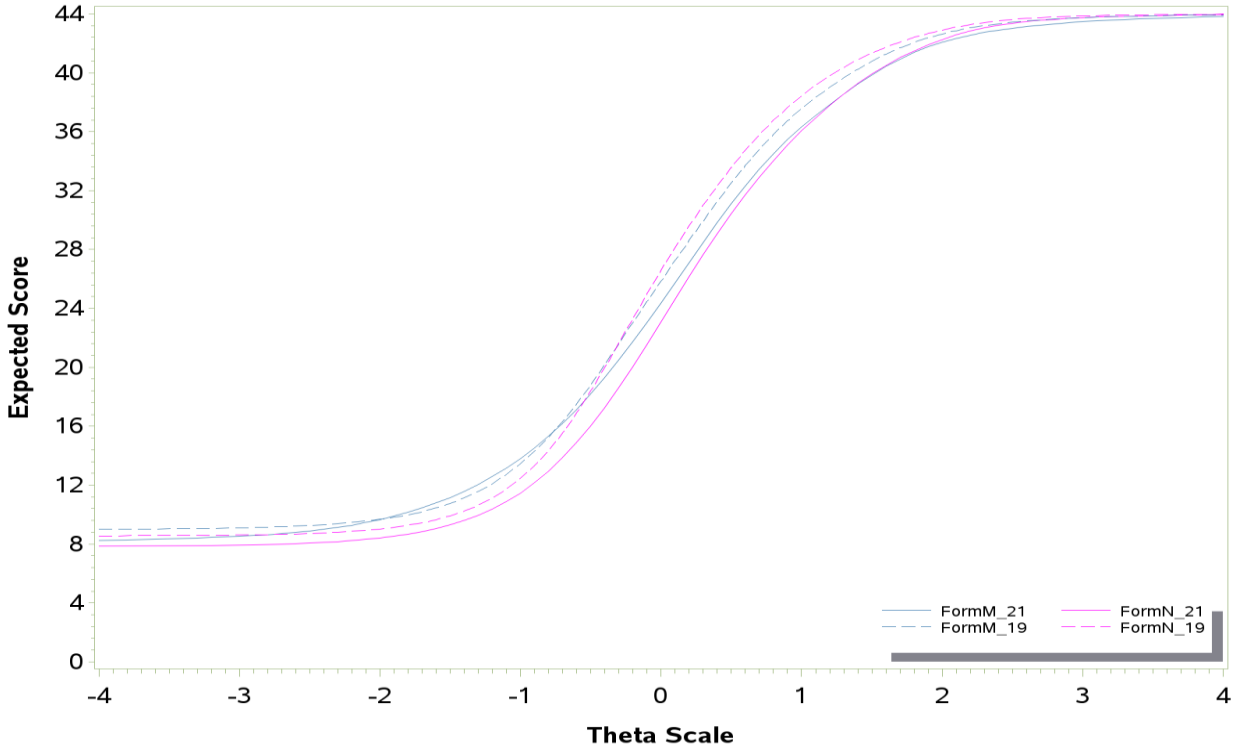


Figure 6.6 TCCs Between 2018–19 FT and 2020–21 OP Parameters, Grade 8 Reading

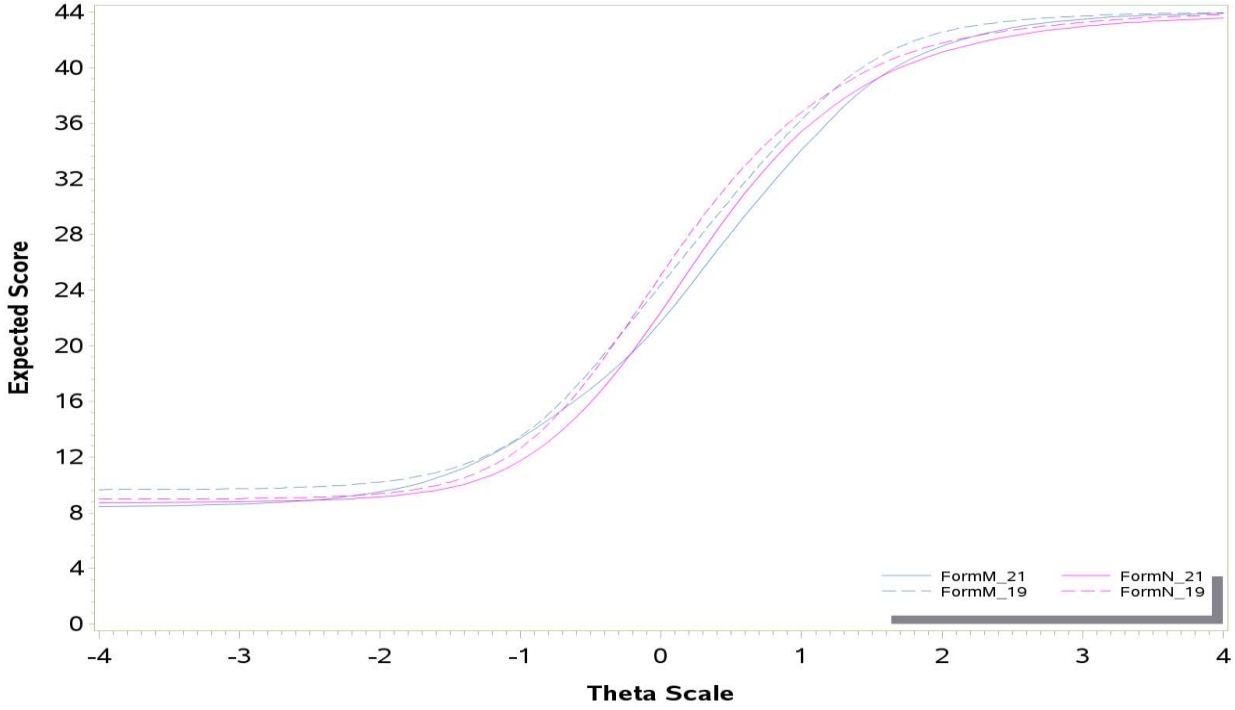
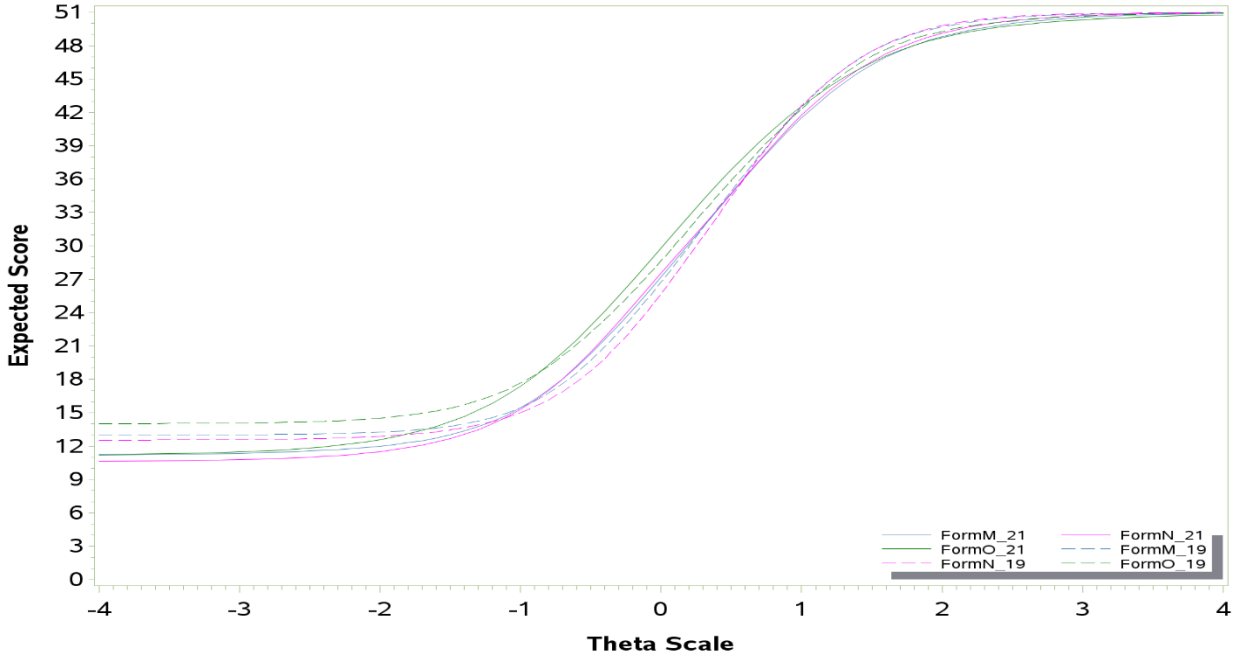


Figure 6.7 TCCs Between 2018–19 FT and 2020–21 OP Parameters, English II



6.4 Impact of Instructional Interruptions

The NCDPI commissioned University of North Carolina, Greensboro Office of Assessment, Evaluation, and Research Services (OAERS) for an independent study to investigate the impact of interruptions to full-time, in-person instruction due to COVID-19 on student learning, particularly performance on EOG mathematics grades 4 and 7 and EOC NC Math 1 tests. The OAERS concluded that on average, scale scores for students who took EOG and EOC mathematics in 2020–21 were about one-half standard deviation lower than those from 2018–19. The percentage of students classified as grade-level proficient for the 2020–21 population overall, and by sub-group were significantly lower than those from 2018–19.

Based on the drift analysis results and findings from the white paper, the NCDPI with recommendation from the NC Technical Advisors decided to use item parameters from the 2018–19 pre-pandemic field-test administration for grades 3–8 reading for scale development and standard setting. For English II, item parameters from the pre-pandemic Fall 2019 operational administration were used for scale development and standard setting. The plan is to conduct a review to check the stability of item parameters and scale in subsequent years. The TIFs/CSEMs for these forms associated with updated 2020–21 IRT parameters are shown in *Appendix 6-A*.

6.5 IRT Summed Score Procedure

IRT parameters calibrated from either field-test or operational administration are used with IRT summed score procedure to create final raw-to-scale conversion tables. During the initial implementation year, students' scores are delayed until after the standard setting workshop is

complete and new performance achievement levels are adopted by the NCSBE 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:

$$L_x(\theta) = \sum_{\sum u_i = x} L(u/\theta) \qquad 6\text{–}1$$

Where $L\left(u/\theta\right) = \prod_i T(u_i/\theta)$ 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:

$$P_x = \int L_x(\theta) d(\theta) \qquad 6\text{–}2$$

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

$$E(\theta/x) = \frac{\int \theta L_x(\theta) d(\theta)}{P_x} \qquad 6\text{–}3$$

With posterior standard deviation given by

$$SD(\theta/x) = \sum u_i) = \left\{\frac{\int [\theta - E(\theta/x)]^2 L_x(\theta) d(\theta)}{P_x}\right\}^{1/2} \qquad 6\text{–}4$$

The values computed using *E (θ/x)* may be tabulated and used as the IRT raw-to-scale 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.

6.6 Score Comparability Across Forms

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 operational administration of *Edition 5* EOG the NCDPI administered at least two base forms. To ensure the grade specific base forms are equivalent and on the same scale, these forms were developed using the same statistical and content specifications. The processes of embedding field-test items with operational forms and spiraling of test forms at the classroom level ensured random equivalent distribution of test forms across the state. This allowed the NCDPI to perform a single group calibration of each form assuming the same joint population density to generate final IRT item parameter estimates on a common metric. These IRT parameters generated separately from each form but assumed to be on the common IRT scale are used with the summed score procedure to create raw-to-scale scores tables for each base form. The resulting raw to scale conversion tables are considered statistically equivalent and any form differences are accounted for so that scores from students taking different base forms are comparable. These processes allow the NCDPI to report reliable estimates of students' proficiency on EOG or EOC assessments for valid interpretation and uses.

6.7 Raw to Scale Scores

The NCDPI administers multiple forms of EOG and EOC within each grade every test cycle. In 2020–21 administration, the NCDPI randomly spiraled multiple 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 sample and test a broader range of grade-level content standards for each grade.
- 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.

A main limitation 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 specification 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. 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 allows for a valid comparison of students' performance across forms and between years even though students are 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. Raw scores across forms offer no inherent interpretative meaning of students' performance because the different sets of raw scores are not on the same reportable scale. A difference of one raw score point between group of students who took different forms does not imply students with a higher raw score performed better compared to those with the lower raw score.

The NCDPI only uses raw scores in the context of creating raw to scale tables. The NCDPI strongly advises 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.

6.8 Automated Decentralized Scoring

6.8.1 Selected and Short Response Items

All items on EOG and EOC assessments, with the exception of CR items, are designed so scoring could be automated. The NCDPI's reporting group gets final answer keys once all edits and checks on forms have been completed. These keys are then updated into the custom scoring software program and final tests are performed to ensure all items are being scored correctly. At the start of each testing window, a new version of the custom scoring program (WinScan) is made available to all PSUs for them to update their automated scoring routine.

For paper-based test forms, the PSU's test coordinator establishes the schedule for receiving, scanning and scoring EOG/EOC tests at the district level. The PSU's test coordinator upon receipt of student response answer 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 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 PSU 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, PSU test coordinators can generate individual student reports and other custom built-in reports of their PSU and school data. 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.

6.8.2 Constructed Response Scoring

This section briefly describes the scoring process for constructed response (CR) items administered operationally in 2020–21 and beyond. Questar Assessment Inc. (QAI) was the 2020–21 scoring vendor for the NCDPI. Starting from 2021–22, Cognia will be the scoring vendor.

Transportation and Processing

There are three operational CR items in each EOC English II form. The forms are administered in computer mode with a small number of paper accommodated forms. For scoring CR items in paper mode, districts/schools receive shipping labels from QAI to ship answer documents directly to QAI's facility. For CR items administered on computer, the student test records are transferred daily as Online Response Data Files via NCDPI's secured File Transfer Protocol (FTP) site. The FTP site serves two primary purposes: exchanging administrative documentation and exchanging student test material. The Student Test Data File Report with scored data are delivered by QAI to NCDPI within 14 business days after the administration has ended.

Rater Selection, Training and Qualification

AERA/APA/NCME (2014) Standard 4.20 specifies the following:

> *"The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters' scoring"* (p. 92).

1. **Project Staffing**

In general, QAI uses a hierarchy of Scoring Directors, Team Leaders, and Scorers. Scoring Directors are chosen for a project based on the following qualifications:

- 4-year degree
- Content expertise
- Previous project experience
- Experience with score point
- Ability to work under pressure to meet deadlines
- Ability to travel, facilitate, and interact with client
- Good work ethic and integrity
- Good verbal and written communication skills
- Evaluations
- Schedule flexibility

The Scoring Directors have the overall responsibility for the training of the project and content as well as the scoring expectations. They undergo extensive specialized training to prepare them for their roles as scoring experts and monitors by working with QAI or department content specialists.

Team Leaders report directly to the Scoring Directors and are typically in charge of a team of 10–12 scorers, depending on the item(s) and content area. They are specifically trained on the requirements and processes for scorer monitoring and intervention, including interpreting score point reports such as Reader Reliability (RR) and Score Point Distribution (SPD) reports, conducting read behinds, holding one-on-one discussions, and scoring.
Team Leaders (TLs) are selected based on:

- 4-year degree
- Content knowledge
- Previous project experience
- Experience with score point (QAI proprietary system)
- Evaluations

Scorers must have fulfilled the following requirements:

- 4-year degree (in a related field in the content area for which they will be scoring
- Attend an open house for an introduction to Questar philosophy
- Complete an application process, complete with references
- Complete a sample of the content area for which they are applying
- Complete a one-on-one interview with Questar scoring staff

2. Training

Training Materials

AERA/APA/NCME (2014) Standard 6.8 states, *those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented* (p. 118).

Training materials for North Carolina include responses scored during range finding that represent the full range of score points as determined by the range finding committees, including responses that exemplify the nuances of the rubric (e.g., differentiation of a low "3" from a high "2").

Training materials consisted of the following:

- One **Passage**
- One **Prompt and Rubric**
- One **Scoring Guide (or Guide Set)** containing approximately 10 items with a minimum of 3 anchor responses (1 for each score point). During training, the Scoring Guide was discussed response by response within the group setting to identify any nuances of individual responses that have been selected as exemplary. This phase also includes a discussion of often seen acceptable and unacceptable details for each item.
- A **Training Set** containing 10 responses representing a variety of score points in random order. The training set was scored independently by each scorer, and each response was discussed by the group. This set is used as a learning tool to assess whether the scorer understands the nuances as discussed in the Scoring Guide.
- A **Qualifying Set** containing 10 responses representing a variety of score points in random order. The qualifying set is scored independently by each scorer, and each response is discussed by the group. This set was used to determine whether a scorer is eligible to continue on to scoring. Meeting the qualification standards on this set demonstrates that the scorer will be able to apply the necessary skills to score.

Team Leader Training

AERA/APA/NCME (2014) Standard 6.9 specifies that "*those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected"* (p. 118).

To meet this requirement, NCDPI's scoring vendor, QAI, had their Scoring Directors carefully select and train only the most qualified people to be Team Leaders. The Team Leaders were trained prior to scorers, so they were familiar with all of the training materials and the scoring procedures prior to scorer training.

Scorers were divided into teams, and each scorer was assigned a unique scorer identification number. That identification number allowed for the tracking of scorer performance via the scorer quality control reports throughout the online scoring.

Once the training staff was confident that the scorers understood and had an awareness of the need to be sensitive to the performances of students, nondisclosure forms were signed, and training began.

Scorers, like Team Leaders, were required to meet the qualification standards before scoring student responses. Any scorer who was unable to meet the qualifying standards was dismissed–a stipulation understood by all scorers when they are hired. The qualification standard entailed 80% exact agreement on rubrics. Prior to actual scoring, the scorers did the following:

- signed a nondisclosure agreement
- acknowledged the QAI harassment policy
- reviewed NCDPI expectations and goals
- set aside any biases they may have about students, student work, and the scoring criteria presented
- trained to use the Score Point online scoring system

Once scorers were instructed on the above, individual training included the following process:

- Scorers were trained on the Scoring Guide, including discussion of the rubric, presenting the task or item (i.e., graphics and all related assets) and reviewing the eligible score points, followed by group participation and discussion of each response using examples and annotations as appropriate. Questions by scorers were addressed as a group for consistent messaging and decisions.
- Scorers then completed a training set independently to assess their grasp of the scoring.
- Each response in the training set was reviewed with the group with an explanation and examples as needed to ensure scorer consistency on the nuances of each response and score point.
- Scorers completed a qualifying set independently. Results using the qualification criteria determined if they were allowed to score that particular task type.
- In addition, each nonscorable code was explained, and examples were provided as available. All nonscorable answers were assigned a code. Examples included blank (BL), illegible (IL), foreign language (FL), repeating prompt (RP), off topic (OT), incoherent (IC), and other reasons (OR).
- Protocol for "alerting" responses that require attention was discussed at this time.

Following the successful completion of training and qualifying, scoring center staff activated individual scorers in the system, allowing them to score student responses.

3. Qualification

In order to score an item, the scorer had to meet the qualifications standards for scoring. The qualification standard for all items was 80% exact agreement. Successful completion of training also requires a minimum acceptable agreement rate of 80% on the task. A scorer can be dismissed if retraining does not elicit satisfactory results or if it is determined that a scorer is not accurately scoring student responses.

Monitoring the Scoring Process

Scoring Directors and Team Leaders live monitor the scoring process in terms of valid responses, ongoing training, one-on-one discussion, and read behinds. There are two kinds of read behinds used: random read behinds and prescribed read behinds. The random read behinds are a part of the daily ongoing monitoring process, while prescribed read behinds are done in case something arises during the scoring. The read behinds may result in a change in a student's score. QAI also produces item reliability and score point distribution reports weekly as a part of monitoring reliability and validity of the scoring. The report includes the number of responses scored, agreement rates, and score distribution.

Inter-rater Agreement

There were three operational CR items in each form. The NCDPI requires 10% of the random responses receive two readings as a part of the inter-rater agreement calculation. Table 6.3 shows exact and adjacent agreement rates for the operational English II CR items for the 2020–21 administration. The results indicate that the exact agreement rates by item range from 86.4% to 94.9% for students who took the test online and 100% for students who took the tests on paper/pencil format with exact and adjacent agreement rates of 99.4% or higher. These high agreement rates add to the validity of the English II scores.

Table 6.3 Rater Agreement Rates by Administration and Mode, 2020–21

Agreement Rates (%), Online						Agreement Rates (%), Paper			
Form	Item	N	Exact	Adjacent	Exact and Adjacent	Form	Item	N	Exact
M	#1	3615	89.7	9.8	99.5	A	#1	75	100
M	#2	3553	92.7	7.1	99.8	A	#2	73	100
M	#3	3542	87.9	11.8	99.7	A	#3	72	100
N	#1	3506	89.9	9.5	99.4	B	#1	68	100
N	#2	3500	86.7	13	99.7	B	#2	70	100
N	#3	3471	86.4	13.4	99.8	B	#3	71	100
O	#1	3380	94.9	4.6	99.5				
O	#2	3439	87.3	11.4	98.7				
O	#3	3466	93.1	6.7	99.8				

6.9 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 key and 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 scores

students' responses and compares their results to the final reported score. The goal is to have a 100% agreement rate between scores from the official scoring software and the independent check.

The second step involves reviewing score distributions, distractor analysis by plotting response probability of each answer option against number correct score making sure that the response probability for the correct answer increases for higher ability students, and review of CTT item statistics. For repeat forms a residual analysis is also performed to check if the item is maintaining its base year statistical property. In 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. This also results in rescoring for all students who took the affected form.

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 PSU 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 rationale and procedures used for establishing cut score should be documented*" (p. 107). For the first operational administration of the 2020–21 *Edition 5* EOG Reading and EOC English II 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 evidence in summarized version of the *Edition 5* Reading and English II final standard setting are presented in the ensuing sections.

7.1 Standard Setting Activities

English II: On August 4–6, 2020, a committee of 14 North Carolina educators participated in a **virtual standard setting** for the North Carolina's EOC English II test. The achievement standards were approved by the North Carolina State Board of Education on September 10, 2020.

Grades 3–8 Reading: On July 12–16, 2021, 38 North Carolina educators participated in an **in-person standard setting** for the North Carolina's EOG grades 3–8 Reading tests. The achievement standards were approved by the North Carolina State Board of Education on August, 2021.

The purpose of the standard settings 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. The item mapping procedure (Lewis, Green, Mitzel, Baum &

[1]Copyright © 2019 Data Recognition Corp.

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. The full standard setting technical reports produced by DRC[1] can be found in *Appendix 7-A* for English II and *Appendix 7-B* for general EOGs.

7.1.1 Participants' Characteristics

English II: Of the 14 participants, approximately 71% (10) were general education teachers, 21% (3) were special education teachers (with EC certification), and 7% (1) curriculum staff member. In terms of education, 50% had more than 15 years in education, 50% held a bachelor's degree, 50% held a master's degree or higher. Gender wise, approximately 79% were female and 21% were male. About 57% identified as white, 21% black, and 21% being of two or more races. Similarly, of the 14 participants, 57% worked in school districts in rural areas, 29% in suburban areas, and 14% in urban areas.

The EOC English II examination is usually administered twice a year—once in the fall and once in the spring—so students may test when they complete a course aligned to the new North Carolina English II content standards. The first English II examination was administered in Fall 2019 for students who completed the course in Fall 2019 semester. However, due to the COVID-19 pandemic, all Spring 2020 test administrations were canceled. Since Fall 2019 student counts were closely representative of the total student population and the test administration was before the pandemic, the EOC English II standard setting used item parameters computed from the 2019 Fall operational administration data for the standard setting.

Grades 3–8 Reading: Of the 38 participants, approximately 50% came from rural communities (50%), with the remaining from suburban (26%) or urban (24%) communities. Similarly, approximately half (51%) had 16 years or more of experience in education. All participants held a bachelor's degree or higher with 62.8% held a master's degree or higher. Most of the participants (93%) were female. Two-thirds of the participants (69%) identified as white, one-quarter (24%) as black, and the remainder (7%) as other ethnicity.

The EOG grades 3–8 Reading examination is administered in the spring. Due to the COVID-19 pandemic, all Spring 2020 test administrations were canceled. The Spring 2021 data were post-pandemic with variation in instructional practices. Therefore, item parameters from Spring 2019 field-test were used for the standard setting.

Given the circumstances surrounding the summer 2021 standard setting, the results for both EOG and EOC will be evaluated in future administrations.

7.1.2 Opening Session and Introductions

For English II standard setting, general student population, all participants began the workshop with a virtual opening session led by the NCDPI. For grades 3–8 Reading standard setting, on each workshop day, all participants for both general and alternate assessments began the workshop with a single opening session led by the NCDPI. During these opening sessions, the

NCDPI's chief of Test Development welcomed the participants to the workshop and described the purpose of the workshop and subsequently 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, each grade level panel spent the remainder of the day discussing 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 panelists 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 Reading and English II 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 7.1*) 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 Reading and English II.

Table 7.1 Policy Achievement Level Descriptors (ALDs) for General Reading

Not Proficient	Level 3	Level 4	Level 5
Students who are not proficient 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 Reports.

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 reading 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 RP67GA to the Ordered Item Booklets for the general reading tests, as this criterion allowed for OIBs to be constructed that included a selection of easy, medium, and difficult items. The RP67GA implies that a minimally competent examinee (MCE) should have at least a 67 percent chance of getting the items correct if the items are before the bookmark and a less than 67 percent chance of getting the items correct if the items are after the bookmark.

The North Carolina educators and stakeholders worked individually and collectively to recommend achievement standards for the North Carolina English II and Reading tests. The achievement standards for the English II were approved by the North Carolina State Board of Education on September 10, 2020 and for grades 3–8 Reading on August 5, 2021.

7.1.5 Across-Grade Articulation and Final ALD Cuts

The across-grade articulation of the achievement standards occurred after the grades 3–8 reading standard setting. The impact data from 2018–19 (pre-pandemic) and 2020–21 were presented to the participants. 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. *Table 7.2* shows the final approved ALD cuts for the North Carolina English II and grades 3–8 Reading.

Table 7.2 Final Cuts and Proficiency Distributions

Grade	Cuts			Proficiency Distributions			
	Level 3	Level 4	Level 5	Not Proficient	Level 3	Level 4	Level 5
English II	549	555	565	42.4%	23.5%	28.3%	5.8%
3	540	546	551	56.0%	18.7%	15.2%	10.1%
4	544	548	556	54.9%	14.1%	22.5%	8.5%
5	550	554	560	57.6%	13.4%	18.0%	11.0%
6	552	558	567	54.6%	21.6%	18.3%	5.5%
7	554	559	566	53.2%	18.6%	17.6%	10.5%
8	557	563	572	51.7%	20.7%	21.7%	5.9%

The raw score ranges for the proficiency levels are shown in *Table 7.3*. Notice that, for some forms in a grade the raw score ranges are different indicating some forms are slightly easier or difficult than others. This 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 7.3 Raw Score Ranges Across Proficiency Levels, 2020–21

Grade/Level	Forms	Not Proficient		Level 3		Level 4		Level 5	
		Min	Max	Min	Max	Min	Max	Min	Max
3	B	0	22	23	28	29	32	33	40

	C	0	22	23	28	29	**33**	**34**	40
	D	0	22	23	28	29	32	33	40
4	A	0	24	25	28	29	34	35	40
	B	0	**25**	**26**	29	30	34	35	40
5	A	0	24	25	28	29	33	34	40
	B	0	24	25	28	29	33	34	40
	C	0	25	26	29	30	33	34	40
6	A	0	22	23	29	30	37	38	44
	B	0	22	23	29	30	**36**	**37**	44
	C	0	**23**	**24**	30	31	37	38	44
7	A	0	25	26	30	31	36	37	44
	B	0	25	26	30	31	36	37	44
8	A	0	23	24	29	30	36	37	44
	B	0	23	24	29	30	36	37	44
English II	M	0	26	27	34	35	46	47	54
	N	0	26	27	35	36	**47**	**48**	54
	O	0	26	27	35	36	46	47	54

7.2 Evaluation of the Standard Setting Workshop

Since the 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 7.4* for English II and *Table 7.5* for grades 3–8 Reading. 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% in English II and 97% in grades 3–8 Reading 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 7.4 Standard Setting Workshop Evaluation Results, English II

Statement	Strongly Disagree	Disagree	Agree	Strongly Agree	Agree + Strongly Agree

During the workshop, my opinions were considered.	0%	0%	0%	100%	100%
The facilitator provided clear instructions.	0%	0%	0%	100%	100%
The descriptions of the threshold students were useful during the process.	0%	0%	14%	86%	100%
The achievement standards represent a reasonable profile of achievement at each level.	0%	0%	36%	64%	100%
My group's work was reflected in the presentation of recommendations.	0%	0%	0%	100%	100%
Overall, I valued the workshop as a professional development experience.	0%	0%	7%	93%	100%

Table 7.5 Standard Setting Workshop Evaluation Results, Grades 3–8 Reading

Statement	Strongly Disagree	Disagree	Agree	Strongly Agree	Agree + Strongly Agree
The achievement standards represent a reasonable profile of performance at each level.	0%	5%	37%	58%	95%
The facilitator provided clear instructions.	0%	0%	26%	74%	100%
The threshold students were useful during the process.	0%	3%	39%	58%	97%
I believe this process will yield defensible cut scores.	0%	8%	21%	71%	91%
My opinions were valued by my group.	0%	0%	18%	82%	100%
Overall, I valued the workshop as a professional development experience.	0%	0%	8%	92%	100%

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 standard setting workshop for both English II held on August 4–6, 2020 and grades 3–8 Reading held on July 12–16, 2021. Dr. Cizek is an expert in the field and is also a member of the North Carolina Technical Advisory Committee. His report regarding the standard setting workshop is summarized below. The detail report is available in *Appendix 7-C* for English II and in *Appendix 7-D* for grades 3–8 Reading.

For both workshops, 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. Overall, Dr. Cizek observed no issues of concern during the standard setting process. No events occurred that would weaken confidence in the validity of the panelists' recommendations.

For English II, Dr. Cizek concluded that "Overall, the workshop produced well-articulated ALDs and cut score recommendations that can be considered to be valid and reliable estimates of appropriate performance standards for the English II assessment."

Similarly, for grades 3–8 Reading Dr. Cizek indicated that "overall, the workshops followed best practices in the area of standard setting; the procedures as implemented followed the plan that had been reviewed and approved by the state's technical advisors; and the activities produced cut score recommendations that reliably and accurately reflect the intended performance expectations for North Carolina students and the expert content judgments of North Carolina educators."

CHAPTER 8 2020–21 TEST RESULTS AND REPORTS

The instructional and assessment contexts surrounding the 2020–21 school year in North Carolina varied in terms of instructional practices, for example, in-person, remote, and mixed instructional format; waiver for testing in 2019–20 and accountability reporting in 2020–21; and varying participation rates across schools and districts. Therefore, the NCDPI urges caution for interpreting summary results presented in this chapter for comparison. Furthermore, one should be cautious for referencing results from 2020–21 in future administrations as the contexts are likely to vary.

With the above context, this chapter documents test level summary results for the EOG Reading and EOC English II tests based on reported scale scores and achievement levels from 2020–21 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 Reading and English II assessments from 2020–21 are summarized in *Figure 8.1* through *Figure 8.7*. These scores are based on results from all eligible EOG and EOC general administration and students with approved NCDPI accommodations such as braille, large print, read-aloud and extended time.

The population scale score means for EOG grades 3–8 Reading and EOC English II are set to be 538, 542, 547, 550, 552, 556, and 550 with a standard deviation of 10 as scaling constants to transform the theta scale to score scale. Note that the *Edition 5* scale scores are grade specific and are not reported in 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. The results show that the scale score distributions for the 2020–21 population have similar distributional properties as the scaling parameters.

The grade 3 results are tied to North Carolina's Read to Achieve legislative initiative. Under the initiative, third-grade students who are not reading at grade level by the end of third grade receive special help, including summer reading camp and other interventions to make sure that they can read well enough to be able to do fourth-grade work. With the obligation to report grade 3 results by the end of Spring, the NCDPI linked *Edition 5* tests to *Edition 4* scale for reporting. Note that the standard setting occurred in the Summer of 2021. The results presented below for grade 3 are based on Edition 4 scale. New scale will be implemented from 2021–22 administration and beyond.

Figure 8.1 Grade 3 Reading Scale Score Distribution, Spring 2021

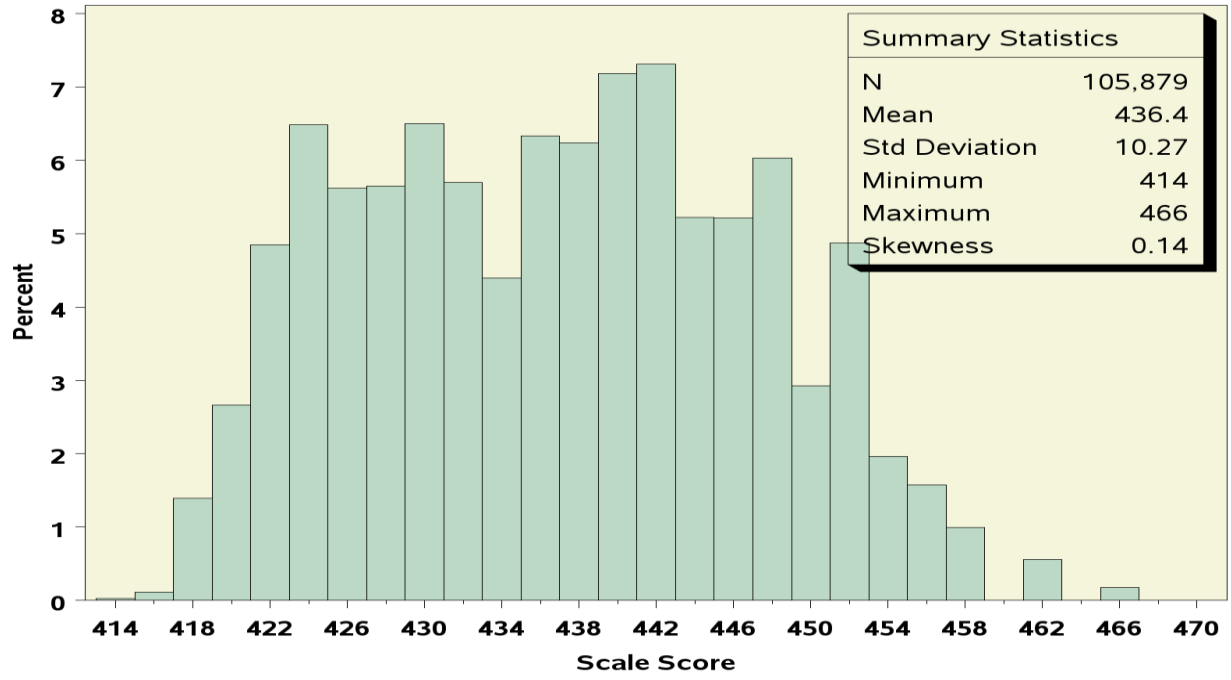


Figure 8.2 Grade 4 Reading Scale Score Distribution, Spring 2021

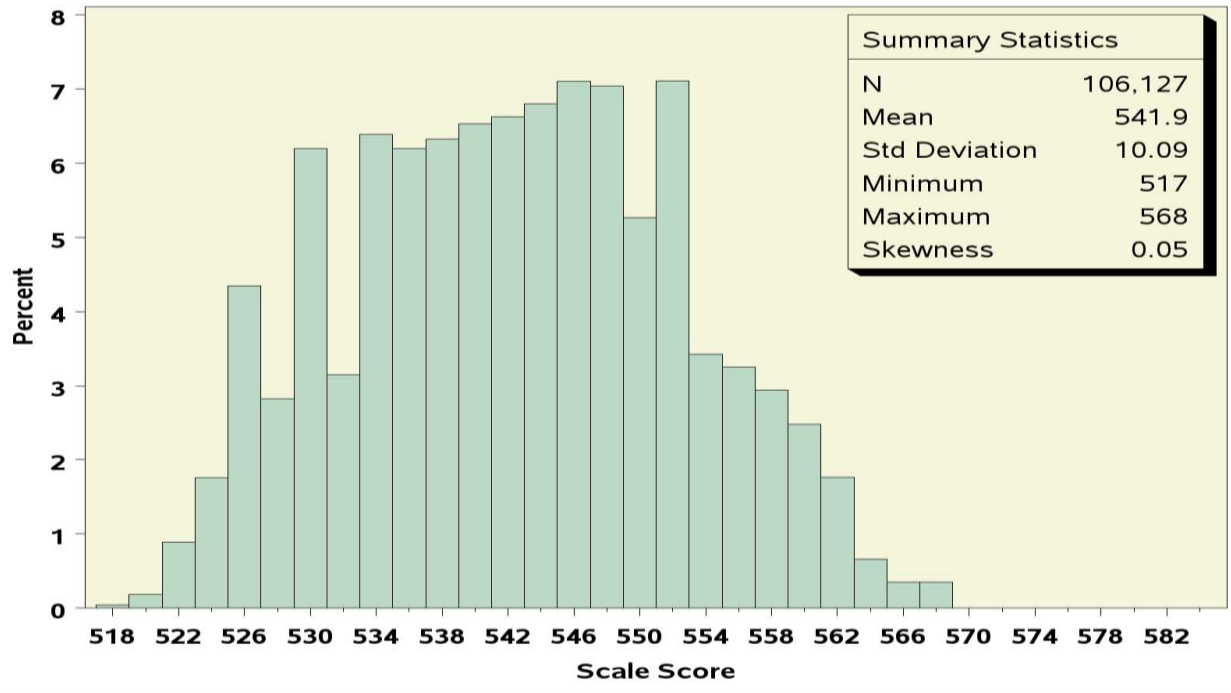


Figure 8.3 Grade 5 Reading Scale Score Distribution, Spring 2021

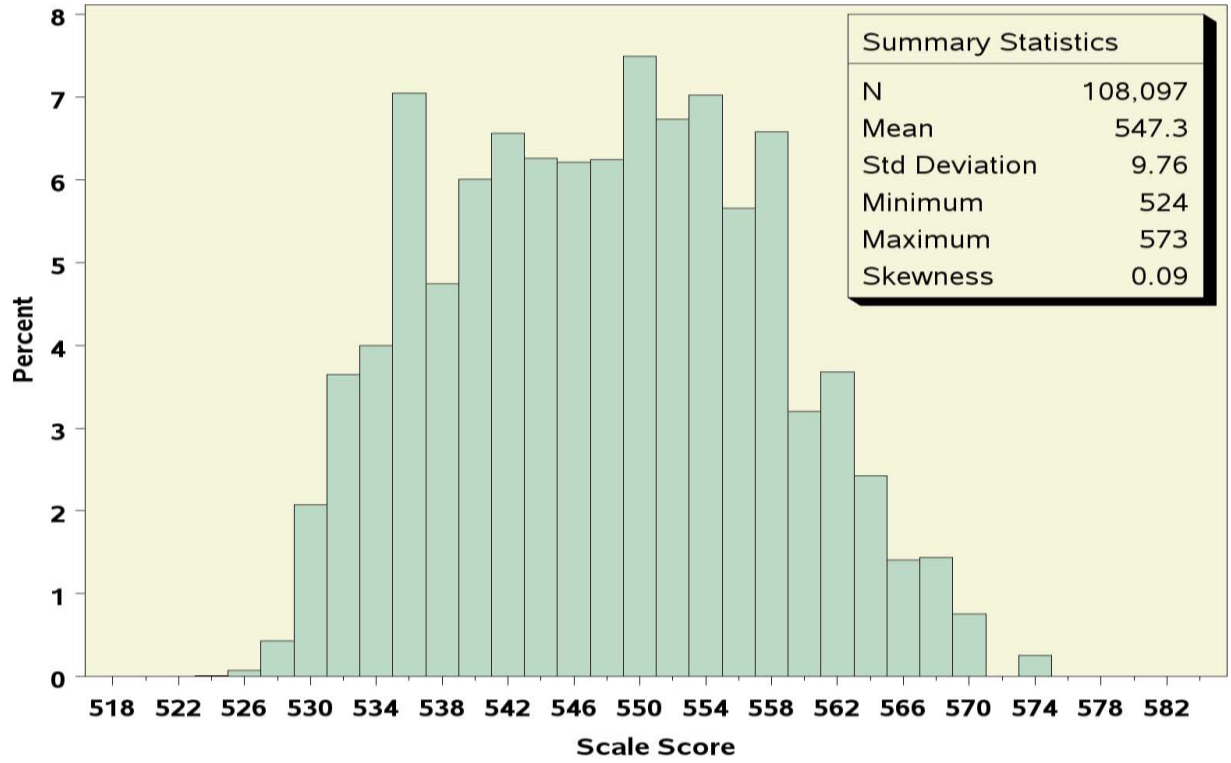


Figure 8.4 Grade 6 Reading Scale Score Distribution, Spring 2021

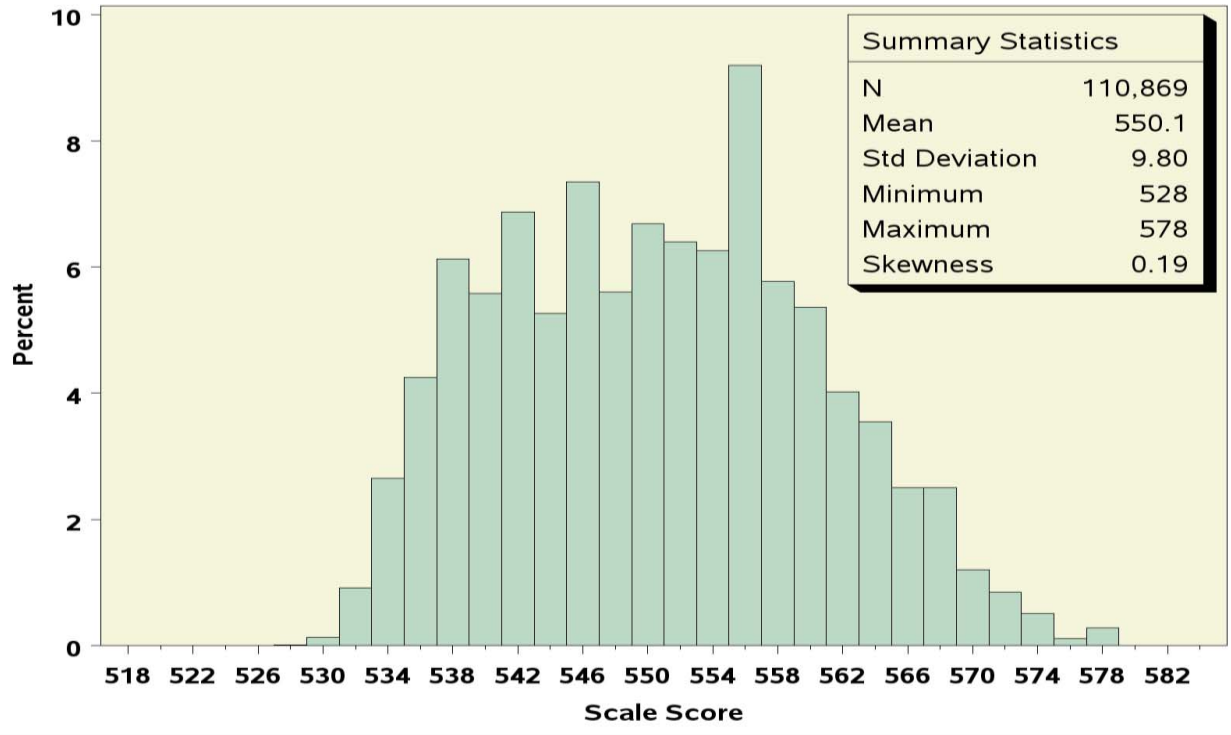


Figure 8.5 Grade 7 Reading Scale Score Distribution, Spring 2021

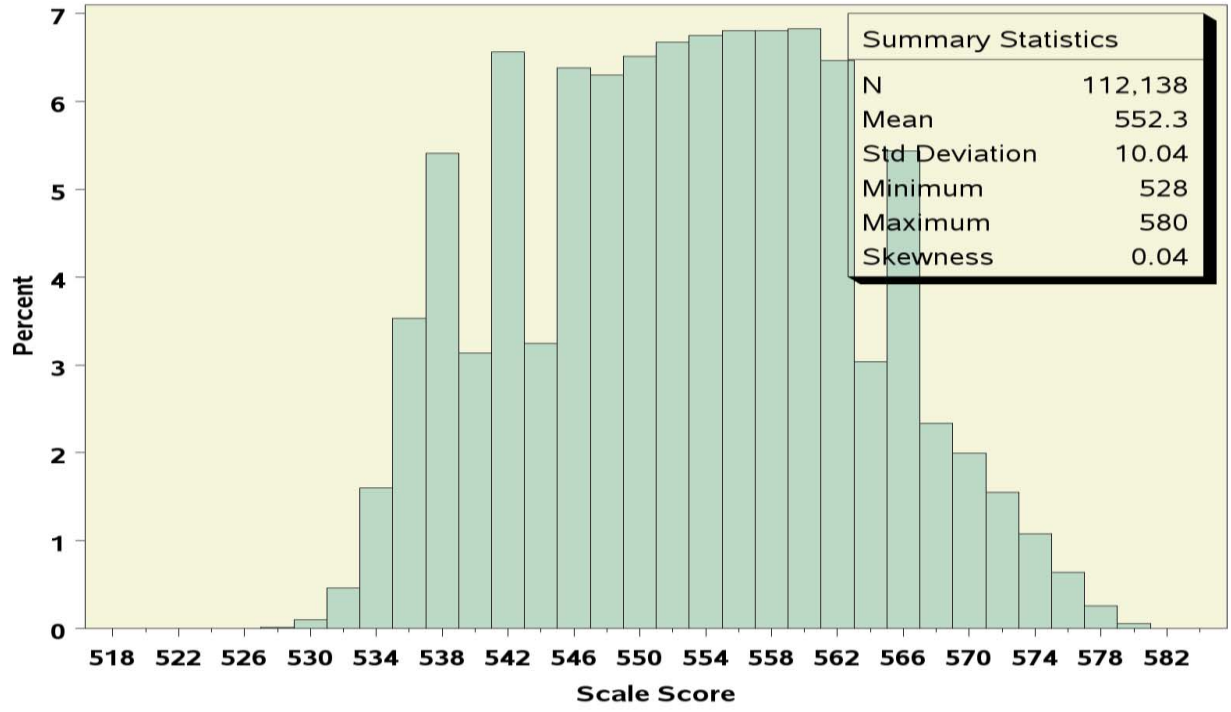


Figure 8.6 Grade 8 Reading Scale Score Distribution, Spring 2021

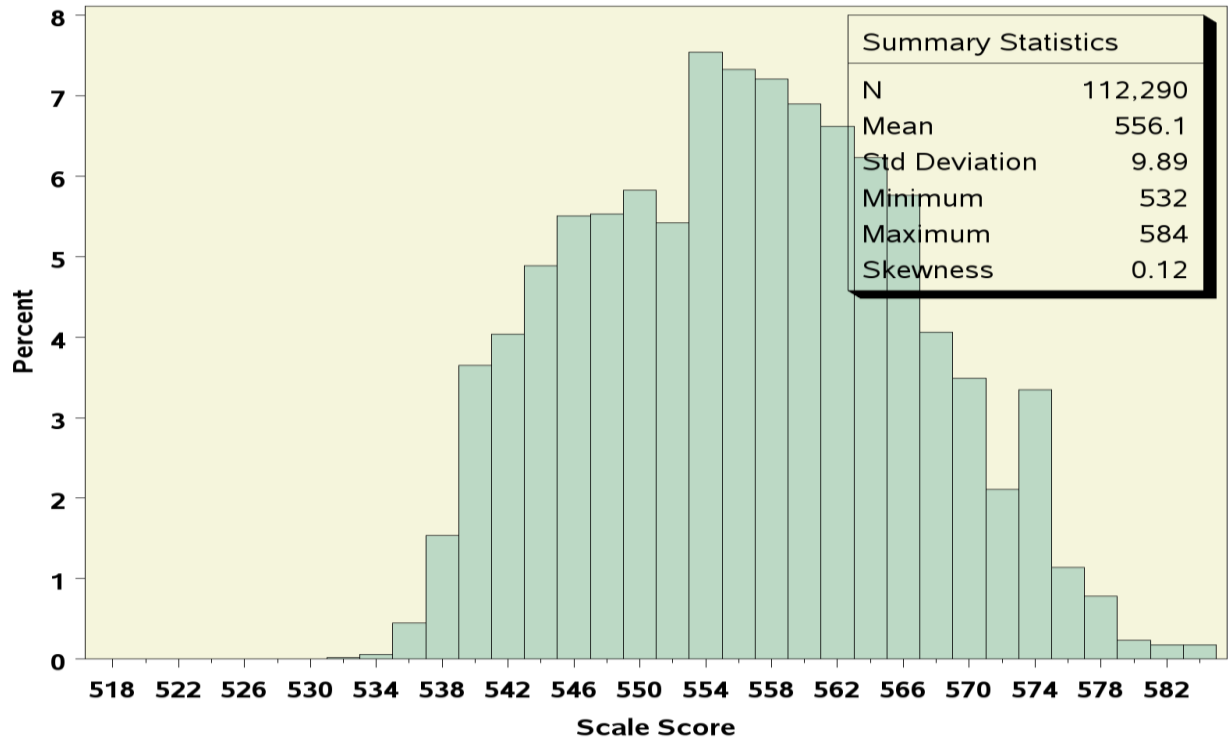
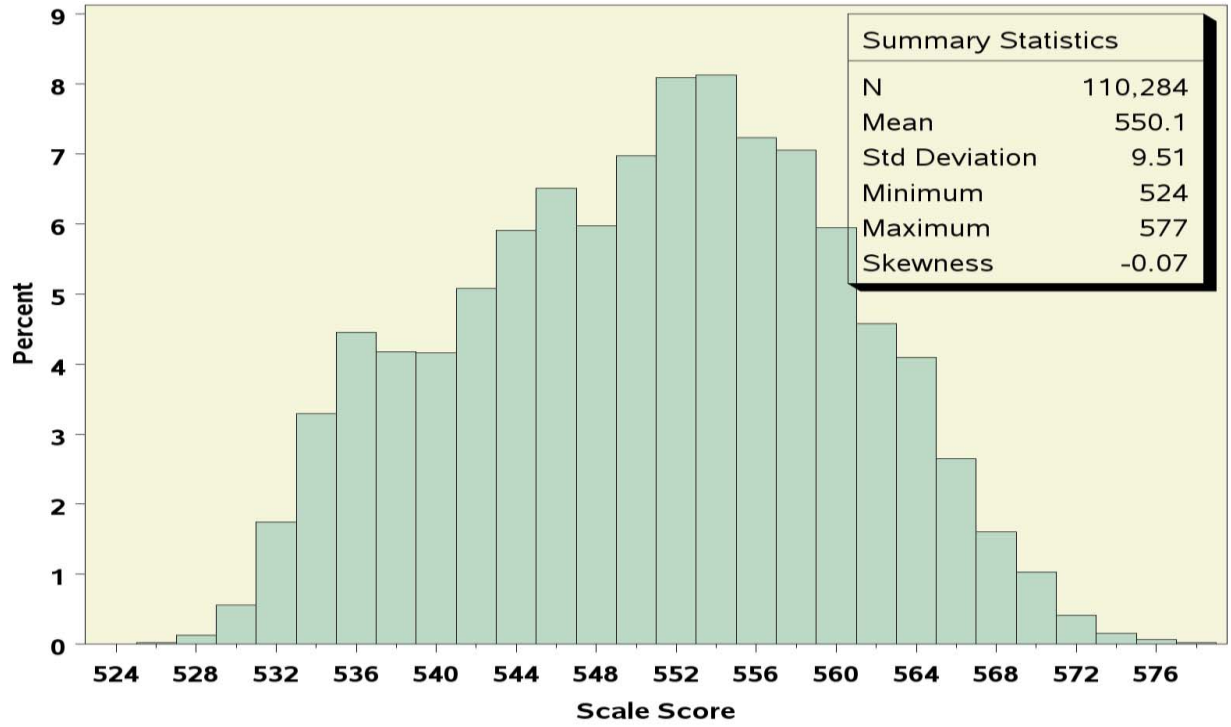


Figure 8.7 English II Scale Score Distribution, 2020–21



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. *Table 8.1* through *Table 8.3* show the summary score distributions for the EOG Reading and English II assessments from 2020–21 administration by major accommodation subgroups described in Section 5.5. Read aloud, either computer or sign language interpretation or teacher read, is not allowed in Reading tests.

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 English II. "Regular Administration" in these tables refer to students who did not receive any NCDPI approved accommodations. Each accommodation category includes all students who received one or more accommodation classified under Section 5.5.

Table 8.1 Grades 3–5 Reading Scale Score by Accommodation Subgroups, Spring 2021

Grade	Subgroups	N	Statistics		Range		Percentile		
			Mean	SD	Min	Max	25th	Median	75th
3	1 Regular Administration	96,232	437.2	10.2	414	466	429	438	445
	2 Assistive Devices	292	428.3	8.9	416	454	422	426	434
	3 Special Environment	8,994	428.6	7.9	414	465	423	427	433
	4 Special Print	246	431.3	9.7	415	461	424	428	438
4	1 Regular Administration	95,759	542.8	9.9	517	568	535	543	550
	2 Assistive Devices	389	533.2	9.1	517	561	526	531	539
	3 Special Environment	9,607	533.8	8.4	517	568	527	533	539
	4 Special Print	247	535.3	10	518	564	527	533	541
5	1 Regular Administration	97,644	548.1	9.6	524	573	541	549	555
	2 Assistive Devices	451	538.4	8.1	524	570	533	536	542
	3 Special Environment	9,638	539.6	8	524	570	533	538	544
	4 Special Print	245	541.2	9.1	525	567	535	540	547

Table 8.2 Grades 6–8 Reading Scale Score by Accommodation Subgroups, Spring 2021

Grade	Subgroups	N	Statistics		Range		Percentile		
			Mean	SD	Min	Max	25th	Median	75th
6	1 Regular Administration	100,930	550.9	9.7	528	578	543	551	558
	2 Assistive Devices	705	540	6.9	529	578	536	538	542
	3 Special Environment	9,093	542.7	7.8	528	578	537	541	547
	4 Special Print	126	547.3	9.7	530	573	539	548	556
7	1 Regular Administration	101,810	553.1	9.9	528	580	546	553	560
	2 Assistive Devices	800	541.4	7.4	530	575	536	540	545
	3 Special Environment	8,971	544.5	8.4	528	578	538	542	549
	4 Special Print	464	550.7	9.8	530	578	543	550	558
8	1 Regular Administration	102,741	556.8	9.7	532	584	549	557	564
	2 Assistive Devices	822	545	6.6	534	576	540	544	548
	3 Special Environment	8,146	548.6	8.3	532	584	542	547	554
	4 Special Print	471	554.1	9.8	532	584	547	554	561

Table 8.3 English II Scale Score by Accommodation Subgroups, 2020–21

Subject	Subgroups	N	Statistics		Range		Percentile		
			Mean	SD	Min	Max	25th	Median	75th
English II	1 Regular Administration	101,375	550.8	9.3	524	577	544	551	557
	2 Assistive Devices	1,267	537.6	6.2	525	570	534	536	540
	3 Special Environment	7,210	542.5	8.7	526	575	536	541	548
	4 Special Print	231	547.6	9.2	530	572	541	547	555

8.1.2 Scale Score by Gender

Table 8.4 through *Table 8.6* summarize EOG and EOC scale score by gender. In all grade levels, male students were a slightly larger proportion (about 51%) of students who took EOG and EOC in North Carolina during 2020–21 school year. Scale score distributions are similar between female and male students for the most part with female students on average scoring higher than male students ranging from 0.8 scale score point in grade 5 to 2.4 scale score points in English II.

Table 8.4 Grades 3–5 Reading Scale Score Descriptive Summary by Gender, Spring 2021

Grade	Gender	N	Statistics		Range		Percentile		
			Mean	SD	Min	Max	25th	Median	75th
3	Female	51,854	436.9	10.1	414	466	429	437	445
	Male	53,910	436.0	10.4	414	466	427	436	444
4	Female	51,851	542.5	9.9	517	568	535	543	550
	Male	54,151	541.4	10.2	517	568	533	541	549
5	Female	52,929	547.7	9.5	524	573	540	548	555
	Male	55,049	546.9	10.0	524	573	539	547	554

Table 8.5 Grades 6–8 Reading Scale Score Descriptive Summary by Gender, Spring 2021

Grade	Gender	N	Statistics		Range		Percentile		
			Mean	SD	Min	Max	25th	Median	75th
6	Female	53,989	551.0	9.7	528	578	543	551	558
	Male	56,865	549.3	9.8	528	578	541	549	556
7	Female	54,837	553.2	9.8	528	580	546	553	560
	Male	57,208	551.5	10.2	528	580	543	552	559
8	Female	54,765	557.0	9.7	532	584	549	557	564
	Male	57,415	555.2	10.0	532	584	547	555	563

Table 8.6 EOC English II Scale Score Descriptive Summary by Gender, 2020–21

Grade	Gender	N	Statistics		Range		Percentile		
			Mean	SD	Min	Max	25th	Median	75th
English II	Female	54,353	551.3	9.2	525	577	545	552	558
	Male	55,730	548.9	9.6	524	577	541	549	556

8.1.3 Scale Score by Major Ethnic Groups

Table 8.7 through *Table 8.9* show the breakdown of EOG and EOC scale scores by major reportable ethnic groups from 2020–21 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 45% to 49% of all students across all levels, Black students representing about 24% to 25%, and Hispanic students making about 19% to 20%. 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 between the two ethnic groups ranged from 0.5 to 0.7 standard deviation across all EOG and EOC.

Table 8.7 Grades 3–4 Reading Scale Score Descriptive Summary by Ethnicity, Spring 2021

Grade	Ethnicity	N	Statistics	Range	Percentile

			Mean	SD	Min	Max	25th	Median	75th
3	1. Black	25,647	432.4	9.0	414	466	425	431	439
	2. Hispanic	21,234	433.0	9.3	414	466	426	431	440
	3. White	47,660	439.7	10.0	414	466	432	441	447
	4. Others	11,223	438.2	10.5	414	466	430	439	446
4	1. Black	25,805	537.8	8.9	517	568	530	537	544
	2. Hispanic	21,129	538.6	9.3	517	568	531	538	545
	3. White	47,885	545.2	9.7	517	568	539	546	552
	4. Others	11,183	543.6	10.4	517	568	536	544	551
5	1. Black	26,391	543.2	8.6	524	573	536	542	549
	2. Hispanic	22,034	544.1	9.0	524	573	537	543	551
	3. White	48,687	550.5	9.4	524	573	544	551	557
	4. Others	10,866	549.3	10.0	525	573	541	550	557

Table 8.8 Grades 6–8 Reading Scale Score Descriptive Summary by Ethnicity, Spring 2021

Grade	Ethnicity	N	Statistics		Range		Percentile		
			Mean	SD	Min	Max	25th	Median	75th
6	1. Black	27,594	546.4	8.6	528	578	540	545	553
	2. Hispanic	22,445	547.3	9.0	528	578	540	546	554
	3. White	49,884	553.0	9.6	528	578	546	554	560
	4. Others	10,931	552.4	10.4	529	578	544	553	560
7	1. Black	27,762	548.4	9.0	528	580	541	548	555
	2. Hispanic	22,408	549.6	9.4	528	578	542	549	557
	3. White	51,192	555.2	9.7	528	580	548	556	562
	4. Others	10,683	554.6	10.5	528	580	547	555	562
8	1. Black	27,072	552.2	8.7	532	584	546	551	558
	2. Hispanic	22,525	553.4	9.2	532	584	546	553	560
	3. White	51,934	558.8	9.7	532	584	552	559	566
	4. Others	10,649	558.3	10.5	532	584	550	558	566

Table 8.9 EOC English II Scale Score Descriptive Summary by Ethnicity, 2020–21

Grade	Ethnicity	N	Statistics		Range		Percentile		
			Mean	SD	Min	Max	25th	Median	75th

English II	1. Black	26,240	546.2	8.7	524	577	539	546	553
	2. Hispanic	20,714	547.4	9.0	524	577	540	548	554
	3. White	53,435	552.7	9.1	525	577	547	553	559
	4. Others	9,694	551.9	9.8	524	577	545	553	559

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 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 Achievement Levels Distribution

Beginning in 2020–21 with *Edition 5* of EOG and EOC, the NCDPI transitioned to classify and report student performance 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.

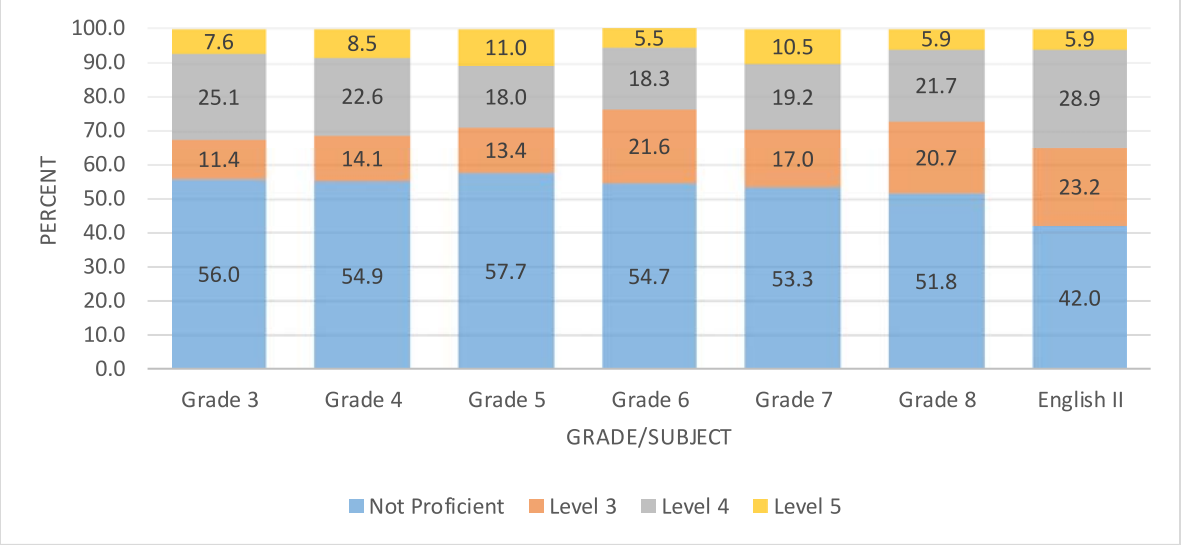
These policy descriptors are used to summarize performance expectations for students at each level. For a detailed explanation of what students in each performance level are expected to be able to do refer to the full achievement level descriptors in *Appendix 8-B* for English II and *Appendix 8-C* for grades 3–8 Reading. 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 or above are considered and reported to have met grade level performance expectations. Students classified as Level 4 or above are further designated to be on track to be Career-and-College Ready (CCR). This subset of Level 4 or above 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 NCSBE 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 8.8 shows the summary of proportion of students by achievement level classifications from the 2020–21 North Carolina EOG reading and EOC English II assessments. The stacked bar graph shows the distribution by grade or course. For example, in EOG grade 3, 56% students are classified as Not Proficient, 11.4% Level 3, 25.1% Level 4 and 7.6% Level 5. Also, for state accountability reporting purposes, 44.1% (Level 3 and up) of NC grade 3 students who took the EOG reading assessment are considered to have met grade level content expectations. While about 32.7% (Level 4 and up) 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-D*.

Figure 8.8 State Level Achievement Level Classifications (%) by Grade, 2020–21

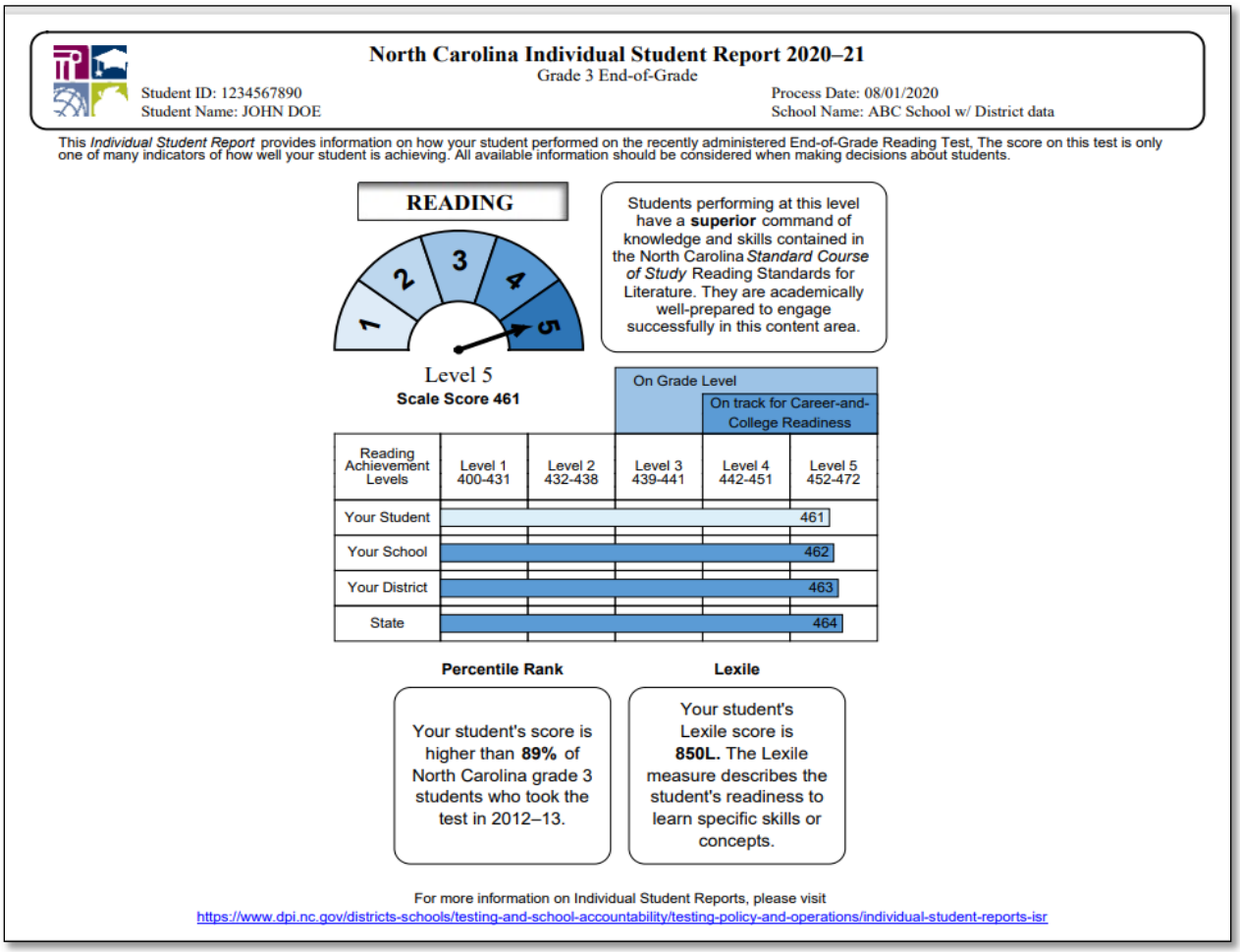


8.2 Score Reports

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 8.9* is designed to inform students and their parents/guardians 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 EOG and EOC tests.

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 8.9 Individual Student Report (ISR)



Additional customized reports for internal use only by districts and schools are generated aimed to provide teachers and school administrators with in-depth and disaggregated data of their students and school performance to help inform instructional practices. 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 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-E*). 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

WinScan Reports	Audience			
	Teacher	School	District	State
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 PSU, 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. PSU should not use private or personal accounts to store students' PII. PSUs who 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 FedEx). 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 EVIDENCE

This chapter presents additional validity evidence collected in support of the interpretation of Edition 5 EOG Reading and EOC English II test scores. The first two sections present validity evidence in support of the internal structure of the 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 the test scores. The sections towards the end document content validity evidence summarized from the alignment study and evidence based on relation to other variables summarized from the EOG/EOC Lexile framework linking study, while the last part presents summary of procedures used to ensure the assessments are accessible and fair for all students.

9.1 Reliability of the 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 grades 3–8 Reading and English II forms by grade and major demographic variables and overall for 2020–21 administration. Across all forms, overall reliability estimates based on the 2020–21 population ranged from 0.89 to 0.92. Reliability index for subgroups (gender and ethnicity) is also consistent with overall index across forms for the most part, they are consistently higher than the 0.85 threshold. The alpha for accommodation subgroups (EDS, SWD, and ELs) is lower, as low as 0.62 for EL students in grade 6 Form B. Note that the sample size for the accommodation subgroups was small and variation was low.

Table 9.1 EOG Reading Reliabilities (Alpha) by Form and Subgroup

Grade	Form	Gender		Ethnicity[1]			Accommodations[1]			All
		Female	Male	Black	Hispanic	White	EDS	SWD	ELs	
3	N	0.90	0.90	0.86	0.87	0.90	0.86	0.84	0.80	0.90
	O	0.91	0.91	0.88	0.88	0.91	0.88	0.86	0.83	0.91
	P	0.89	0.90	0.85	0.87	0.90	0.86	0.84	0.81	0.89
4	M	0.91	0.91	0.88	0.89	0.91	0.89	0.87	0.85	0.91
	N	0.90	0.90	0.87	0.88	0.90	0.87	0.85	0.82	0.90
5	M	0.90	0.90	0.86	0.88	0.90	0.87	0.80	0.73	0.90
	N	0.91	0.91	0.88	0.89	0.91	0.88	0.83	0.75	0.91
	O	0.90	0.91	0.88	0.89	0.90	0.88	0.84	0.79	0.90
6	M	0.90	0.90	0.86	0.87	0.90	0.86	0.78	0.67	0.90
	N	0.89	0.89	0.85	0.86	0.89	0.85	0.77	0.62	0.89
	O	0.89	0.90	0.86	0.87	0.90	0.86	0.78	0.67	0.89
7	M	0.89	0.90	0.86	0.88	0.89	0.87	0.80	0.75	0.89
	N	0.90	0.90	0.87	0.88	0.90	0.87	0.78	0.74	0.90
8	M	0.87	0.88	0.83	0.85	0.88	0.83	0.73	0.64	0.88
	N	0.89	0.89	0.86	0.87	0.89	0.86	0.77	0.70	0.89
English II	M	0.91	0.92	0.89	0.90	0.91	0.89	0.83	0.75	0.92
	N	0.92	0.92	0.90	0.90	0.91	0.90	0.84	0.79	0.92
	O	0.91	0.92	0.90	0.90	0.91	0.90	0.84	0.75	0.91

[1]Reliabilities estimates are displayed only for major ethnic groups and accommodations.

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 reading and EOC English II forms are shown in *Table 9.2*.

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 reading is 540 (see *Table 9.2*). A student who took Form A and scored 540 with a SE of 3 has a 68% probability that his or her true score or ability ranges from 537 to 543 (540±1x3) 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 534 to 546 (540±2x3). For most of the EOG and EOC scale score cuts in the middle range, particularly at the Level 3 and Level 4, the conditional standard errors are between 2 and 4. Cuts at the LOSS and HOSS have the conditional SEs between 5 and 6. 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 9.2 Conditional Standard Errors (SE) at Achievement Level Cuts by Form

Grade	Form	Min		Level 3		Level 4		Level 5		Max	
		LOSS	SE	Cut	SE	Cut	SE	Cut	SE	HOSS	SE
3	N	515	6	540	3	546	3	551	3	564	6
	O	517	6	540	3	546	3	551	3	564	6
	P	516	6	540	3	546	3	551	3	564	5
4	M	518	5	544	3	548	3	556	3	568	6
	N	517	5	544	3	548	3	556	4	568	6
5	M	525	5	550	3	554	3	560	3	573	5
	N	524	5	550	3	554	3	560	3	573	6
	O	524	5	550	3	554	3	560	3	573	6
6	M	529	6	552	3	558	3	567	3	578	6
	N	528	6	552	3	558	3	567	3	578	6
	O	528	6	552	3	558	3	567	3	578	6
7	M	528	6	554	3	559	3	566	3	580	6
	N	527	6	554	3	559	3	566	3	580	6
8	M	532	6	557	3	563	3	572	3	584	5
	N	532	6	557	3	563	3	572	3	584	6
English II	M	526	5	549	3	555	2	565	3	577	5
	N	525	5	549	3	555	2	565	3	577	5
	O	524	5	549	3	555	2	565	3	577	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). The classification consistency analyses for the North Carolina EOG Reading and EOC English II followed the HB procedures.

Table 9.3 shows the decision accuracy and consistency indexes for achievement levels at each grade. Overall, the values indicate good classification accuracy (ranging from 0.89 to 0.97) and consistency (from 0.85 to 0.95). For example, EOG grade 3 reading Form B has an accuracy rate of 0.92 at Level 3 cut which means if a student who is classified as Level 3 were to take a non-overlapping, equally difficult form a second time, there is a 92% (**bolded**) probability that the student would still be classified as Level 3. The higher classification consistency also entails smaller standard error and higher reliability.

[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.

Table 9.3 Classification Accuracy and Consistency Results, EOG and EOC Tests

Grade		Level 3		Level 4		Level 5	
		Acc.	Con.	Acc.	Con.	Acc.	Con.
3	N	**0.92**	0.89	0.94	0.91	0.95	0.93
	O	**0.93**	0.90	0.93	0.91	0.94	0.92
	P	**0.92**	0.88	0.93	0.91	0.95	0.93
4	M	0.92	0.89	0.92	0.89	0.94	0.91
	N	0.92	0.88	0.92	0.89	0.94	0.92
5	M	0.92	0.89	0.93	0.90	0.95	0.93
	N	0.92	0.89	0.93	0.90	0.94	0.92
	O	0.92	0.89	0.92	0.89	0.92	0.90
6	M	0.91	0.88	0.93	0.91	0.97	0.95
	N	0.91	0.88	0.93	0.91	0.96	0.95
	O	0.91	0.88	0.93	0.90	0.96	0.95
7	M	0.91	0.87	0.91	0.88	0.94	0.92
	N	0.91	0.88	0.92	0.89	0.94	0.92
8	M	0.89	0.85	0.92	0.89	0.96	0.95
	N	0.91	0.87	0.92	0.89	0.95	0.93
English II	M	0.92	0.89	0.93	0.90	0.96	0.94
	N	0.92	0.89	0.93	0.90	0.96	0.95
	O	0.92	0.89	0.93	0.90	0.95	0.94

Note: Acc. = Accuracy; Con. = Consistency

9.4 Unidimensionality of EOG and EOC Assessments

North Carolina EOG Reading and EOC English II assessments are designed based 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 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 9.7* show the PCA results for all reading assessments 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 9.4* through *Table 9.6*. Based on the PCA results, the average ratio of the first to the second eigenvalue across grades ranged from about 3.6 to about 9.6. Also, on average the first principal component accounts for about 24% to 38% of the total variance.

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 EOG Reading and EOC English II 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 6% or less total variance across all 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 Reading items at each test measured an overall Reading 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 plots 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 Reading and English II score using a unidimensional scale.

Figure 9.1 Grade 3 PCA Scree Plot and Cumulative Variance by Form

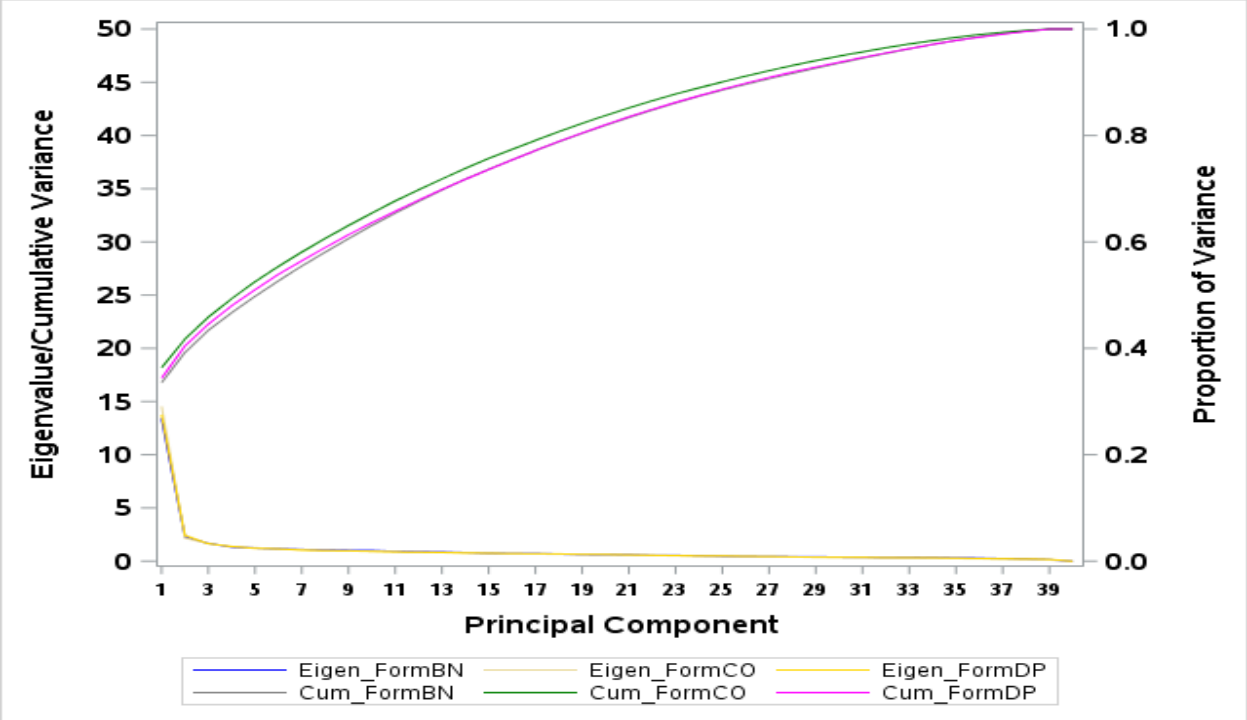


Figure 9.2 Grade 4 PCA Scree Plot and Cumulative Variance by Form

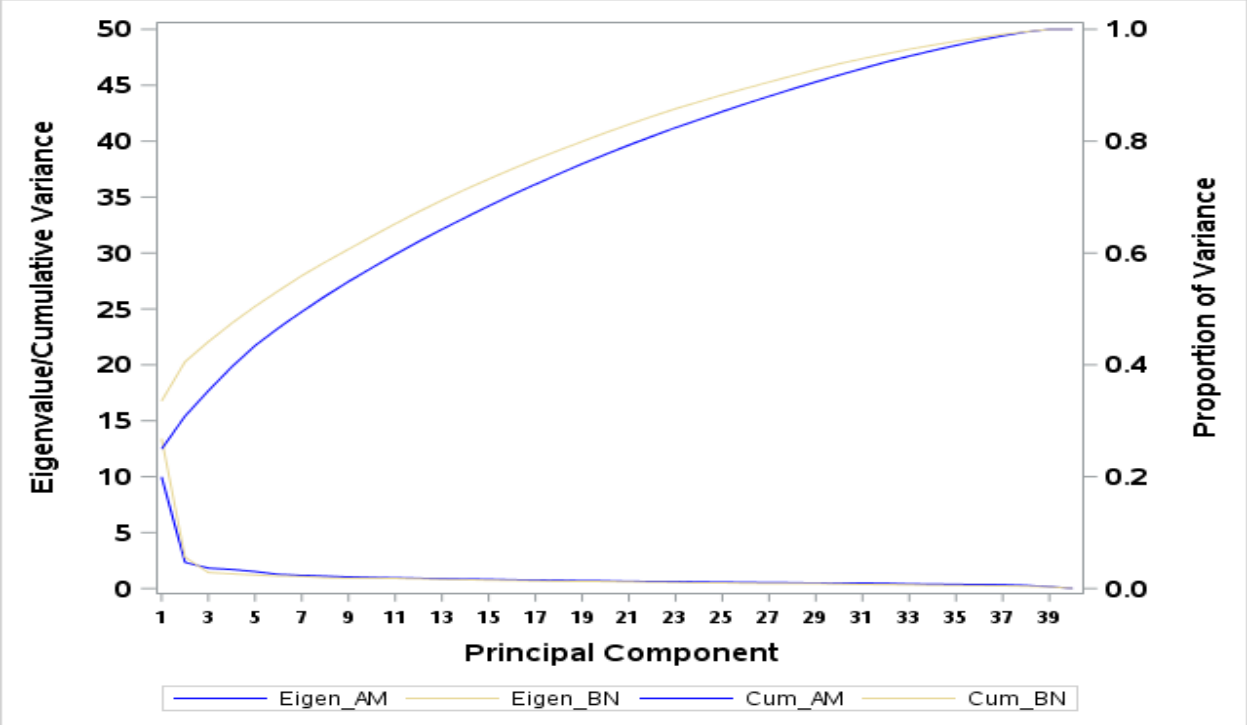


Figure 9.3 Grade 5 PCA Scree Plot and Cumulative Variance by Form

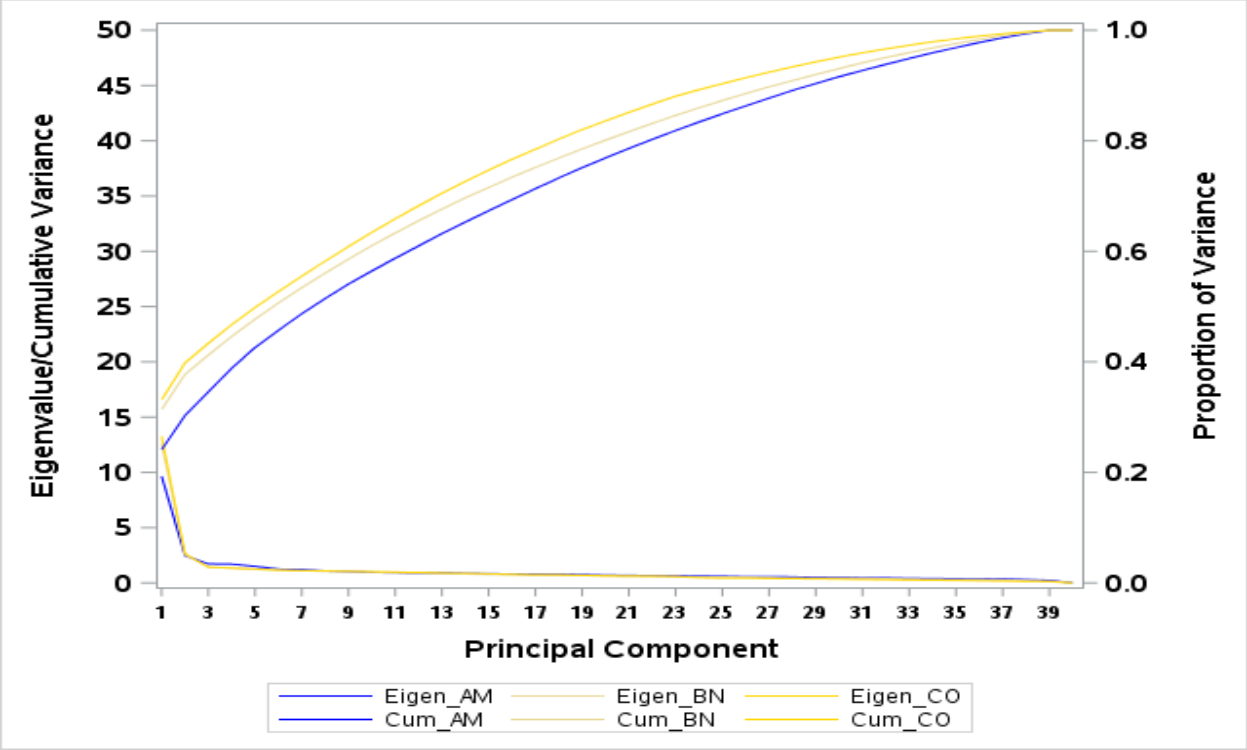


Figure 9.4 Grade 6 PCA Scree Plot and Cumulative Variance by Form

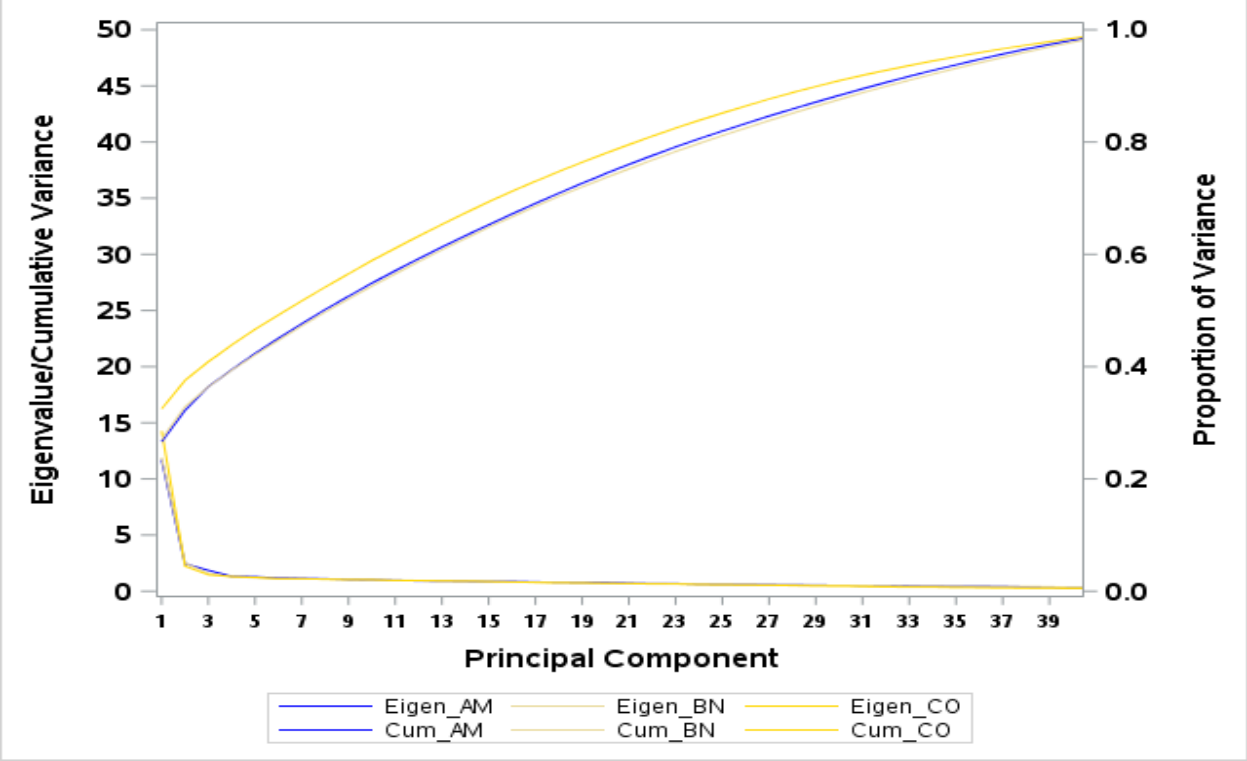


Figure 9.5 Grade 7 PCA Scree Plot and Cumulative Variance by Form

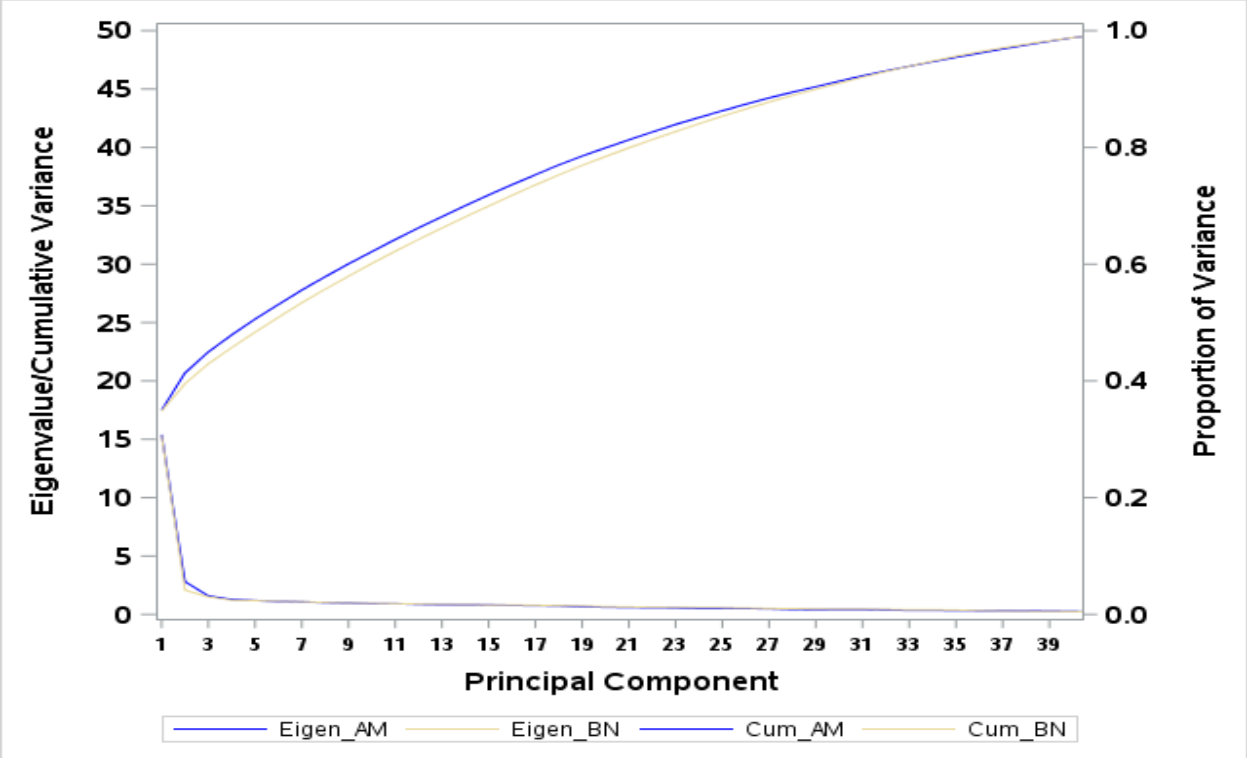


Figure 9.6 Grade 8 PCA Scree Plot and Cumulative Variance by Form

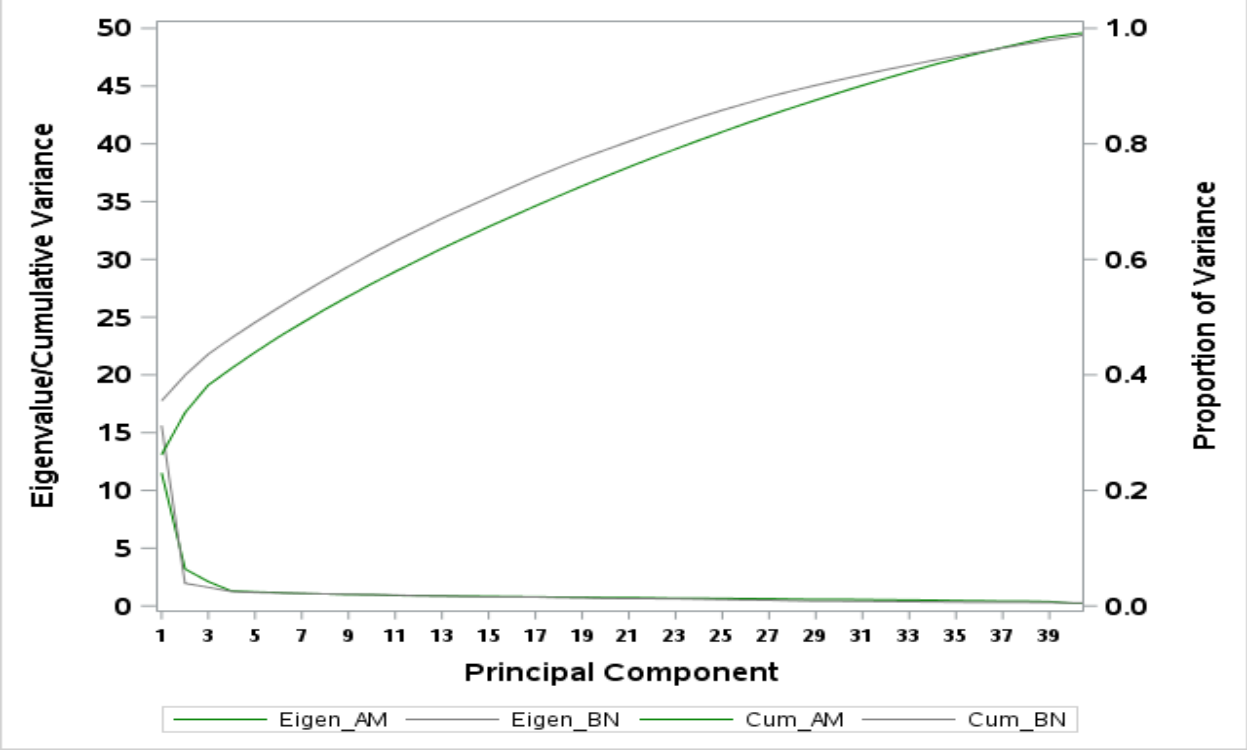


Figure 9.7 EOC English II PCA Scree Plot and Cumulative Variance by Form

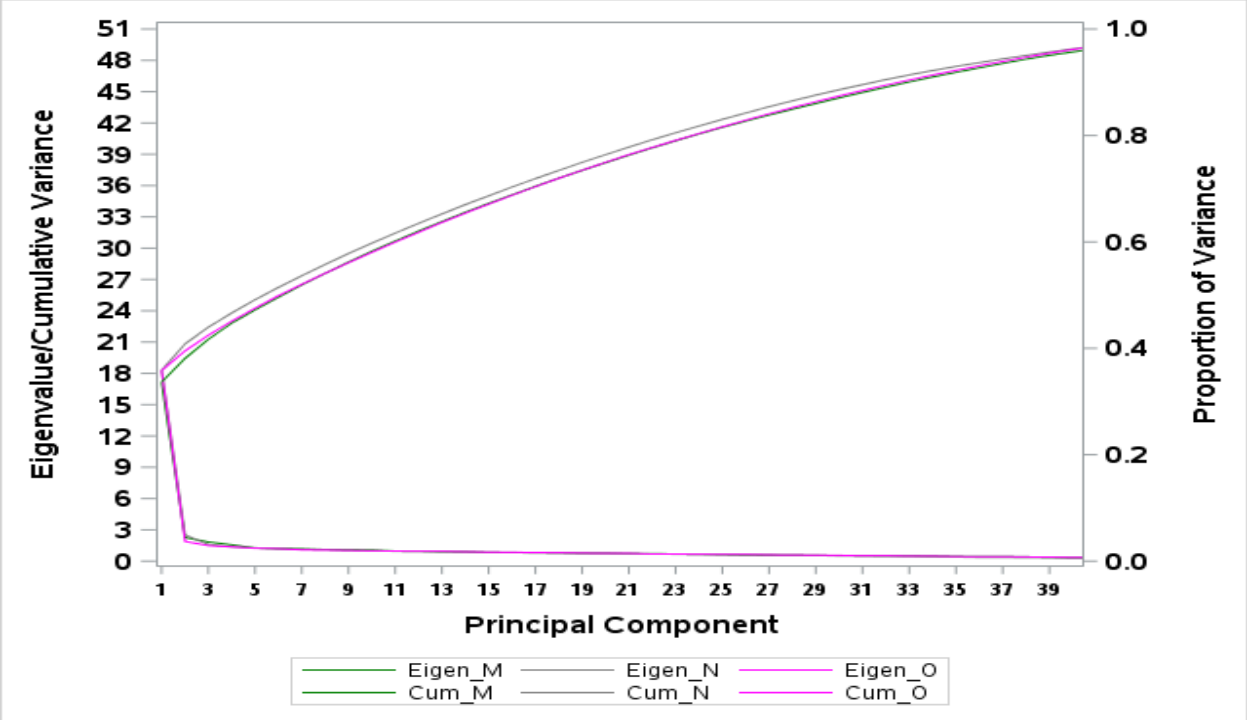


Table 9.4 Grades 3–5 Reading Principal Component and Variance by Form

Grade	Form	Component	Eigen	Variance	Cumulative Variance
3	M	1	13.4	34%	34%
		2	2.3	6%	39%
		3	1.7	4%	43%
	N	1	14.5	36%	36%
		2	2.2	5%	42%
		3	1.6	4%	46%
	O	1	13.8	34%	34%
		2	2.4	6%	40%
		3	1.6	4%	45%
4	M	1	10.0	25%	25%
		2	2.3	6%	31%
		3	1.8	5%	35%
	N	1	13.4	34%	34%
		2	2.8	7%	41%
		3	1.4	4%	44%
5	M	1	9.7	24%	24%
		2	2.5	6%	30%
		3	1.7	4%	35%
	N	1	12.6	31%	31%
		2	2.5	6%	38%
		3	1.4	4%	41%
	O	1	13.3	33%	33%
		2	2.6	7%	40%
		3	1.4	4%	43%

Table 9.5 Grades 6–8 Reading Principal Component and Variance by Form

Grade	Form	Component	Eigen	Variance	Cumulative Variance
6	M	1	11.7	27%	27%
		2	2.5	6%	32%
		3	1.9	4%	36%
	N	1	11.9	27%	27%
		2	2.5	6%	33%
		3	1.6	4%	36%
	O	1	14.3	32%	32%
		2	2.2	5%	38%
		3	1.5	3%	41%
7	M	1	15.4	35%	35%
		2	2.8	6%	41%
		3	1.6	4%	45%
	N	1	15.3	35%	35%
		2	2.1	5%	40%
		3	1.5	3%	43%
8	M	1	11.5	26%	26%
		2	3.2	7%	33%
		3	2.1	5%	38%
	N	1	15.6	36%	36%
		2	2.0	4%	40%
		3	1.6	4%	44%

Table 9.6 EOC English II Principal Component and Variance by Form

Course	Form	Component	Eigen	Variance	Cumulative Variance
English II	M	1	17.2	34%	34%
		2	2.3	4%	38%
		3	1.8	4%	42%
	N	1	18.3	36%	36%
		2	2.5	5%	41%
		3	1.6	3%	44%
	O	1	18.3	36%	36%
		2	1.9	4%	40%
		3	1.5	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 evidence for the North Carolina assessments started from the item writing and test development phase where TMSs from NCSU-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 blueprints, performance expectations in terms of cognitive complexities or DOKs and creating a test reflecting target difficulty.

A formal alignment study quantifying the degree of alignment in the major outcome variables is planned to be available by October 2022.

9.6 Evidence Regarding Relationships with External Variables

One of the primary intended uses of the EOG and EOC assessments is to provide data to measure students' achievement and progress relative to readiness as defined by CCR standards. For the assessments to provide evidence of this type of achievement, it is important that reading passages are an appropriate measure of college and career reading.

In 2017, the NC state Board of Education adopted revisions to the Standard Course of Study in grades 3–8 Reading and English II that best aligned with the appropriate content for CCR to be implementation for schools in 2018–2019 (NCDPI, 2021b) administration. In order to understand the external context of the NC EOG and EOC results, the NCDPI commissioned MetaMetrics, Inc. for linking revised NC EOG Reading and NC EOC English II scales with the Lexile Framework® for Reading scale. The initial plan of using 2020–21 data for linking was reevaluated primarily because students' educational experiences during 2020–21 school year were atypical due to the COVID-19 pandemic. Since pre-equated parameters for the items

existed from pre-pandemic administration, MetaMetrics and the NCDPI designed and conducted the linking study using the pre-equated item parameters from the 2019 administration under advisement from the NCDPI technical advisory committee. The 2019 NC Ready EOG Reading and NC EOC English II assessments had an established link from the 2013 linking study. The use of pre-equated measures was also recommended by multiple professional organizations such as the National Council of Measurement in Education and The Chief State School Officers. Details on the methodology and results of the study are outlined in the updated Lexile Linking Technical Report (*Appendix 9-B*).

9.6.1 The Lexile Framework for Reading

The Lexile Framework is a tool that can help teachers, parents, and students locate challenging reading materials. Text complexity (difficulty) and reader ability are measured in the same unit—the Lexile. Text complexity is determined by examining such characteristics as word frequency and sentence length. Items and text are calibrated using the Rasch model. The typical range of the Lexile Scale is from 200L to 1600L, although actual Lexile measures can range from below zero (BR 150L) to above 1600L.

MetaMetrics has collected validity evidence over the past three decades to show that the Lexile Framework measures reading comprehension and text difficulty. This evidence includes demonstrating strong relationships between (1) the Lexile Framework and other measures of reading comprehension (e.g., other standardized assessments); (2) the Lexile Framework and Basal readers; and (3) the Lexile Framework and the difficulty of reading test items.

9.6.2 Linking the NC Assessments to the Lexile Framework

The EOG Reading tests consist of 40 operational items in grades 3–4, and 44 operational items in grades 5–8. The EOC English II Test consists of 51 operational items. The EOG and EOC tests are scaled horizontally, ranging from 500 to 600. It is important to note that, even though the reported EOG Reading tests scale ranges for the 2021 version are similar to the previous edition, the reported scale scores do not have the same meaning between editions.

The 2021 tests along with pre-equated item parameters were provided to MetaMetrics on both the 2019 and 2021 reporting scales. A one-to-one relationship between these scales did not exist at every scale score point. In such occurrences, MetaMetrics averaged 2019 score value associated with each 2021 reported score for each grade and course for linking the 2021 edition scale to 2019 edition scale. The Lexile linking formula established in 2013 was applied to these average 2019 scores. This provided a direct correspondence of Lexile Reading measures between the 2019 EOG Reading and EOC English II scales with the 2021 EOG Reading and EOC English II scales, respectively. Then a concordance table was established between the 2021 EOG Reading/EOC English II scale scores and the Lexile scale. The concordance table is an optimal solution in this scenario as the property of symmetry between the EOG Reading and EOC English II scale scores and the Lexile scale is maintained.

Table 9.7 presents the achievement level cut scores on the EOG Reading/EOC English II assessments and the associated Lexile measures. There are three achievement level cuts: Level 3,

Level 4, and Level 5. The values in the table are the cut scores associated with the minimal score for each category.

Table 9.7 North Carolina EOG Reading and EOC English II Performance Level Cut Scores and the Associated Lexile Measures[3]

Grade	Level 3		Level 4		Level 5	
	NC EOG Reading/EOC English II Scale Score	Lexile Measure	NC EOG Reading/EOC English II Scale Score	Lexile Measure	NC EOG Reading/EOC English II Scale Score	Lexile Measure
3	540	725L	546	860L	551	985L
4	544	840L	548	935L	556	1125L
5	550	985L	554	1075L	560	1220L
6	552	1030L	558	1180L	567	1400L
7	554	1075L	559	1195L	566	1370L
8	557	1145L	563	1300L	572	1515L
English	549	1240L	555	1405L	565	1655L

Figure 9.8 shows the EOG Reading and EOC English II Lexile reading measures and the Lexile reading use norms. The normative information for the Lexile Framework for Reading is based on linking studies conducted with the Lexile Framework and the results of assessments that report directly in the Lexile metric (N = 3,888,110). The EOG Reading and EOC English II scale scores as expressed in the Lexile metric are very similar to the 25th, 50th, and 75th percentiles trends across the grade ranges. In grade 3 and English II the 75th and 50th percentiles are slightly above the Lexile user norms and the 25th percentile is slightly lower than the Lexile user norms. For the remainder of grades, the selected percentiles for EOG Reading are slightly below the Lexile user norms. Overall, the EOG Reading and English II show very similar patterns with that of the Lexile user norms.

[3] The table is different from that presented in original report. This version was updated to reflect the current five achievement level cuts currently used by NCDPI

Figure 9.8 Selected Percentiles (25th, 50th, and 75th) Plotted for the EOG Reading/EOC English II Lexile Reading Measures in Relation to the Lexile Measure Norms

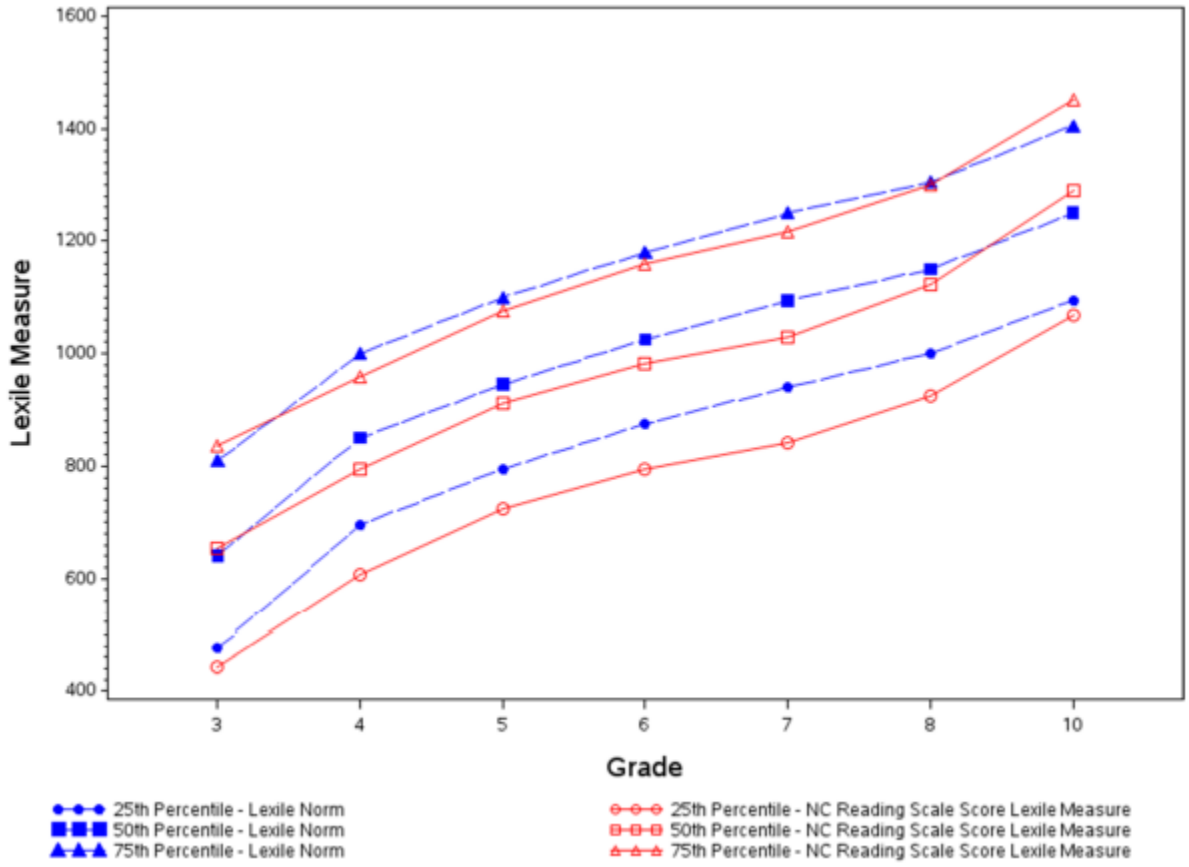


Table 9.8 provides the scale score and Lexile reading measure ranges for each of the achievement levels.

Table 9.8 NC EOG Reading and NC EOC English II achievement level scale score ranges and associated Lexile reading measures

	Not Proficient		Level 3		Level 4		Level 5	
Grade /Test Level	NC EOG/EOC Scale Score Range	Lexile Measure Range	NC EOG/EOC Scale Score Range	Lexile Measure Range	NC EOG/EOC Scale Score Range	Lexile Measure Range	NC EOG/EOC Scale Score Range	Lexile Measure Range
3	515-539	115L-700L	540-545	725L-835L	546-550	860L-965L	551-564	985L-1200L
4	517-543	210L-820L	544-547	840L-910L	548-555	935L-1100L	556-568	1125L-1300L
5	524-549	370L-960L	550-553	985L-1055L	554-559	1075L-1195L	560-573	1220L-1400L
6	528-551	445L-1005L	552-557	1030L-1160L	558-566	1180L-1380L	567-578	1400L-1500L
7	527-553	395L-1055L	554-558	1075L-1180L	559-565	1195L-1355L	566-580	1370L-1600L
8	532-556	530L-1125L	557-562	1145L-1265L	563-571	1300L-1500L	572-584	1515L-1700L
Eng. II	524-548	595L-1210L	549-554	1240L-1370L	555-564	1405L-1630L	565-577	1655L-1980L

9.6.3 The Lexile Framework and College- and Career-Reading Demands

MetaMetrics also conducted research on the reading demands of the EOG and EOC tests that are typically associated with CCR and developed a Lexile-based reading text complexity range for each grade band. *Figure 9.9* shows the relationship between the "Level 3" achievement level/proficiency standard for each test level established on the EOG Reading and EOC English II tests and the "stretch" reading demands. At each grade, the lowest score in the Level 3 range is the cut point and the highest score in the Level 3 range is the last score before the Level 4 cut point, with a dashed line connecting them. *Figure 9.9* helps contextualize the proficiency level set by the NCDPI by showing that students classified as "Level 3" and above on the EOG Reading and EOC English II should be able to read text that they are likely to encounter as they prepare for college and careers.

Figure 9.9 Comparison of EOG Reading and EOC English II "Level 3" Achievement Level with College and Career Reading Levels

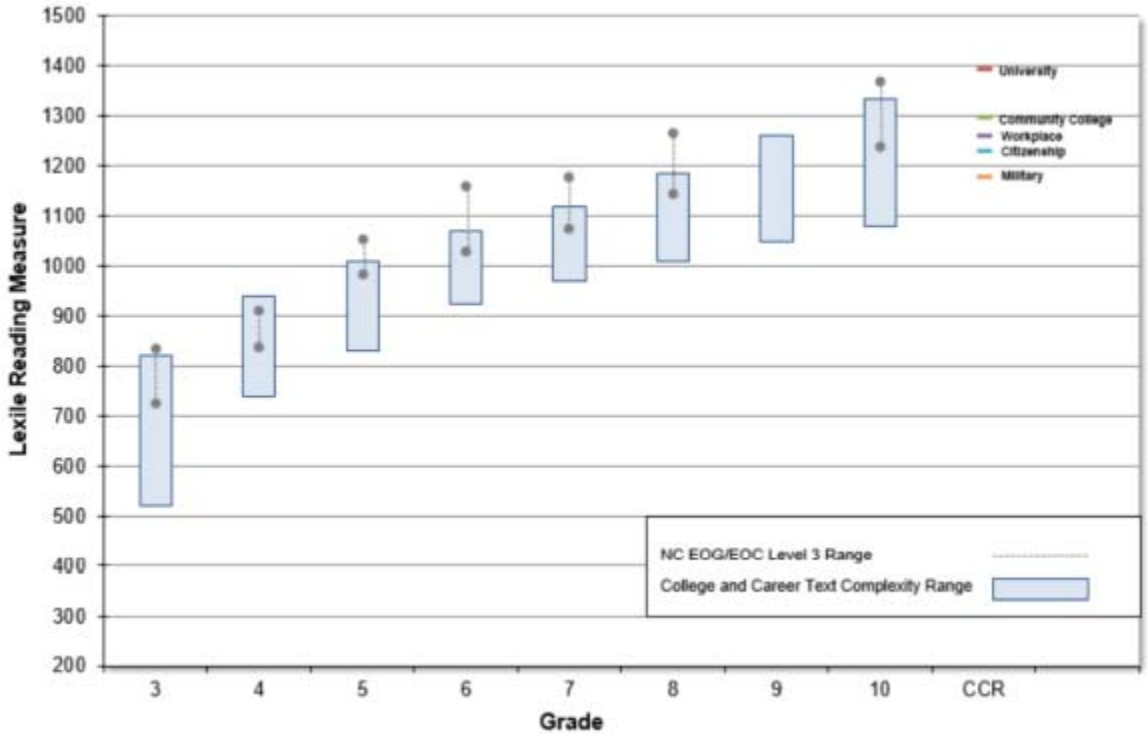
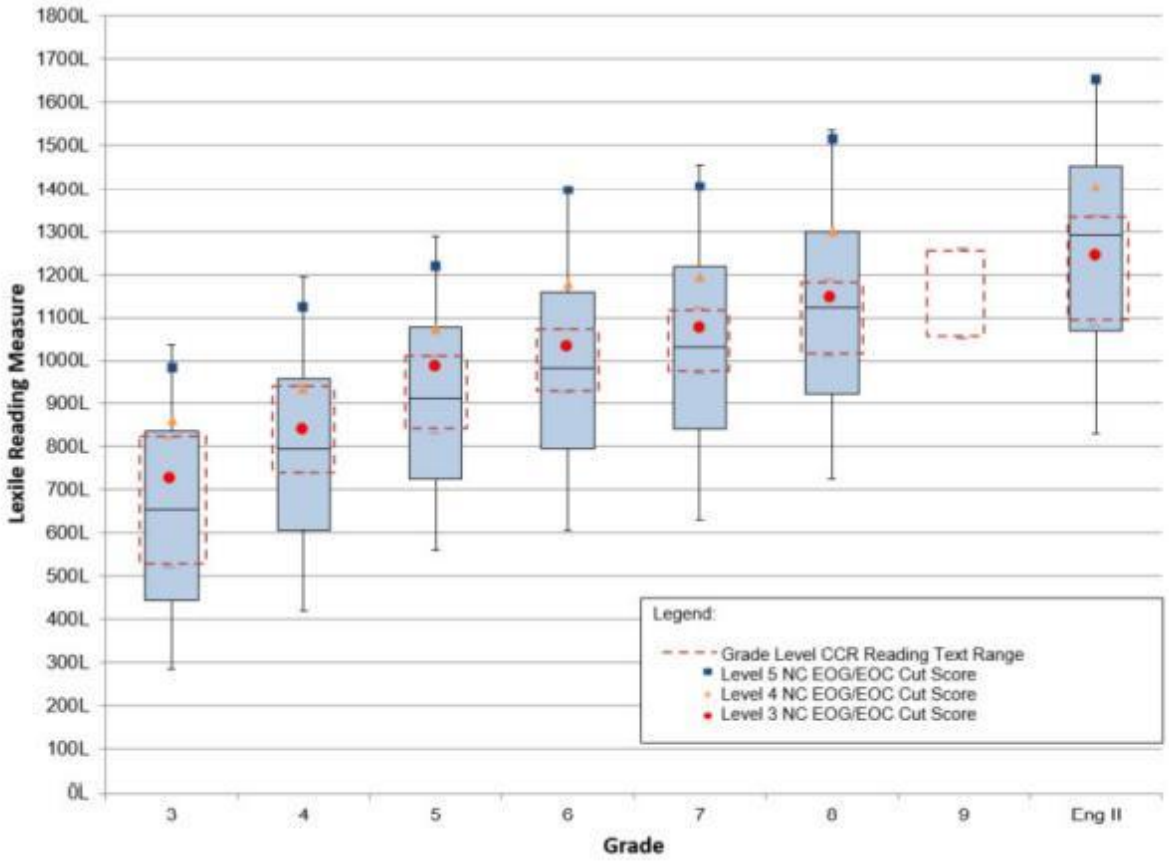


Figure 9.10 compares the distribution of 2020–2021 student performance expressed as Lexile Reading measures (blue boxes) to the North Carolina achievement level cut scores and CCR levels. For each test level, the blue box refers to the interquartile range. The line within the blue box indicates the median. The end of each whisker represents the 5th percentile at the low end and the 95th percentile at the high end of the distribution of students' Lexile Reading measures. For each grade/test level the achievement level cut scores are provided. Across grades, most student scores fall within or above the CCR text demand ranges. For each grade level the median student score is below the Level 3 achievement level. For English II, the Level 3 cut is slightly

below median. Combining student results with criterion referenced indicators provides information to reference when matching students with reading texts.

Figure 9.10 NC EOG Reading and NC EOC English II 2020–2021 Student Performance Expressed as Lexile Reading Measures Overlayed with the Achievement Level Descriptors and Grade Level CCR Reading Text Ranges



9.6.4 Conclusions

The NC assessments were linked to the Lexile Framework as a means of collecting evidence on the rigor of the NC assessments and external validity of the EOG and EOC assessments. This study showed that the reading levels of the NC assessments are aligned with expectations of CCR as measured by the Lexile Framework. In addition, these results provided evidence that the rigor of reading measured by the NC assessment has increased since the previous version of the assessment.

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 students' 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 trains item writers 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[4]
    - 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
    - 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

[4] See discussions on fairness review in Chapter 4

- 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 a *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.

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 NCATP. 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 Term	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 Term	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.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick, Statistical theories of mental test scores (pp. 397-479). Reading, MA: Addison-Wesley.

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.

CCSSO (2020). Restart & Recovery: Assessments in Spring 2021. (nciea.org).

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 Algebra I/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.

North Carolina Department of Public Instruction (2021b). Standard Course of Study. https://www.dpi.nc.gov/districts-schools/classroom-resources/academic-standards/standard-course-study

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika monograph supplement. 1969;17(4):2–2. doi: 10.1002/j.2333-8504.1968.tb00153.x.

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

https://www.ncleg.net/Sessions/2013/Bills/Senate/HTML/S812v7.html

Appendix 1-B The North Carolina Academic Standards Review Commission Report Dec2015

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

Appendix 1-C North Carolina Testing Code of Ethics

https://www.dpi.nc.gov/documents/testing-code-ethics

APPENDIX 2

Appendix 2-A Reading and English II Test Specification Meeting Agendas, Survey Form, and Demographic Information of Participants

English Language Arts
Test Specifications Meeting Agenda
February 26, 2018
Jane S. McKimmon Center, NC State University

8:30am	**Registration—Main Lobby**
9:00am	**Welcome and Introductions** Dr. Tammy Howard, Dan Auman • Meeting purpose • Substitute Teacher Form, Stipend Form, Demographics Form • Testing Code of Ethics and Test Security Agreement • Travel Reimbursement
9:30am	**Summative Assessment Psychometric Overview** Dr. Kinge Mbella
10:15am	**Break**
10:30am	**Overview of Revised ELA Standards** DPI-Curriculum & Instruction and Exceptional Children Divisions
11:30am	**Prioritizing Standards Overview** Dan Auman
11:45am	**Lunch** (on your own)
12:45pm	**Prioritize Standards—ROUND 1** (Breakout Groups—General and EC: Grades 3-5, Grades 6-8, and Grades 9-12) • Prioritize Assessable Standards • Recommend Weighting by Domain
2:15pm	**Break** (on your own)
2:30pm	**Prioritize Standards—ROUND 2** (Breakout Groups) • Prioritize Assessable Standards • Recommend Weighting by Domain
3:15pm	**Recommend Percent by Item Type—Discussion** (Large Group) Dan Auman, Kinge Mbella
3:45pm	**Summary of Recommendations and General Considerations** Dan Auman
4:00 pm	**Meeting Adjourned**

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-A Demographic Characteristics of the Test Specification Meeting Participants

Category	Sub-Category	Reading/English II (N=57)	
		N	%
Gender	Female	54	95%
	Male	3	5%
Ethnicity	Asian	3	5%
	Black	7	12%
	Native American	1	2%
	Hispanic	2	4%
	White	42	74%
	Mixed	2	4%
Highest Degrees Earned	BA/BS	20	35%
	J.D./Ed.D/Ph.D	1	2%
	MA/MS/M.Ed	36	63%
District Size	Large	18	32%
	Large/Medium	1	2%
	Medium	21	37%
	Small	14	25%
	Blank	3	5%
Urbanicity	Rural	26	46%
	Suburban	19	33%
	Urban	9	16%
	Blank	3	5%

*Some participants did not declare some of the demographic characteristics

Appendix 2-B General Definition of ELA DOK Level

https://www.nciea.org/publications/DOKreading_KH08.pdf

Appendix 2-C A Guide for using Webb's DOK with Common Core State Standards

Webbs-DOK-Flip-Chart.pdf (casciac.org)

Appendix 2-D North Carolina Annual Testing Program Test Development Process

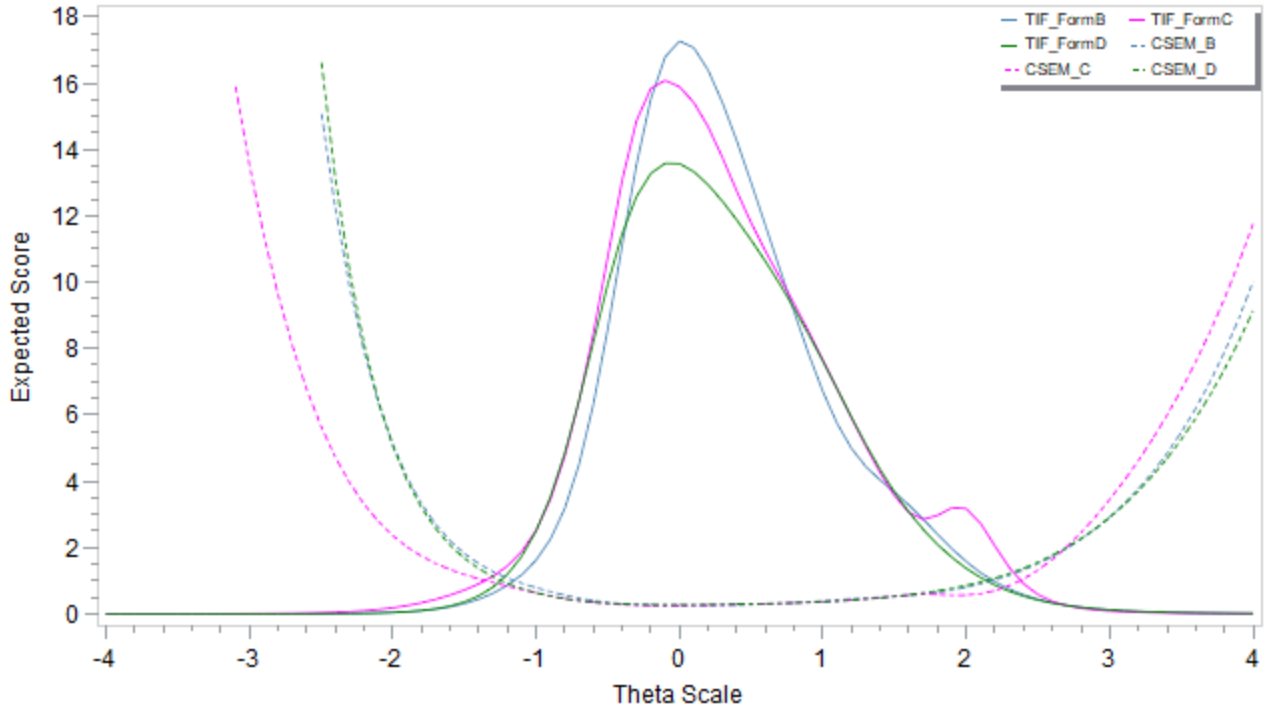
Assessment Development Process (nc.gov)

APPENDIX 4

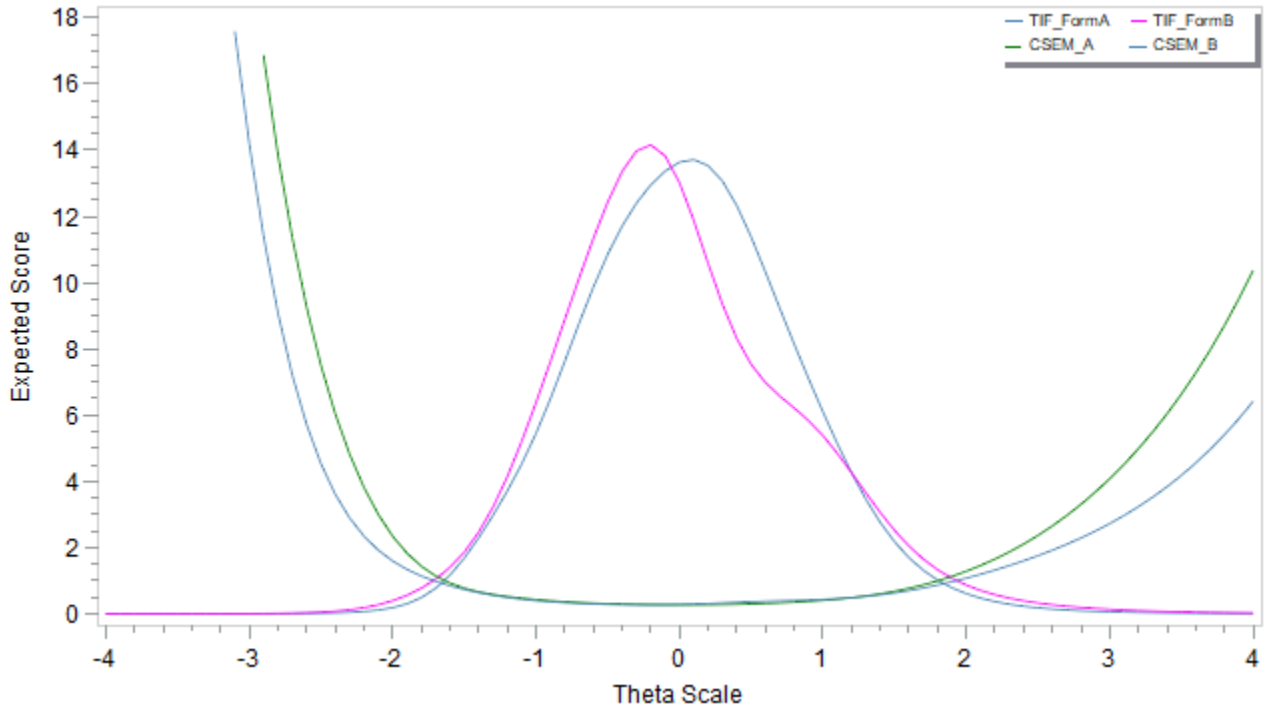
Appendix 4-A Field-Test TIFs and CSEMs

Test Information Functions and Conditional Standard Error of Measurement

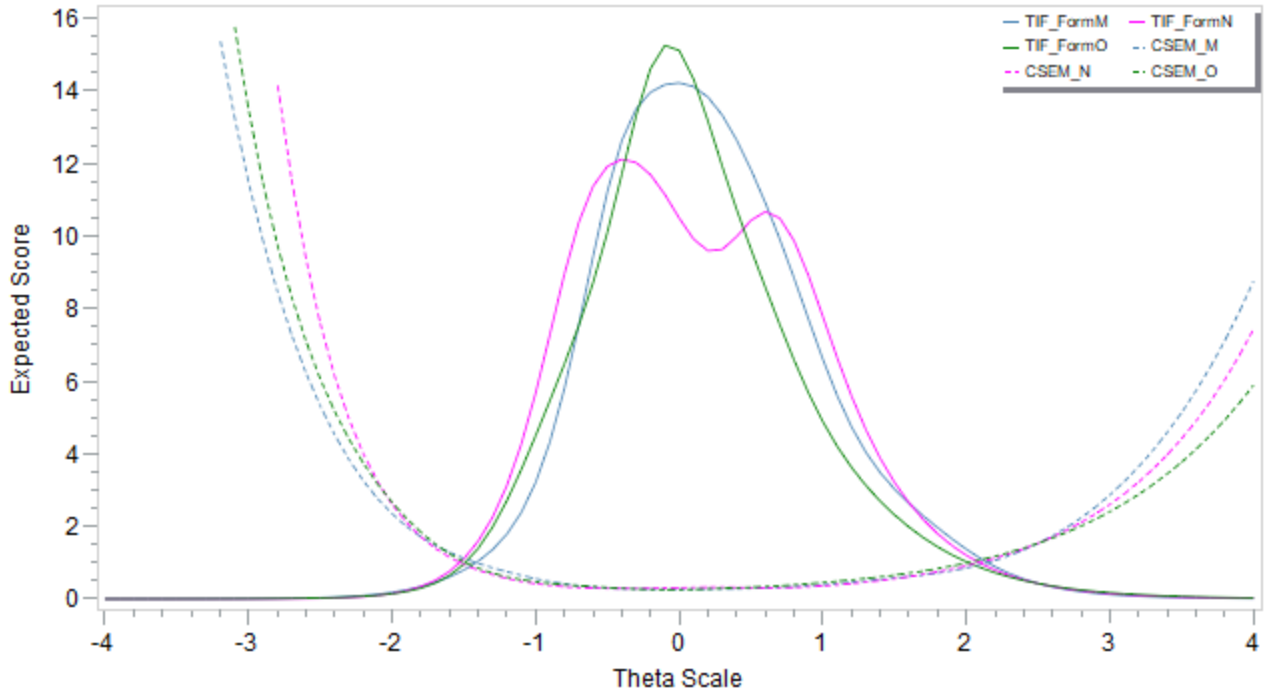
TIFs and CSEMs Based on Field Test Item Parameters, Grade 3



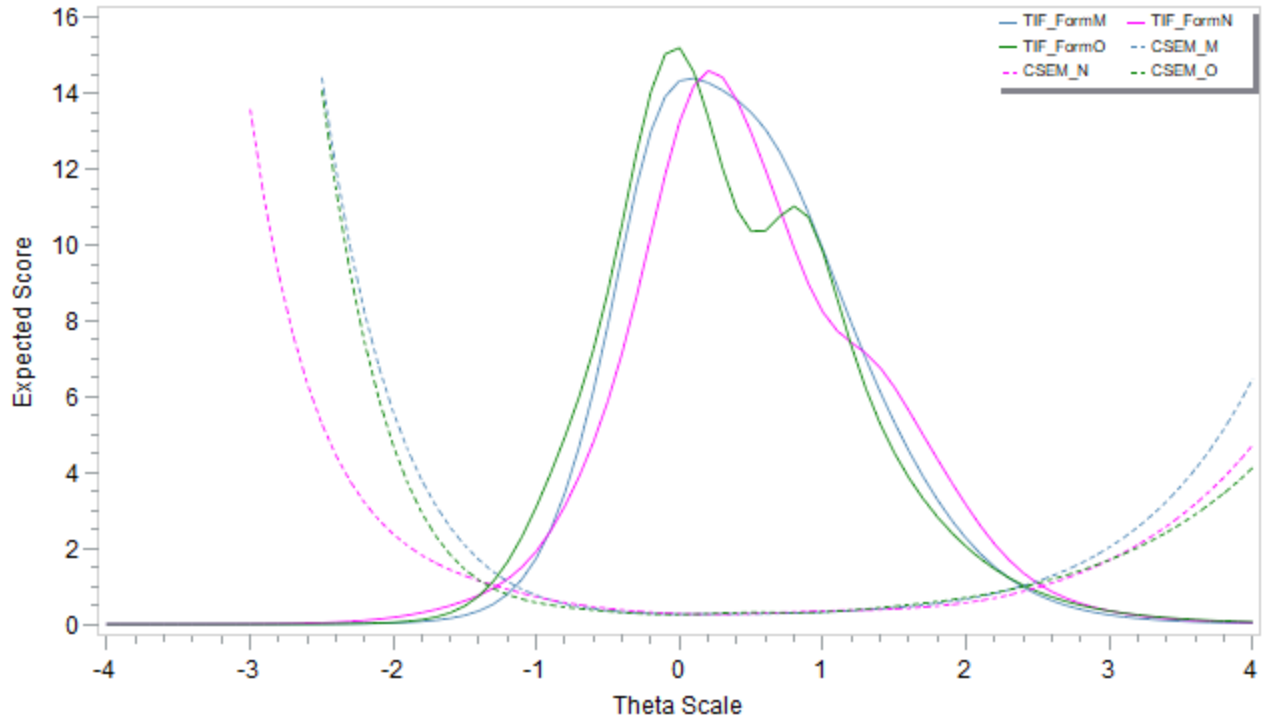
TIFs and CSEMs Based on Field Test Item Parameters, Grade 4



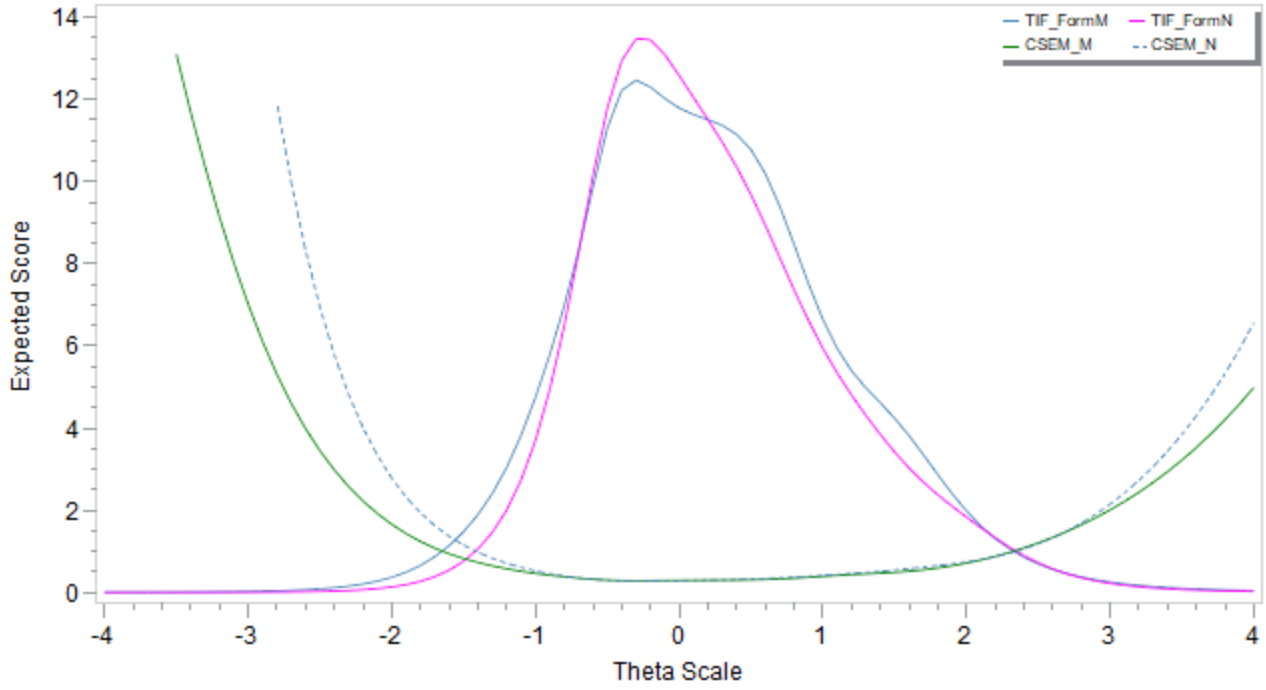
TIFs and CSEMs Based on Field Test Item Parameters, Grade 5



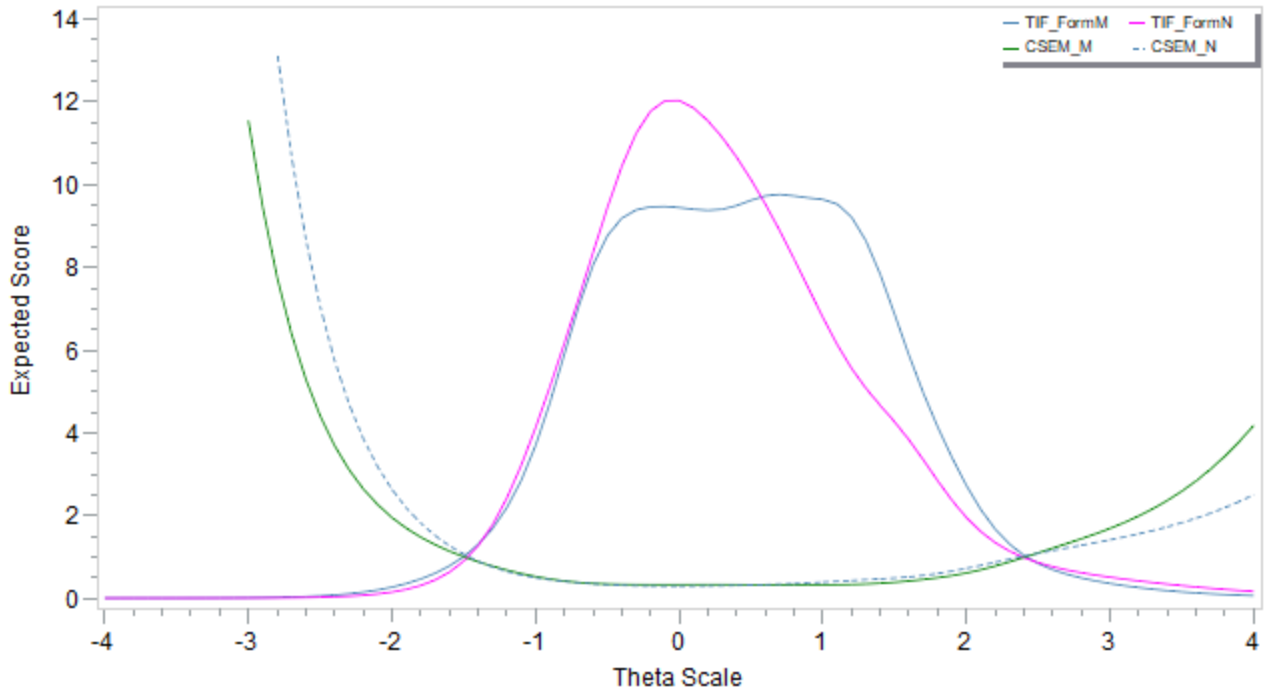
TIFs and CSEMs Based on Field Test Item Parameters, Grade 6



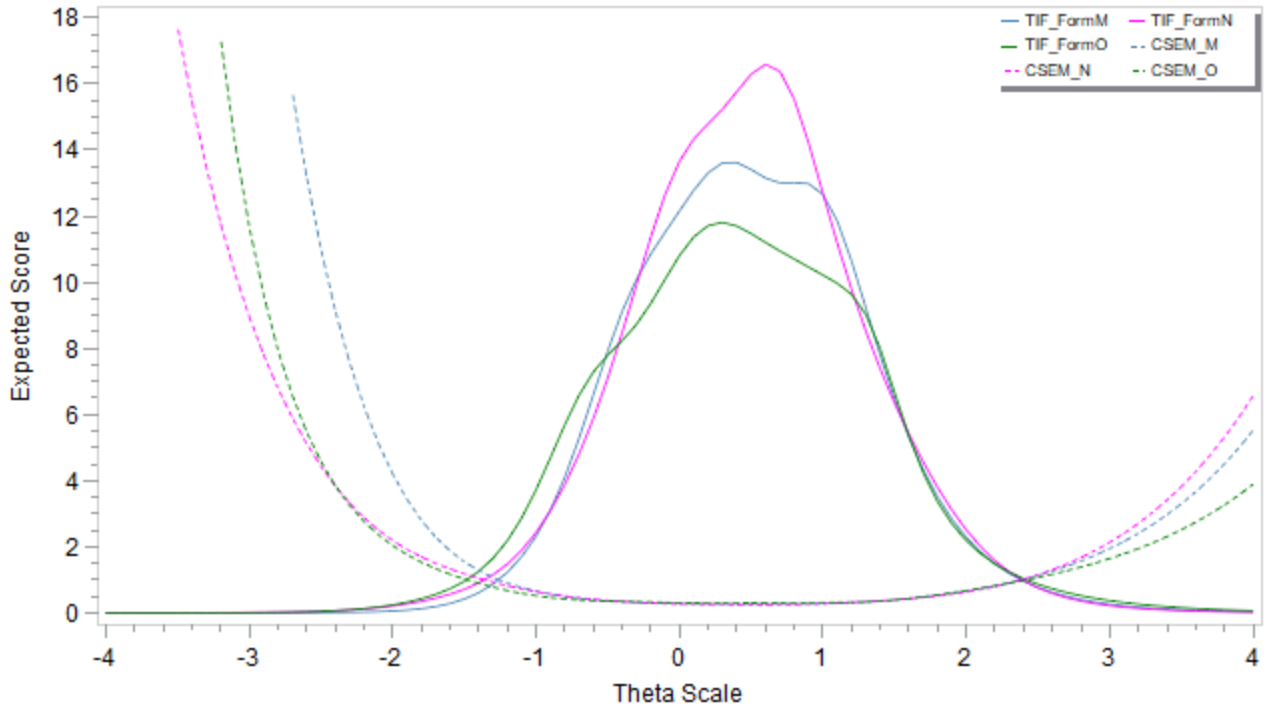
TIFs and CSEMs Based on Field Test Item Parameters, Grade 7



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



TIFs and CSEMs Based on Field Test Item Parameters, English II



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 4-C Example of CR Items

https://www.dpi.nc.gov/media/9704/open

APPENDIX 5

Appendix 5-A The Proctor's Guide

https://center.ncsu.edu/ncaccount/pluginfile.php/1543/course/section/561/2020%20Proctor%20Guide.pdf

Appendix 5-B Guidelines for Testing English Learners Students

https://www.dpi.nc.gov/districts-schools/testing-and-school-accountability/testing-policy-and-operations/testing-students-identified-english-learners

Appendix 5-C Guidelines for Testing Students with Disability

https://www.dpi.nc.gov/districts-schools/testing-and-school-accountability/testing-policy-and-operations/testing-students-disabilities

Appendix 5-D Testing Security Protocols and Procedures

https://www.dpi.nc.gov/media/12865/open

Appendix 5-E North Carolina Test Coordinators' Policies and Procedures Handbook

https://www.dpi.nc.gov/media/12865/open

APPENDIX 6

Appendix 6-A Test Information Functions (TIFs) and Standard Error of Measurements (SEMs)

Figure 6.1 Reading Grade 3 TIFs and CSEMs 2020-21 Operational Forms

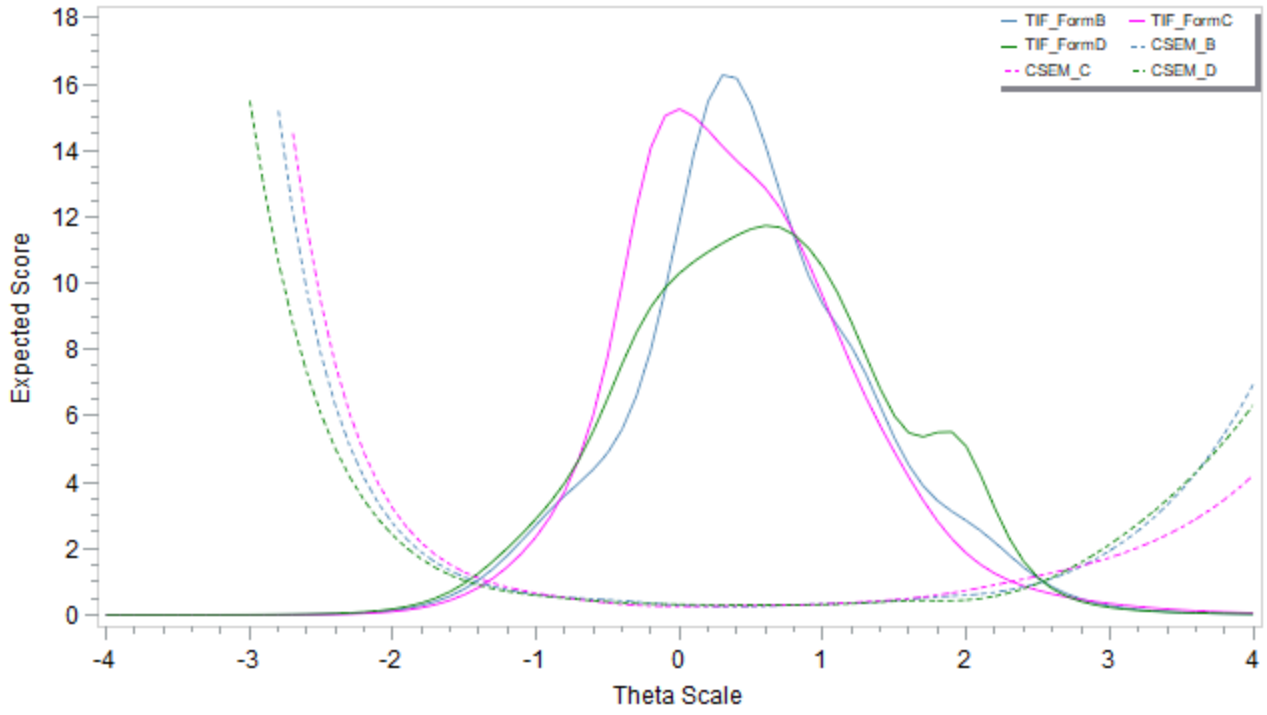


Figure 6.2 Reading Grade 4 TIFs and CSEMs 2020-21 Operational Forms

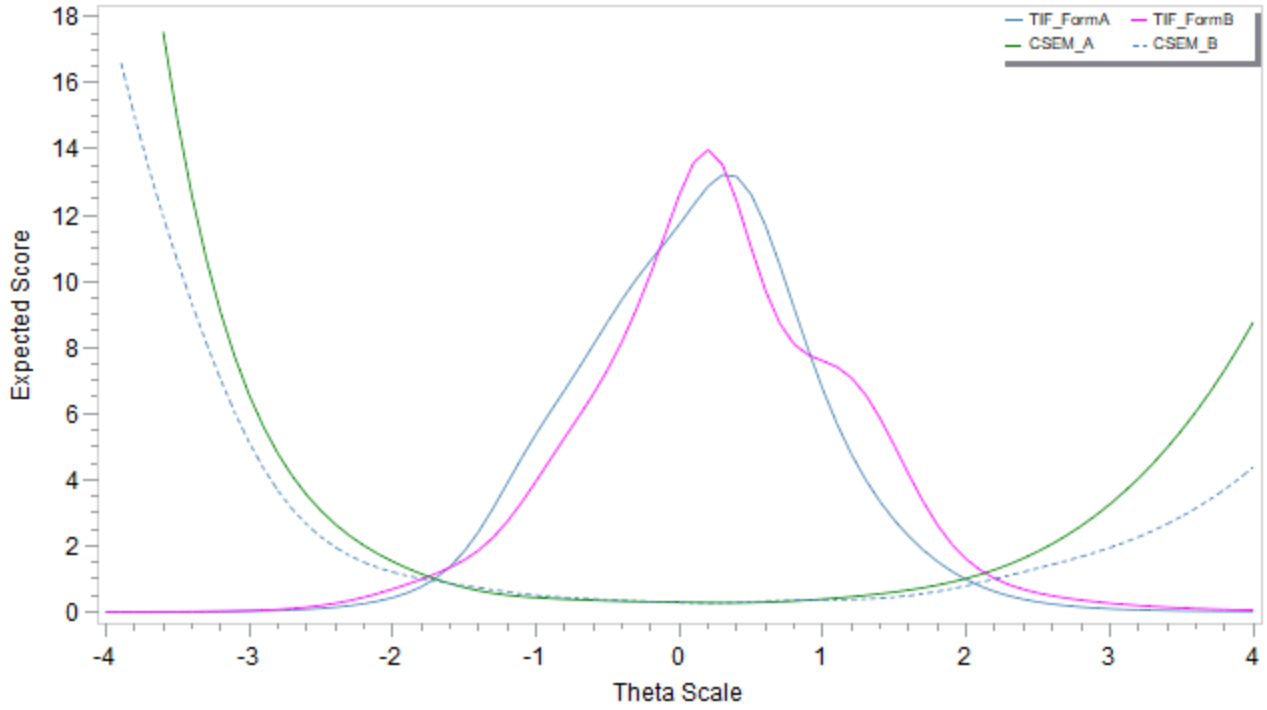


Figure 6.3 Reading Grade 5 TIFs and CSEMs 2020-21 Operational Forms

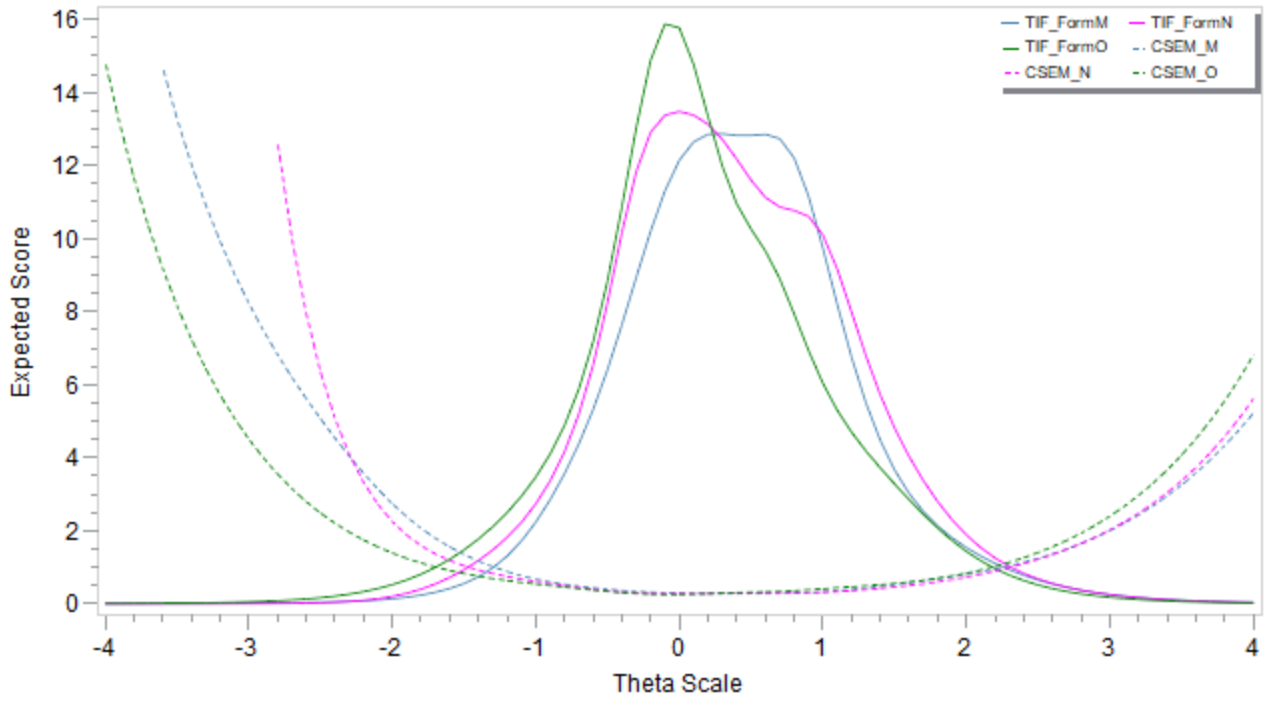


Figure 6.4 Reading Grade 6 TIFs and CSEMs 2020-21 Operational Forms

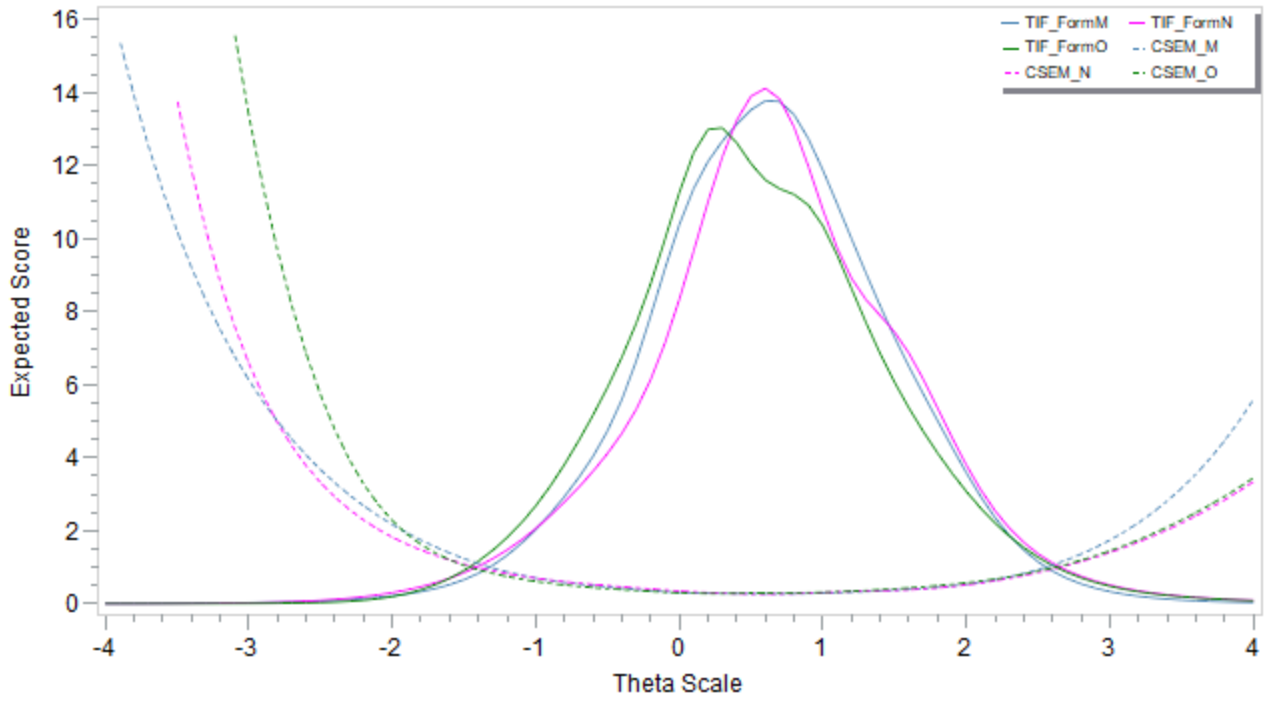


Figure 6.5 Reading Grade 7 TIFs and CSEMs 2020-21 Operational Forms

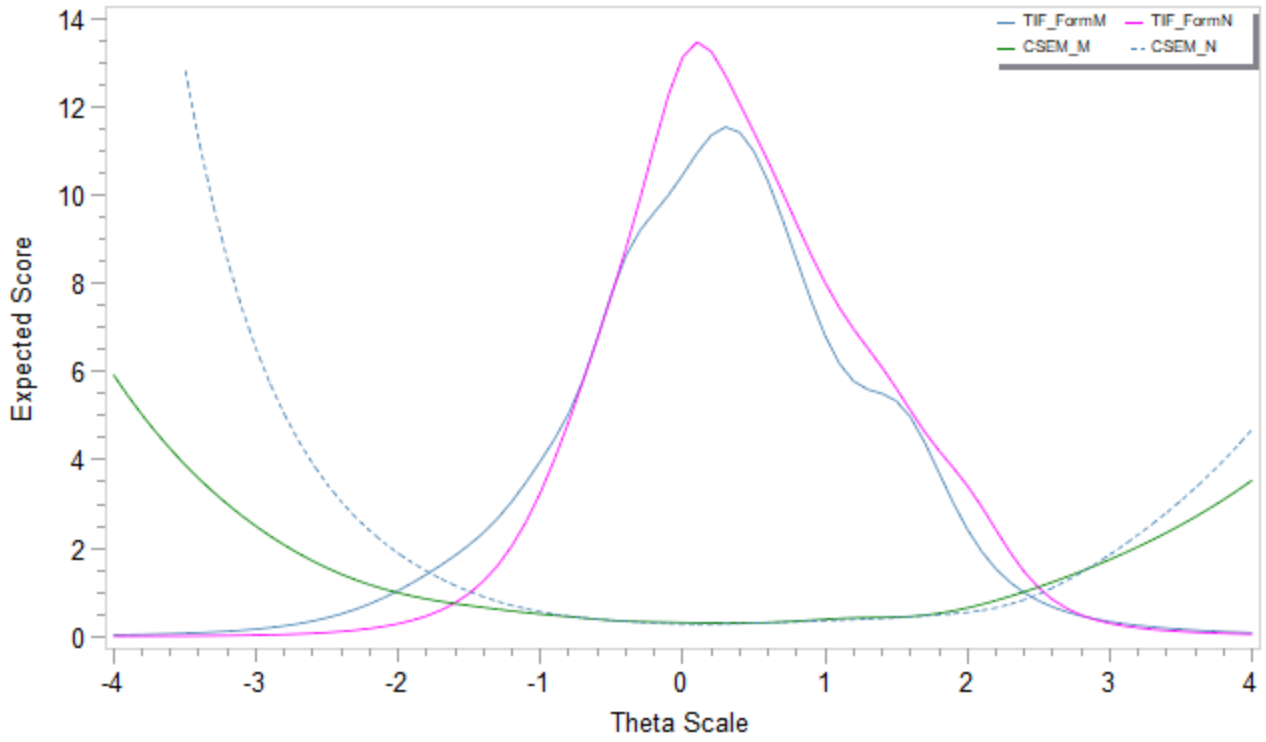


Figure 6.6 Reading Grade 8 TIFs and CSEMs 2020-21 Operational Forms

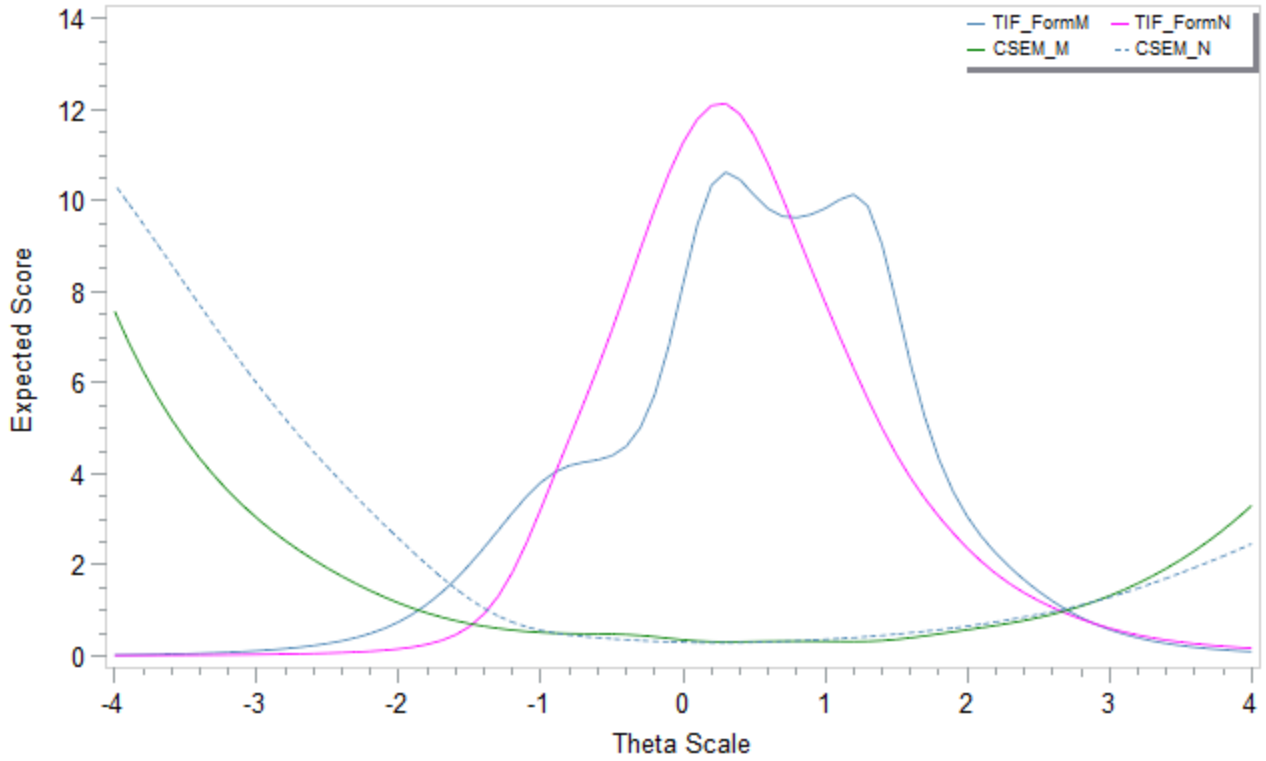
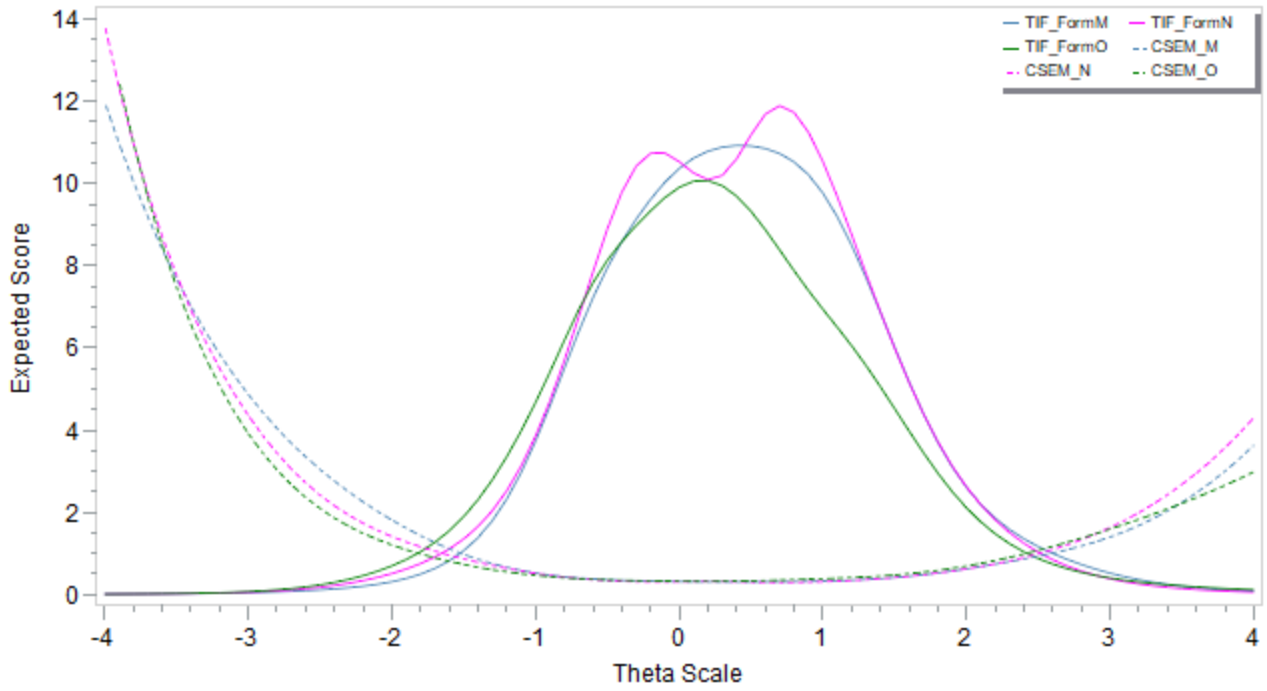


Figure 6.7 English II TIFs and CSEMs 2020-21 Operational Forms


TIF_FormM
TIF_FormN
TIF_FormO
CSEM_M
CSEM_N
CSEM_O
Expected Score
Theta Scale
14
12
10
8
6
4
2
0
-4
-3
-2
-1
0
1
2
3
4

APPENDIX 7

Appendix 7-A North Carolina EOC English II Standard Setting 2020 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 EOG Grades 3–8 Reading Standard Setting 2021 Technical Report

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

Appendix 7-C External Evaluation Report of EOC English II Standard Setting 2020

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

Appendix 7-D External Evaluation Report of EOG Grades 3–8 Standard Setting 2021

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 Reading 2020–21 Scale Score by Subgroups

Table 1. 2020–21 Reading Scale Scores by Subgroups, Grades 3-5

Grade	Type	Category	N	Statistics		Range		Percentile		
				Mean	SD	Min	Max	25th	Median	75th
EOG 3	SWD	Regular	93,368	437.6	10.1	414	466	430	438	445
		Students with Disability	12,522	429.1	8.7	414	466	423	427	434
	EDS	Not Economically Disadvantaged	61,939	439.4	10.2	414	466	431	440	447
		Economically Disadvantaged	43,951	432.8	9.1	414	466	426	432	440
	Els	Regular	92,678	437.4	10.3	414	466	429	438	445
		Other	1,851	434.4	11.7	414	466	423	432	445
		English Language Learner	11,361	430.5	8.0	414	465	424	430	436
EOG 4	SWD	Regular	93,139	543.1	9.7	517	568	536	543	550
		Students with Disability	13,025	533.1	8.6	517	568	526	531	538
	EDS	Not Economically Disadvantaged	62,916	544.6	9.9	517	568	538	545	552
		Economically Disadvantaged	43,248	538.0	9.0	517	568	531	537	545
	ELs	Regular	92,729	542.7	10.0	517	568	535	543	550
		Other	2,417	541.4	11.5	517	568	530	543	551
		English Language Learner	11,018	535.4	8.0	517	564	529	535	541
EOG 5	SWD	Regular	94,879	548.5	9.4	524	573	541	549	555
		Students with Disability	13,284	538.4	7.5	524	573	533	536	542
	EDS	Not Economically Disadvantaged	64,445	549.8	9.6	524	573	543	551	557
		Economically Disadvantaged	43,718	543.5	8.7	524	573	536	543	550
	ELs	Regular	94,126	548.0	9.7	524	573	540	549	555
		Other	4,771	547.6	9.3	524	573	541	549	554
		English Language Learner	9,266	539.1	6.6	524	569	535	538	543

Table 2. 2020–21 Reading Scale Scores by Subgroups, Grade 6-8

Grade	Type	Category	N	Statistics		Range		Percentile		
				Mean	SD	Min	Max	25th	Median	75th
EOG 6	SWD	Regular	97,687	551.3	9.5	528	578	544	552	558
		Students with Disability	13,241	541.5	7.1	528	578	537	540	545
	EDS	Not Economically Disadvantaged	66,064	552.6	9.8	528	578	545	553	560
		Economically Disadvantaged	44,864	546.5	8.6	528	578	540	546	553
	ELs	Regular	96,641	550.9	9.8	528	578	543	551	558
		Other	6,428	549.7	8.5	529	578	544	550	556
		English Language Learner	7,859	541.3	6.0	529	577	537	540	545
EOG 7	SWD	Regular	99,149	553.6	9.6	528	580	547	554	561
		Students with Disability	12,976	542.7	7.5	528	580	537	541	547
	EDS	Not Economically Disadvantaged	68,525	554.7	9.9	528	580	548	555	562
		Economically Disadvantaged	43,600	548.6	9.0	528	580	541	548	555
	ELs	Regular	97,321	553.1	10.0	528	580	546	553	561
		Other	6,474	552.4	8.9	528	580	547	553	559
		English Language Learner	8,330	543.7	6.9	528	578	538	542	548
EOG 8	SWD	Regular	99,527	557.2	9.6	532	584	550	557	564
		Students with Disability	12,766	546.9	7.3	532	584	542	546	551
	EDS	Not Economically Disadvantaged	70,588	558.2	9.9	532	584	551	559	565
		Economically Disadvantaged	41,705	552.4	8.8	532	584	546	552	559
	ELs	Regular	104,274	556.7	9.8	532	584	549	557	564
		Other	1,990	552.3	9.9	532	584	544	552	560
		English Language Learner	6,029	546.5	6.3	532	584	542	546	550

Table 3. 2020–21 Reading Scale Scores by Subgroups, English II

Grade	Type	Category	N	Statistics		Range		Percentile		
				Mean	SD	Min	Max	25th	Median	75th
English II	SWD	Regular	98,540	551.3	9.0	524	577	545	552	558
		Students with Disability	11,267	540.5	7.2	524	573	535	539	545
	EDS	Not Economically Disadvantaged	74,057	552.0	9.3	524	577	546	553	559
		Economically Disadvantaged	35,750	546.4	8.7	525	577	539	546	553
	ELs	Regular	104,129	550.7	9.3	524	577	544	551	557
		Other	1,476	542.9	9.8	524	575	535	540	551
		English Language Learner	4,202	539.5	6.2	525	565	535	538	544

Appendix 8-B EOC English II Achievement Level Ranges and Descriptors

EOC English II Achievement Level Descriptors | NC DPI

Appendix 8-C EOG Grades 3–8 Achievement Level Ranges and Descriptors

https://www.dpi.nc.gov/media/5868/open

Appendix 8-D Grades 3–8 Reading and English II 2020-21 Proficiency Classification by Subgroup

Table 1. 2020-21 Reading Proficiency Classifications by Subgroups, Grades 3-5

Grade	Type	Category	N	Not Proficient	Level 3	Level 4	Level 5
3	SWD	Regular	93,368	51.0	12.1	28.2	8.7
		Students with Disability	12,522	83.8	5.3	8.8	2.1
	EDS	Not Economically Disadvantaged	61,939	43.1	12.2	32.9	11.8
		Economically Disadvantaged	43,951	71.4	9.9	16.2	2.5
	ELs	Regular	92,678	51.5	11.7	28.0	8.8
		Other	1,851	60.0	6.6	25.6	7.8
		English Language Learner	11,361	81.4	8.7	9.3	0.7
4	SWD	Regular	93,139	50.5	15.3	24.8	9.5
		Students with Disability	13,025	86.6	5.4	6.4	1.6
	EDS	Not Economically Disadvantaged	62,916	43.4	15.3	28.8	12.5
		Economically Disadvantaged	43,248	71.6	12.2	13.5	2.7
	ELs	Regular	92,729	51.7	14.6	24.3	9.5
		Other	2,417	50.6	13.4	27.4	8.6
		English Language Learner	11,018	82.8	9.5	6.9	0.8
5	SWD	Regular	94,879	53.1	14.6	19.9	12.3
		Students with Disability	13,284	90.3	4.4	3.9	1.4
	EDS	Not Economically Disadvantaged	64,445	46.3	15.0	22.7	16.0
		Economically Disadvantaged	43,718	74.4	11.0	10.9	3.6
	ELs	Regular	94,126	54.4	14.0	19.4	12.2
		Other	4,771	53.9	17.3	20.4	8.5
		English Language Learner	9,266	92.6	4.9	2.2	0.3

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. 2020-21 Reading Proficiency Classifications by Subgroups, Grades 6-8

Grade	Type	Category	N	Not Proficient	Level 3	Level 4	Level 5
6	SWD	Regular	97,687	49.9	23.7	20.3	6.1
		Students with Disability	13,241	89.9	6.5	3.0	0.6
	EDS	Not Economically Disadvantaged	66,064	43.9	24.1	24.0	8.1
		Economically Disadvantaged	44,864	70.6	18.0	9.9	1.6
	ELs	Regular	96,641	51.5	22.6	19.9	6.1
		Other	6,428	55.6	26.9	15.4	2.1
		English Language Learner	7,859	93.7	5.3	0.9	0.2
7	SWD	Regular	99,149	48.5	18.5	21.3	11.7
		Students with Disability	12,976	90.2	5.5	3.4	1.0
	EDS	Not Economically Disadvantaged	68,525	43.0	18.3	23.9	14.8
		Economically Disadvantaged	43,600	69.5	15.0	11.8	3.7
	ELs	Regular	97,321	50.2	17.6	20.7	11.6
		Other	6,474	52.6	22.0	18.6	6.8
		English Language Learner	8,330	90.5	6.6	2.6	0.3
8	SWD	Regular	99,527	47.0	22.5	24.0	6.5
		Students with Disability	12,766	89.1	7.1	3.3	0.5
	EDS	Not Economically Disadvantaged	70,588	42.3	22.4	27.0	8.3
		Economically Disadvantaged	41,705	67.9	18.0	12.5	1.6
	ELs	Regular	104,274	49.2	21.7	22.9	6.3
		Other	1,990	64.9	18.0	14.6	2.6
		English Language Learner	6,029	92.7	5.5	1.7	0.0

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. 2020-21 English II Proficiency Classifications by Subgroups

Grade	Type	Category	N	Not Proficient	Level 3	Level 4	Level 5
English II	SWD	Regular	98,540	36.3	25.1	32.0	6.6
		Students with Disability	11,267	85.7	9.5	4.5	0.4
	EDS	Not Economically Disadvantaged	74,057	33.2	24.0	34.7	8.1
		Economically Disadvantaged	35,750	58.2	22.5	17.6	1.7
	ELs	Regular	104,129	39.0	24.3	30.5	6.3
		Other	1,476	70.2	13.9	14.2	1.7
		English Language Learner	4,202	90.7	7.6	1.7	0.0

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-E Interpretive Guide to the Score Reports for the North Carolina End-of Grade Assessments, 2018–19

https://www.dpi.nc.gov/media/9654/open

APPENDIX 9

Appendix 9-A Two Factors Exploratory Factor Analysis with Simple Structure

Grade 3

Order	Form B/N		Form C/O		Form D/P	
	Factor		Factor		Factor	
	1	2	1	2	1	2
1	0.78	-0.14	0.15	0.01	0.77	-0.30
2	0.57	0.12	0.92	-0.04	0.71	-0.15
3	0.64	-0.05	0.05	0.02	0.55	0.30
4	0.68	-0.44	0.81	0.07	0.80	-0.02
5	0.72	-0.03	0.75	-0.12	0.75	-0.24
6	0.84	0.13	0.74	-0.09	0.07	-0.09
7	0.81	-0.01	0.61	-0.43	0.74	-0.07
8	0.51	0.14	0.66	-0.38	0.47	0.03
9	0.55	-0.22	0.76	-0.35	0.65	0.01
10	0.75	0.21	0.43	0.10	0.73	0.20
11	0.82	0.17	0.70	-0.19	0.64	-0.04
12	-0.03	-0.28	0.23	0.20	0.79	-0.27
13	0.68	0.40	0.43	0.44	0.80	-0.13
14	0.74	0.17	-0.14	-0.22	0.72	0.13
15	0.63	0.42	0.89	-0.07	0.65	0.31
16	0.70	0.21	0.80	0.17	0.85	-0.04
17	0.65	0.36	0.55	-0.16	0.37	-0.13
18	0.60	0.34	0.78	-0.21	0.87	-0.12
19	-0.01	0.05	0.43	0.30	0.41	0.51
20	-0.05	0.03	0.57	0.02	0.31	0.50
21	0.38	0.41	0.75	0.11	0.51	0.13
22	-0.06	0.23	0.68	0.02	0.69	0.27
23	0.78	-0.37	-0.39	-0.20	0.83	-0.30
24	0.46	0.09	0.91	0.00	0.26	0.44
25	0.66	-0.12	0.84	0.12	0.47	-0.08
26	0.24	0.16	0.64	0.21	-0.50	-0.25
27	0.41	-0.21	0.36	0.15	0.73	0.12
28	-0.09	0.08	0.69	0.16	0.53	-0.13

Order	Form B/N		Form C/O		Form D/P	
	Factor		Factor		Factor	
	1	2	1	2	1	2
29	-0.52	0.15	0.37	0.13	0.11	0.43
30	0.22	-0.21	0.65	0.21	0.22	0.36
31	0.63	-0.06	0.51	0.42	-0.50	0.31
32	0.28	-0.11	-0.09	0.65	0.45	0.46
33	0.11	0.52	0.18	0.07	0.06	0.31
34	0.76	-0.13	0.08	0.25	0.71	-0.18
35	0.25	-0.13	0.06	0.00	0.27	-0.07
36	0.62	0.04	0.84	-0.09	0.58	-0.15
37	0.54	-0.18	0.41	0.14	0.11	0.21
38	0.61	-0.35	0.37	0.38	0.29	0.14
39	0.84	-0.37	0.74	-0.25	0.77	-0.18
40	0.65	-0.15	0.77	-0.14	0.07	0.06

Grade 4

Order	Form A/M		Form B/N	
	Factor		Factor	
	1	2	1	2
1	0.26	0.30	0.31	-0.44
2	0.61	-0.05	0.62	-0.14
3	0.65	-0.30	0.79	-0.12
4	0.31	-0.05	0.71	0.05
5	0.57	-0.24	0.47	-0.14
6	0.63	-0.20	0.67	-0.16
7	0.51	0.06	0.73	-0.03
8	0.43	0.03	0.78	-0.03
9	0.03	-0.39	0.42	-0.13
10	0.30	-0.37	0.73	0.09
11	0.79	-0.12	0.57	0.26
12	0.25	0.43	0.78	-0.29
13	0.07	0.17	0.76	0.18
14	0.28	0.09	0.73	-0.42
15	0.76	-0.01	0.39	-0.12
16	0.56	-0.01	0.19	0.26
17	-0.32	0.40	-0.10	-0.16

Order	Form A/M		Form B/N	
	Factor		Factor	
	1	2	1	2
18	0.29	0.20	-0.20	0.47
19	0.55	0.25	0.11	0.56
20	0.61	0.45	0.74	-0.20
21	-0.03	0.31	0.09	0.36
22	0.31	0.08	-0.43	-0.06
23	0.56	-0.19	0.69	-0.20
24	0.72	0.18	0.52	0.37
25	0.16	0.46	-0.09	0.40
26	0.84	-0.27	0.08	0.45
27	0.42	0.36	0.63	0.37
28	0.47	0.00	0.35	0.27
29	0.73	-0.17	0.21	0.38
30	0.54	0.05	0.52	0.17
31	0.04	-0.21	0.74	0.15
32	0.45	0.01	0.74	0.15
33	0.51	-0.24	0.77	0.37
34	0.64	0.05	0.61	-0.28
35	0.31	0.12	0.63	0.28
36	0.55	0.37	0.63	-0.17
37	0.25	0.36	0.79	-0.19
38	0.72	0.08	0.35	-0.02
39	0.71	-0.10	0.63	0.00
40	0.12	-0.20	0.82	0.07

Grade 5

Order	Form A/M		Form B/N		Form C/O	
	Factor		Factor		Factor	
	1	2	1	2	1	2
1	-0.29	-0.02	-0.19	-0.17	0.44	-0.41
2	0.69	-0.25	0.70	-0.07	-0.11	0.23
3	0.48	0.03	0.15	0.40	0.25	-0.12
4	0.19	-0.16	0.66	-0.29	0.80	-0.03
5	0.64	-0.24	0.37	0.05	0.43	0.26
6	0.58	0.03	0.26	0.63	0.35	0.00

Order	Form A/M		Form B/N		Form C/O	
	Factor		Factor		Factor	
	1	2	1	2	1	2
7	0.38	0.18	0.67	0.36	0.76	-0.04
8	0.56	-0.16	0.67	0.19	0.17	0.04
9	0.57	-0.11	-0.36	0.02	0.84	-0.03
10	0.25	0.36	0.38	-0.06	0.79	-0.16
11	0.32	0.56	0.11	0.58	0.55	-0.09
12	-0.15	0.00	0.54	-0.25	0.30	-0.46
13	0.25	0.12	0.84	-0.01	0.58	-0.18
14	0.04	-0.21	0.08	0.26	0.91	0.03
15	0.64	0.22	0.03	0.26	0.24	0.11
16	0.34	0.26	0.41	0.39	0.82	-0.09
17	0.14	-0.57	0.53	-0.10	0.88	0.01
18	0.60	0.14	0.74	0.16	0.42	0.08
19	0.70	-0.09	0.30	0.19	0.84	0.07
20	0.57	0.22	0.62	-0.22	0.11	0.32
21	0.68	-0.34	0.73	-0.11	-0.02	0.28
22	0.76	0.10	0.67	-0.19	0.78	0.27
23	0.30	-0.07	0.78	-0.20	0.76	-0.31
24	0.73	0.21	0.90	-0.07	0.46	0.42
25	0.53	0.33	0.54	-0.19	0.33	0.30
26	0.71	-0.14	0.79	-0.02	0.78	0.23
27	0.64	-0.23	0.62	-0.02	0.03	0.35
28	0.42	-0.11	0.64	0.16	0.41	0.12
29	0.55	-0.07	0.62	0.04	0.70	0.17
30	0.65	-0.09	0.59	0.28	0.26	0.05
31	0.33	0.54	0.42	0.36	0.67	0.39
32	0.29	-0.07	0.65	0.14	0.75	0.08
33	0.58	-0.01	0.26	0.21	0.37	0.40
34	0.73	-0.34	0.69	-0.40	0.04	-0.14
35	0.52	0.55	0.65	0.21	0.48	0.55
36	0.14	-0.07	0.51	0.14	0.40	-0.37
37	0.47	-0.26	0.08	-0.27	0.64	-0.33
38	-0.03	0.19	0.79	-0.23	0.72	-0.40
39	0.08	0.05	0.65	-0.30	0.73	-0.36
40	0.00	-0.04	-0.06	0.08	0.58	0.07

Grade 6

Order	Form A/M		Form B/N		Form C/O	
	Factor		Factor		Factor	
	1	2	1	2	1	2
1	0.27	0.20	0.47	-0.22	0.77	-0.35
2	0.77	-0.33	0.28	-0.26	0.65	-0.09
3	0.16	0.07	0.61	-0.46	0.47	-0.21
4	0.49	0.11	0.61	0.14	0.80	0.04
5	0.08	0.04	0.72	-0.21	0.53	0.32
6	0.65	-0.31	0.39	0.39	0.27	0.47
7	0.32	-0.48	0.47	-0.45	0.43	0.23
8	0.35	-0.20	0.09	0.04	0.41	0.34
9	0.62	0.10	0.31	-0.21	0.55	-0.07
10	0.37	0.34	-0.21	0.24	0.08	0.25
11	0.60	-0.26	0.19	0.05	0.50	0.01
12	0.44	-0.25	0.28	0.31	0.27	-0.16
13	0.74	-0.30	0.63	-0.14	0.66	-0.08
14	0.78	-0.07	0.62	-0.29	0.39	-0.32
15	0.64	0.25	0.25	0.12	0.22	0.36
16	0.41	0.20	0.11	0.05	0.37	0.48
17	0.84	0.03	0.16	0.34	0.08	-0.27
18	0.54	0.03	0.75	-0.16	0.62	-0.11
19	0.67	-0.14	-0.10	-0.15	0.32	0.46
20	0.75	0.03	-0.06	0.44	0.65	0.10
21	0.14	0.44	0.60	0.18	0.79	-0.03
22	0.57	0.04	0.69	0.12	0.81	0.12
23	0.57	-0.32	0.60	-0.21	0.10	0.34
24	0.21	0.20	0.18	-0.16	-0.14	0.15
25	0.02	-0.03	0.76	0.04	0.34	-0.15
26	-0.09	0.26	0.19	-0.24	0.13	0.20
27	0.56	0.03	0.66	0.17	0.78	0.15
28	-0.40	0.22	0.60	0.06	0.78	0.22
29	-0.11	0.04	0.67	0.02	0.86	-0.20
30	0.56	0.32	0.58	0.36	0.77	0.04
31	0.75	0.08	0.73	-0.24	0.78	-0.02
32	0.28	0.08	0.72	0.18	0.38	-0.14
33	0.58	0.19	0.64	0.02	0.61	0.04
34	0.65	-0.18	0.74	0.14	-0.06	-0.02
35	0.58	0.32	-0.19	0.25	-0.04	0.19

Order	Form A/M		Form B/N		Form C/O	
	Factor		Factor		Factor	
	1	2	1	2	1	2
36	0.71	0.03	0.53	0.10	0.82	0.09
37	0.48	0.20	0.60	0.28	0.30	0.19
38	0.68	0.08	0.53	0.33	-0.20	-0.26
39	0.33	0.53	0.66	0.31	0.70	-0.09
40	0.41	0.15	0.64	-0.29	0.81	-0.14
41	0.49	0.19	0.33	0.19	0.83	-0.23
42	0.57	-0.05	0.69	0.07	0.63	-0.21
43	0.07	0.04	0.66	0.05	0.73	-0.27
44	0.13	-0.56	-0.15	0.42	0.86	0.09

Grade 7

Order	Form A/M		Form B/N	
	Factor		Factor	
	1	2	1	2
1	0.71	-0.25	0.47	-0.24
2	0.54	0.31	-0.03	-0.33
3	-0.21	-0.23	0.54	0.16
4	0.78	0.07	0.59	-0.13
5	0.64	-0.02	-0.15	0.43
6	0.70	0.21	0.24	0.25
7	0.62	0.06	0.10	-0.22
8	0.70	-0.31	0.82	-0.16
9	0.85	-0.32	-0.51	0.26
10	-0.09	-0.49	0.71	-0.04
11	0.43	0.11	0.83	-0.11
12	0.77	-0.29	0.22	-0.26
13	0.35	-0.34	0.82	-0.16
14	0.03	0.16	0.66	0.16
15	0.67	0.24	0.85	-0.16
16	0.40	0.01	0.20	0.18
17	0.21	0.37	0.62	0.00
18	-0.15	0.39	0.62	0.30
19	0.44	0.27	0.37	0.08
20	0.72	0.30	0.83	-0.03
21	-0.17	0.48	0.84	0.05
22	0.36	0.26	-0.05	0.41

Order	Form A/M		Form B/N	
	Factor		Factor	
	1	2	1	2
23	0.84	-0.11	0.30	-0.06
24	0.50	0.32	0.66	0.00
25	0.61	-0.25	0.70	0.27
26	0.80	0.04	0.84	0.13
27	0.72	0.23	0.08	0.19
28	0.58	-0.10	0.69	-0.01
29	0.92	-0.08	0.27	0.39
30	0.68	0.04	0.10	-0.03
31	0.65	0.18	0.69	0.08
32	0.19	0.23	0.69	-0.13
33	0.70	0.15	0.26	0.45
34	0.69	-0.31	0.75	-0.20
35	0.70	0.17	0.50	0.35
36	0.18	0.29	0.60	-0.04
37	0.78	0.07	0.77	0.12
38	0.59	0.00	0.89	0.18
39	0.54	0.38	0.62	0.37
40	0.77	-0.23	0.62	-0.28
41	0.62	-0.50	0.40	0.01
42	0.18	0.03	0.72	-0.22
43	-0.08	-0.16	0.52	-0.18
44	0.79	-0.11	0.65	-0.09

Grade 8

Order	Form A/M		Form B/N	
	Factor		Factor	
	1	2	1	2
1	0.39	0.02	0.78	-0.08
2	0.22	-0.24	0.77	0.09
3	0.60	-0.23	0.85	-0.28
4	0.33	0.44	0.43	0.17
5	0.43	0.40	0.37	0.03
6	0.66	-0.32	-0.03	-0.72
7	-0.21	0.02	0.58	0.31
8	0.01	0.43	0.78	-0.17
9	0.28	-0.13	0.66	0.05

Order	Form A/M		Form B/N	
	Factor		Factor	
	1	2	1	2
10	0.12	-0.09	0.56	0.00
11	0.00	0.01	0.75	-0.18
12	0.76	-0.26	0.41	0.03
13	-0.38	0.13	0.06	0.11
14	0.62	-0.12	0.64	-0.29
15	0.27	0.55	0.68	-0.06
16	0.58	0.27	0.81	-0.20
17	0.47	-0.02	0.75	-0.03
18	0.57	-0.44	0.39	-0.10
19	0.53	-0.05	0.54	-0.05
20	0.10	0.54	0.82	-0.14
21	-0.15	-0.10	0.79	-0.13
22	-0.08	0.31	0.58	0.25
23	0.81	-0.37	0.33	0.35
24	0.54	0.19	0.63	0.02
25	0.79	0.24	0.78	-0.02
26	0.82	0.27	0.00	-0.22
27	0.46	0.28	0.82	-0.02
28	0.70	0.13	0.39	0.36
29	0.39	0.58	0.62	0.22
30	0.33	0.18	0.59	0.11
31	0.58	0.02	0.46	0.18
32	0.57	0.03	0.59	0.12
33	0.63	-0.08	0.78	-0.03
34	0.83	-0.25	0.75	-0.12
35	0.42	0.12	0.60	0.06
36	0.53	0.04	0.08	0.44
37	0.11	0.20	0.52	0.06
38	-0.16	0.47	0.73	0.07
39	0.67	0.23	0.74	0.14
40	0.28	-0.12	-0.27	-0.10
41	0.56	0.01	0.27	0.20
42	0.73	0.14	0.61	0.09
43	0.82	-0.28	-0.20	0.26
44	0.46	-0.28	-0.33	0.34

English II

Order	Form M		Form N		Form O	
	Factor		Factor		Factor	
	1	2	1	2	1	2
1	0.43	0.14	-0.20	-0.07	0.12	-0.07
2	0.70	0.03	0.60	0.11	0.55	0.27
3	0.57	0.09	0.39	-0.32	0.18	0.25
4	0.36	0.43	0.89	-0.04	0.67	0.08
5	0.22	0.44	0.51	0.02	0.01	0.34
6	0.04	0.28	0.58	0.22	0.36	0.22
7	0.36	0.03	0.55	0.52	-0.17	-0.09
8	0.81	-0.01	0.45	0.04	0.60	0.00
9	0.41	0.02	0.82	-0.24	0.41	-0.22
10	0.71	0.07	0.40	0.44	0.71	-0.23
11	0.82	-0.15	0.74	-0.47	0.18	0.18
12	0.26	-0.13	0.44	-0.26	0.44	-0.22
13	0.74	0.16	0.81	-0.15	0.67	0.15
14	0.81	-0.14	0.76	-0.17	0.16	0.25
15	0.38	0.08	0.71	-0.15	0.60	-0.20
16	0.62	0.15	-0.03	-0.06	0.81	-0.05
17	0.35	0.23	0.25	-0.28	0.41	0.37
18	0.81	-0.28	0.72	-0.07	0.79	0.08
19	0.77	-0.10	0.33	0.08	0.34	0.26
20	0.56	0.16	0.46	0.10	-0.02	0.32
21	0.60	-0.05	0.19	0.32	0.58	0.05
22	0.84	-0.19	-0.01	0.10	0.76	0.21
23	0.14	-0.45	0.80	-0.07	0.40	0.02
24	0.20	-0.43	0.15	0.34	0.52	-0.07
25	0.60	0.08	0.31	0.23	0.23	0.21
26	0.22	-0.09	0.82	-0.20	0.43	-0.14
27	0.91	-0.12	0.92	-0.07	0.94	0.21
28	0.50	0.32	0.63	0.02	0.63	-0.23
29	0.25	0.42	0.56	0.26	0.45	-0.20
30	0.34	-0.23	0.74	0.10	0.41	-0.45
31	0.38	0.31	0.42	0.25	0.55	0.21
32	0.87	-0.16	0.47	-0.44	0.84	-0.22
33	0.68	-0.22	0.70	-0.22	0.90	-0.20
34	0.60	-0.05	0.61	0.06	0.64	0.07
35	0.63	-0.07	0.87	-0.06	0.35	-0.01

Order	Form M		Form N		Form O	
	Factor		Factor		Factor	
	1	2	1	2	1	2
36	0.74	0.11	0.89	0.02	0.62	0.34
37	0.90	-0.02	0.92	-0.06	0.93	0.06
38	0.30	0.12	0.32	0.39	0.71	-0.27
39	0.32	0.10	0.84	-0.06	0.62	0.20
40	0.67	0.17	0.69	0.43	0.69	0.08
41	0.84	-0.01	0.40	0.36	0.92	-0.06
42	0.64	0.11	-0.08	0.12	0.45	-0.01
43	0.35	0.21	0.30	-0.08	0.83	-0.10
44	0.19	0.21	0.54	0.19	0.80	-0.01
45	0.60	-0.20	0.60	-0.06	0.61	0.03
46	0.36	0.36	0.34	0.30	0.73	-0.08
47	0.92	0.02	0.85	-0.02	0.94	0.09
48	0.60	-0.13	0.76	0.01	0.66	-0.30
49	-0.34	0.13	0.53	0.08	0.53	0.05
50	0.40	0.15	0.65	0.03	0.62	0.00
51	0.47	-0.38	0.33	-0.03	0.39	0.04

Appendix 9-B North Carolina Lexile Linking Report by MetaMetrics

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