



## **Spring 2019 NSCAS Summative ELA, Mathematics, and Science Technical Report**



## Table of Contents

Executive Summary .....	12
Section 1: Introduction .....	15
1.1. NSCAS Overview.....	15
1.2. Background.....	16
1.3. Schedule of Major Events .....	16
1.4. Building a Validity Argument .....	17
1.4.1. Intended Purposes and Uses of Test Results.....	19
1.4.2. Theory of Action .....	19
Section 2: Test Design and Development .....	21
2.1. Test Designs.....	21
2.2. Academic Content Standards.....	24
2.3. Table of Specifications (TOS) .....	24
2.4. Item Types.....	25
2.5. Depth of Knowledge (DOK).....	25
2.6. ALD Development.....	30
2.6.1. Policy ALDs.....	30
2.6.2. Range ALDs.....	31
2.6.3. Reporting ALDs.....	33
2.7. ELA Passage Development .....	34
2.7.1. Passage Specifications .....	34
2.7.2. Readability Measures.....	34
2.8. Item Development.....	35
2.8.1. Item Specifications .....	35
2.8.2. Development Targets .....	36
2.8.3. Item Writer Workshop (IWW).....	37
2.8.4. Item Development Results .....	38
2.8.5. External Content and Bias Review .....	39
2.8.6. Item Retirement.....	41
2.9. Content Alignment .....	41
2.9.1. Alignment and Adaptive Testing .....	41
2.9.2. 2019 Mathematics Alignment Study .....	42
2.10. Universal Design.....	43
2.11. Sensitivity and Fairness .....	43
2.12. Test Construction.....	44
2.13. Data Review.....	44
Section 3: Test Administration and Security .....	46
3.1. User Roles and Responsibilities.....	47
3.2. Administration Training .....	47
3.3. Item Type Samplers.....	48
3.4. Accommodations and Accessibility Features .....	49
3.5. User Acceptance Testing (UAT).....	51
3.6. Student Participation.....	52
3.6.1. Paper-Pencil Participation Criteria.....	52

3.6.2. Participation of English Language Learners (ELLs) .....	52
3.6.3. Participation of Recently Arrived Limited English Proficient (RAEL) Students..	53
3.7. Test Security .....	53
3.7.1. Online Test Security .....	54
3.7.2. Paper-Pencil Test Security .....	54
3.7.2.1. Physical Warehouse Security .....	54
3.7.2.2. Secure Destruction of Test Materials .....	55
3.7.2.3. Shipping Security .....	55
3.7.2.4. Electronic Security of Test Materials and Data .....	55
3.7.3. Caveon Test Security .....	56
3.7.3.1. Monitoring for Disclosure of Test Content .....	56
3.7.3.2. Monitoring for Potential Test Security Violations .....	56
3.8. Partner Support.....	58
Section 4: Scoring and Reporting.....	59
4.1. Scoring Rules.....	59
4.2. Paper-Pencil Scoring .....	59
4.2.1. Scanning of Answer Sheets by EDS.....	59
4.2.1.1. Quality Control of Scanning and Scoring.....	60
4.2.1.2. Quality Control of Image Editing.....	61
4.2.1.3. Quality Control of Answer Document Processing and Scoring .....	61
4.2.2. Scoring by NWEA.....	61
4.3. Score Reporting Methods .....	62
4.4. Report Summary .....	64
4.5. Reporting Process .....	66
4.5.1. Online Reports .....	66
4.5.2. Printed ISRs .....	66
4.5.3. Report Verification.....	67
4.6. Matrix.....	67
Section 5: Constraint-Based Engine.....	70
5.1. Overview.....	70
5.2. Engine Simulations and Evaluation .....	71
5.2.1. Evaluation Criteria .....	72
5.2.2. Blueprint Constraint Accuracy .....	73
5.2.3. Item Exposure Rates.....	75
5.2.4. Score Precision and Reliability .....	77
Section 6: Psychometric Analyses .....	81
6.1. Number of Student Included in the Analyses.....	81
6.2. Classical Item Analyses .....	82
6.2.1. Item Difficulty (P-Value).....	82
6.2.2. Item Discrimination (Item-Total Correlation) .....	83
6.2.3. Item Suppression .....	85
6.3. Differential Item Functioning (DIF).....	86
6.3.1. DIF Methods.....	87
6.3.2. DIF Results .....	89
6.4. IRT Calibration.....	92

6.4.1. Checking Model Assumptions .....	93
6.4.1.1. Unidimensionality.....	93
6.4.1.2. Local Independence.....	95
6.4.1.3. Item Fit.....	96
6.4.2. Summary IRT Item Statistics .....	98
6.5. Vertical Scaling (ELA and Mathematics) .....	99
6.5.1. Linking Item Selection .....	99
6.5.2. Vertical Scaling Process.....	100
6.5.3. Vertical Scale Evaluation.....	100
6.5.4. 2020 Scaling Considerations.....	105
6.6. Post-Equating Check (Science) .....	105
6.6.1. Post-Equating Method.....	105
6.6.2. Post-Equating Results.....	106
6.7. Scaling.....	107
Section 7: Standard Setting.....	110
7.1. Overview.....	110
7.2. ID Matching Method.....	111
7.3. Meeting Process .....	111
7.4. ALD Revision .....	112
7.5. Final Results .....	112
Section 8: Test Results .....	114
8.1. Demographics and Accommodations.....	114
8.2. Administration Mode (Online vs. Paper-Pencil) .....	120
8.3. Testing Time .....	121
8.4. Achievement Level Distributions .....	123
8.5. Descriptive Statistics of Scale Scores .....	123
8.6. Reporting Category Correlations .....	124
8.7. Correlations with MAP Growth .....	129
8.8. Score Differences between 2018 and 2019.....	129
Section 9: Reliability.....	132
9.1. Marginal Reliability.....	132
9.2. Conditional Standard Error of Measurement (CSEM).....	134
9.3. Classification Accuracy .....	135
9.4. Reliability for Fixed Forms (Science).....	141
Section 10: Validity .....	143
10.1. Intended Purposes and Uses of Test Scores .....	143
10.2. Sources of Validity Evidence.....	144
10.3. Evidentiary Validity Framework .....	145
10.4. Interpretive Argument Claims.....	148
10.5. NSCAS Validity Argument.....	149
References .....	151

## List of Tables

Table 1.1. Schedule of Major Events.....	17
Table 2.1. Available NSCAS Summative Assessments in 2019 .....	21
Table 2.2. Number of Items and Points Per Test.....	22
Table 2.3. Horizontal Linking Configuration.....	23
Table 2.4. Item Types for Online ELA and Mathematics.....	25
Table 2.5. ELA Passage Targets and Development by Passage Type and Source.....	34
Table 2.6. Lexile and Word Count Ranges.....	35
Table 2.7. Overall Item Development Targets—ELA and Mathematics .....	36
Table 2.8. Item Development Targets—ELA .....	36
Table 2.9. Item Development Targets—Mathematics.....	37
Table 2.10. IWW Panel Composition—ELA and Mathematics.....	37
Table 2.11. Item Development Results—ELA .....	38
Table 2.12. Item Development Results—Mathematics .....	38
Table 2.13. Item Development Targets vs. Number of Items Developed .....	39
Table 2.14. Item Review Meeting Panel Composition .....	40
Table 2.15. External Item Review Results.....	40
Table 2.16. Data Review Flagging Criteria—Multiple-Choice Items.....	44
Table 2.17. Data Review Flagging Criteria—Non-Multiple-Choice Items.....	45
Table 2.18. Data Review Results .....	45
Table 3.1. User Roles and Responsibilities .....	47
Table 3.2. Fall 2018 Regional Workshop Locations and Participation .....	48
Table 3.3. Summative Test Administration Workshop Dates and Participation.....	48
Table 3.4. Accommodations and Universal Features .....	49
Table 3.5. Statistical Analysis and Potential Incidents.....	56
Table 3.6. Partner Support Communication Options .....	58
Table 3.7. Number of NSCAS Cases to Partner Support in 2018–2019.....	58
Table 4.1. Attemptedness Rules for Scoring .....	59
Table 4.2. Scale Score Ranges.....	62
Table 4.3. Achievement Level Descriptions.....	63
Table 4.4. Reporting Categories .....	63
Table 4.5. Non-Tested Codes (NTCs).....	64
Table 5.1. Blueprint Constraint by Reporting Category—Simulations.....	73
Table 5.2. Blueprint Constraint by Reporting Category—Engine Evaluation.....	74
Table 5.3. Item Exposure Rates—Simulations .....	76
Table 5.4. Item Exposure Rates—Engine Evaluation.....	77
Table 5.5. Mean Bias of the Ability Estimation (True - Estimated)—Simulations .....	78
Table 5.6. Score Precision and Reliability—Simulations .....	78
Table 5.7. Score Precision and Reliability—Engine Evaluation .....	79
Table 5.8. SEM by Deciles—Simulations .....	79
Table 5.9. SEM by Deciles—Engine Evaluation .....	80
Table 6.1. Number of Students Included in the Psychometric Analyses .....	81
Table 6.2. Summary <i>P</i> -Values—Operational Items .....	82
Table 6.3. Summary <i>P</i> -Values—Field Test Items.....	83
Table 6.4. Summary Item-Total Correlations—Operational Items.....	84

Table 6.5. Summary Item-Total Correlations—Field Test Items .....	84
Table 6.6. Flagging Criteria for MC Items .....	85
Table 6.7. Flagging Criteria for Partial-Credit Items.....	85
Table 6.8. Suppressed Items .....	86
Table 6.9. Focal and Reference Groups for Gender- and Ethnicity-Based DIF.....	86
Table 6.10. DIF Categories for Dichotomous Items.....	88
Table 6.11. DIF Categories for Polytomous Items .....	88
Table 6.12. DIF Results—Operational Items .....	89
Table 6.13. DIF Results—Field Test Items.....	91
Table 6.14. Unidimensionality: Results from PCA of Residuals .....	93
Table 6.15. Local Independence: Summary of Item Residual Correlations.....	96
Table 6.16. Item Fit: Summary of Infit and Outfit MNSQ Statistics for Items .....	97
Table 6.17. Item Fit: Summary of Infit and Outfit ZSTD Statistics for Items .....	98
Table 6.18. Summary IRT Item Statistics—Operational Items.....	98
Table 6.19. Summary IRT Item Statistics—Field Test Items.....	99
Table 6.20. Scale Score Difference Between 2018 and 2019 Final Recommendations.....	101
Table 6.21. Achievement Level Distributions—2019, 2018, and %Difference.....	102
Table 6.22. Scale Scores by Percentile Rank.....	102
Table 6.23. Effect Size and Horizontal Distance.....	103
Table 6.24. Number of Items with a Large Parameter Change .....	105
Table 6.25. Science Pre- and Post-Equating Comparison.....	106
Table 6.26. Science Achievement Level Distribution for 2018 and 2019 .....	106
Table 6.27. Score Range (LOSS and HOSS) and Assigned Score .....	108
Table 6.28. Conversion of Theta to Scale Scores .....	109
Table 7.1. Final Approved Cut Scores and Impact Data—ELA and Mathematics.....	112
Table 8.1. Number of Students Tested by Demographics—Grade 3.....	114
Table 8.2. Number of Students Tested by Demographics—Grade 4.....	115
Table 8.3. Number of Students Tested by Demographics—Grade 5.....	116
Table 8.4. Number of Students Tested by Demographics—Grade 6.....	117
Table 8.5. Number of Students Tested by Demographics—Grade 7 .....	118
Table 8.6. Number of Students Tested by Demographics—Grade 8.....	119
Table 8.7. Number of Students Tested by Administration Mode .....	120
Table 8.8. Testing Time in Minutes—ELA .....	121
Table 8.9. Testing Time in Minutes—Mathematics.....	122
Table 8.10. Testing Time in Minutes—Science .....	122
Table 8.11. Achievement Level Distributions.....	123
Table 8.12. Scale Score Descriptive Statistics .....	123
Table 8.13. Reporting Category Correlations—Grade 3 .....	124
Table 8.14. Reporting Category Correlations—Grade 4 .....	125
Table 8.15. Reporting Category Correlations—Grade 5 .....	125
Table 8.16. Reporting Category Correlations—Grade 6 .....	126
Table 8.17. Reporting Category Correlations—Grade 7 .....	126
Table 8.18. Reporting Category Correlations—Grade 8 .....	126
Table 8.19. Reporting Category Disattenuated Correlations—Grade 3 .....	127
Table 8.20. Reporting Category Disattenuated Correlations—Grade 4 .....	127
Table 8.21. Reporting Category Disattenuated Correlations—Grade 5 .....	127

Table 8.22. Reporting Category Disattenuated Correlations—Grade 6 .....	128
Table 8.23. Reporting Category Disattenuated Correlations—Grade 7 .....	128
Table 8.24. Reporting Category Disattenuated Correlations—Grade 8 .....	128
Table 8.25. Correlation and Descriptive Statistics of NSCAS and MAP Growth Scores .....	129
Table 8.26. Descriptive Statistics of Score Point Differences from 2018 to 2019.....	130
Table 8.27. Minimum Score Point Differences from 2018 to 2019 .....	130
Table 8.28. Changes in Achievement Level from 2018 to 2019.....	131
Table 9.1. Marginal Reliability of Scale Scores—ELA .....	133
Table 9.2. Marginal Reliability of Scale Scores—Mathematics.....	133
Table 9.3. Marginal Reliability of Scale Scores—Science.....	133
Table 9.4. Marginal Reliability: Variance .....	133
Table 9.5. CSEMs at the Proficient Cut Scores .....	134
Table 9.6. Mean CSEMs by Deciles.....	135
Table 9.7. Classification Accuracy by Achievement Level and Reporting Category—ELA .....	136
Table 9.8. Classification Accuracy by Achievement Level and Reporting Category— Mathematics .....	138
Table 9.9. Classification Accuracy by Achievement Level and Reporting Category—Science.....	140
Table 9.10. Cronbach’s Alpha (Internal Consistency) by Demographics for Science Fixed Forms .....	142
Table 10.1. Sources of Validity Evidence for Each NSCAS Test Purpose .....	145
Table 10.2. Sources of Validity Evidence based on Test Content .....	145
Table 10.3. Sources of Validity Evidence based on Response Process .....	146
Table 10.4. Sources of Validity Evidence based on Internal Structure.....	147
Table 10.5. Sources of Validity Evidence based on Other Variables .....	147
Table 10.6. Interpretive Argument Claims, Evidence to Support the Essential Validity Elements .....	148

### **List of Figures**

Figure 1.1. Principled Test Design Process to Support Test Score Interpretations and Uses ....	18
Figure 1.2. NSCAS Theory of Action.....	20
Figure 2.1. Test Development Process .....	21
Figure 2.2. Adaptive Test Design with Horizontal and Vertical Linking .....	22
Figure 2.3. General Item Sequence for ELA and Mathematics.....	24
Figure 2.4. DOK Box Plots for 2019 Operational and Field Test Items—ELA.....	26
Figure 2.5. DOK Box Plots for 2019 Operational and Field Test Items—Mathematics.....	28
Figure 2.6. Range ALD Example: NSCAS Summative ELA Grade 3.....	32
Figure 3.1. CAP Student Login Screen .....	46
Figure 4.1. Reports Landing Page Example—District Assessment Contact .....	66
Figure 4.2. Matrix Example: Percent Proficient.....	68
Figure 4.3. Matrix Example: Scale Score by Demographics .....	69
Figure 4.4. Matrix Example: Scale Score by Sub-Groups.....	69
Figure 5.1. Adaptive Engine Overview .....	70
Figure 5.2. Shadow Test Approach.....	71
Figure 6.1. Mean Scale Score by Grade .....	103



Figure 6.2. Mean Scale Score Differences Between Adjacent Grades .....	103
Figure 6.3. SD of Scale Score by Grade .....	104
Figure 6.4. Effect Sizes between Adjacent Grades .....	104
Figure 6.5. Cumulative Distribution Function (CDF) .....	104

### **List of Appendices**

Appendix A: Table of Specifications (TOS) .....	A-1
Appendix B: Number of Items by Standard Taken to Committee .....	A-70
Appendix C: Data Review Cheat Sheet.....	A-74
Appendix D: Test Administration Training PowerPoints.....	A-79
Appendix E: Sample Reports .....	A-117
Appendix F: Classical Item-Level Statistics .....	A-135
Appendix G: Summary <i>P</i> -Values by Item Type .....	A-303
Appendix H: Summary Item-Total Correlations by Item Type .....	A-309
Appendix I: Differential Item Functioning (DIF) Item-Level Statistics .....	A-315
Appendix J: IRT Item-Level Statistics.....	A-451
Appendix K: Science Pre- and Post-Equating Results.....	A-611
Appendix L: Achievement Level Distributions & Scale Score Descriptive Statistics by Demographics.....	A-617
Appendix M: Marginal Reliability by Demographics.....	A-624
Appendix N: Scatterplots for Scale Score CSEM .....	A-631

## List of Abbreviations

Below is a list of abbreviations that appear in this technical report.

ALD	achievement level descriptor
API	application program interface
CAP	Comprehensive Assessment Platform
CCR	College and Career Readiness
CSEM	conditional standard error of measurement
DIF	differential item functioning
DNU	Do Not Use
DOK	Depth of Knowledge
DRC	Data Recognition Corporation
EDS	Educational Data Systems
ELA	English Language Arts
ELL	English language learner
ESEA	Elementary and Secondary Education Act
ESC	Education Strategy Consulting
ESU	educational service unit
ETS	Educational Testing Service
FRL	free and reduced lunch
FT	field test
GIS	Group Identification Sheet
HL	horizontal linking
HOSS	highest obtainable scale score
ID	Item-Descriptor
ISR	Individual Student Report
IEP	Individualized Education Plan
IRT	item response theory
IWW	item writer workshop
KSAs	knowledge, skills, and abilities
LEP	limited English proficiency
LOSS	lowest obtainable scale score
MC	multiple-choice
MH	Mantel-Haenszel
MLE	maximum likelihood estimation
NCCRS-S	Nebraska College and Career Ready Standards for Science
NCLB	No Child Left Behind
NDE	Nebraska Department of Education
NeSA	Nebraska State Accountability
NSCAS	Nebraska Student-Centered Assessment System
OIB	ordered item book
OP	operational
PCA	principal component analysis
PP	paper-pencil
RAEL	Recently Arrived Limited English Proficient
SD	standard deviation
SEM	standard error of measurement
SFTP	Secure File Transfer Protocol

SGL..... School Group List  
STARS ..... School-based Teacher-led Assessment and Reporting System  
TAC..... Technical Advisory Committee  
TAM ..... Test Administration Manual  
TCC..... test characteristic curve  
TEI ..... technology-enhanced item  
TOS..... Table of Specifications  
TTS ..... text-to-speech  
UAT ..... user acceptance testing  
UDL..... Universal Design for Learning  
VL..... vertical linking  
VOIP ..... Voice Over Internet Protocol

## Executive Summary

This technical report documents the processes and procedures implemented to support the Spring 2019 Nebraska Student-Centered Assessment System (NSCAS) General Summative English Language Arts (ELA), Mathematics, and Science assessments by NWEA® under the supervision of the Nebraska Department of Education (NDE). The technical report shows how the processes, methods applied, and results relate to the issues of validity and reliability and to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). Below is a high-level summary of each section in the technical report.

### Section 1: Introduction

Students taking the ELA and Mathematics tests were placed into one of the following achievement levels: Developing, On Track, or College and Career Readiness (CCR) Benchmark. Students taking the Science tests were placed into one of the following achievement levels: Below the Standards, Meets the Standards, or Exceeds the Standards. The purposes of the 2019 NSCAS Summative assessments are to measure and report Nebraska students' depth of achievement regarding Nebraska's College and Career Ready Standards for ELA and Mathematics in Grades 3–8 and Nebraska's Science standards for Grades 5 and 8; to report if student achievement is sufficient academic proficiency in ELA and Mathematics to be on track for achieving college readiness; to measure students' annual progress toward college and career readiness in ELA and Mathematics; to inform teachers how student thinking differs along different areas of the scale as represented by the achievement level descriptors (ALDs) as information to support instructional planning; and to assess students' construct relevant achievement in ELA, Mathematics, and Science for all students and subgroups of students.

### Section 2: Test Design and Development

Nebraska's College and Career Ready Standards have been adopted by the Nebraska State Board of Education for ELA, Mathematics, and Science in 2014, 2015, and 2017, respectively. The Spring 2019 NSCAS assessments were aligned to the Nebraska's College and Career Ready Standards for ELA and Mathematics in Grades 3 to 8. The design of the NSCAS Summative assessments is based on a principled approach to test design based on ALDs under which the evidence needed to draw a conclusion about where a student is in their learning of content is made explicit in the ALDs and items are developed according to those evidence pieces. To fully represent the constructs being assessed by the NSCAS to determine if students are ready for college and careers, the adherence to specifications, common interpretations of the standards, and an agreed-upon approach for cognitive complexity across all item types were closely monitored during item, passage, and test development.

### Section 3: Test Administration and Security

The Spring 2019 NSCAS Summative testing window was from March 18 to April 26, 2019, with an extension until May 17, 2019, for a select number of districts. The tests were administered online via NWEA's Comprehensive Assessment Platform (CAP) test management system with paper-pencil versions available as an accommodation. Appropriate accommodations and universal features were provided, and test security was adhered to throughout the entire test administration process for both online and paper-pencil testing. User acceptance testing (UAT) was conducted prior to the operational administration to make sure the technology and item functionality were working properly.

#### **Section 4: Scoring and Reporting**

The online ELA and Mathematics assessments were administered adaptively via NWEA's constraint-based engine, whereas all Science assessments, all paper-pencil tests, and all Spanish versions were administered as fixed-form. All tests were scored with maximum likelihood estimation (MLE) scoring. All steps of scoring for online and paper-pencil went through a quality control process. Score reports were produced and delivered at the individual student level, and aggregated reports were delivered at the school, district, and state levels. A web-based portal referred to as the Matrix also provided visualizations for the NSCAS assessments, including combinations of aggregate and disaggregate information of results by demographics and other filtering options.

#### **Section 5: Constraint-Based Engine**

The NWEA constraint-based engine administers items adaptively to match the ability level of each individual student. It has two stages of consideration as it selects the next item that conforms to the Table of Specifications (TOS) while providing the maximum information about the student based on the student's momentary ability estimate: shadow test approach followed by a variation of the weighted penalty model. Pre-administration simulations were conducted prior to the Spring 2019 operational testing window to evaluate the constraint-based engine's item selection algorithm and estimation of student ability based on the TOS. After the Spring 2019 testing window closed, a post-administration evaluation study was then conducted. Overall, the constraint-based engine performed as expected.

#### **Section 6: Psychometric Analyses**

The following post-administration analyses were conducted for the ELA, Mathematics, and Science assessments: classical item analyses, including item difficulty ( $p$ -value), item discrimination, and item suppression; differential item functioning (DIF) based on gender and ethnicity; item response theory (IRT) calibration; a vertical scale evaluation for ELA and Mathematics; and post-equating checks for Science. The average  $p$ -values ranged from 0.4 to 0.6 across content areas and grades, which falls in the target range. The item-total correlation results appear out of bounds from traditional metrics, but this is because ELA and Mathematics were adaptive. Based on item analysis results and flagging criteria, seven items were suppressed from the 2019 scoring and 12 items were removed from the future item pool. One fixed-form item was also removed from the 2019 scoring. There was no suppression for Science. Most items were categorized as DIF Category A (negligible DIF). Operational item parameter means increased by grade for ELA and Mathematics, as can be expected for vertical scales. Based on an evaluation of the vertical scale, NWEA recommended the following for 2019 scoring, which were approved by NDE: pre-equating for ELA Grades 3–8 and Mathematics Grades 3–6 and horizontal linking for Mathematics Grades 7 and 8.

#### **Section 7: Standard Setting**

No standard setting was held in 2019. Nebraska's statewide assessment system for ELA and Mathematics underwent significant changes between 2016 and 2017, so cut scores for ELA and Mathematics were set following the Spring 2018 administration at standard setting and cut score review meetings from July 26–28, 2018, using the Item-Descriptor (ID) Matching method. The purpose of the standard setting was to set new cut scores for Mathematics, whereas the purpose of the cut score review was to validate the existing cut scores for ELA. No changes were made to the Science standards or assessments, so a standard setting was not necessary.

## **Section 8: Test Results**

More than 22,000 students took the assessment in each grade and content area. Of those students across grades, half are males, half are females, two thirds are white, and about one fifth are Hispanic. Among the students across grades, about 45% to 47% are eligible for free and reduced lunch (FRL), 14–16% have limited English proficient (LEP) status, and 13–16% belong to at least one special education (SPED) category. The 2019 NSCAS assessments were administered online to the extent practical. Less than 1% of students took the assessment in the paper-based version. Most students finished tests within 120 minutes. For ELA, 42–51% of students are at Developing and 48–58% of students are at On Track or CCR Benchmark. For Mathematics, 45–52% of students are at Developing and 48–55% of students are at On Track or CCR Benchmark. For Science, 31–37% of students are at Below the Standards and 63–69% are at Meets or Exceeds the Standards. The mean scale score increases with the grade for ELA and Mathematics, as expected. For each grade and content area, Pearson’s correlation coefficients were calculated between reporting category scores to provide information on score dimensionality. In general, the within-content-area reporting category correlations are higher than the across-content-area reporting category correlations.

## **Section 9: Reliability**

The reliability/precision of the 2019 NSCAS assessments was examined through analysis of measurement error in simulated and operational conditions, including constraint-based engine score precision and reliability, marginal reliability, conditional standard error of measurement (CSEM), and Cronbach’s alpha and standard error of measurement (SEM) for fixed forms. Marginal reliability estimates for the total scores are well above 0.80 (.87 or higher), which is typically considered the minimally acceptable level of reliability. The CSEM represents the degree of measurement error in scale score units and are conditioned on the ability of the student. When applied to an adaptive assessment, the CSEM will vary for the same scale score. It is therefore necessary to report averages. The overall CSEM is slightly higher for ELA than for Mathematics. The low CSEM for Science is expected as its conversion slope is smaller than ELA or Mathematics. Results also suggest that item pools have more items in the middle than at both ends and that more difficult items are needed for both ELA and Mathematics, which is consistent with reliability results. The classification accuracy results suggest that accurate classifications are being made for Nebraska students on the NSCAS assessments.

## **Section 10: Validity**

Validating a test score interpretation is not a quantifiable property but an ongoing process, beginning at initial conceptualization of the construct and continuing throughout the entire assessment process. As the technical report progresses, it covers the different phases of the testing cycle and the procedures and processes applied in the NSCAS, as well as the results. The section revisits phases and summarizes relevant evidence and a rationale in support of any test score interpretations and intended uses based on the *Standards*. The validity argument begins with a statement of the assessment’s intended purposes, followed by the evidentiary framework where available validity evidence is provided to support the argument that the test actually measures what it purports to measure (SBAC, 2016).

## Section 1: Introduction

The purpose of this technical report is to summarize the design, development, administration, technical processes, and results of the Spring 2019 Nebraska Student-Centered Assessment System (NSCAS) General Summative assessments in English Language Arts (ELA) and Mathematics for Grades 3–8 and in Science for Grades 5 and 8 to support test users in evaluating the intended purposes, uses, and interpretations of the test scores. NSCAS was designed by the state of Nebraska with support from its vendor NWEA® to meet the requirements of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) and the federal peer review requirements (U.S. Department of Education, 2018) with an emphasis on using a principled assessment design process.

### 1.1. NSCAS Overview

NSCAS is a statewide assessment system that embodies Nebraska’s holistic view of students and helps them prepare for success in postsecondary education, career, and civic life. It uses multiple measures throughout the year to provide educators and decision makers at all levels with the insights they need to support student learning. The NSCAS Summative assessment, developed specifically for Nebraska and aligned to the state content area standards, may be considered the criterion-referenced, summative measure for the assessment system for most of the Nebraska student population in Grades 3–8.

The Spring 2019 NSCAS assessments were administered online with paper-pencil versions available as an accommodation. They included a variety of item types, including multiple-choice and technology-enhanced items. Student scores were reported as composite scale scores, reporting category scale scores, and achievement levels. The ELA and Mathematics assessments were administered using a multi-stage adaptive design, whereas Science was administered in fixed form online. Students taking the ELA and Mathematics tests were placed into one of the following achievement levels based on their final test scores:

- Developing
- On Track
- College and Career Readiness (CCR) Benchmark

Students taking the Science tests were placed into one of the following achievement levels:

- Below the Standards
- Meets the Standards
- Exceeds the Standards

Items for the ELA and Mathematics tests were aligned to the 2014 and 2015 College and Career Ready Standards, respectively, and came from the item bank that the Nebraska Department of Education (NDE) and Nebraska educators have built over the years, including items field tested in Spring 2018. The tests also included newly developed field test items that will be added to the operational pool for the future depending on the field test data and data review. Items for the Science test came from the operational pool that NDE had built over the previous years and were aligned to the 2010 Nebraska Legacy Standards in Science. A pilot assessment was also administered in Spring 2019 to gain feedback from Nebraska students on newly developed performance tasks for use on the new science assessment that will be aligned to the Nebraska College and Career Ready Standards for Science (NCCRS-S; NDE, 2017).

## 1.2. Background

From 2001 to 2009, Nebraska administered a blend of local and state-generated assessments called the School-based Teacher-led Assessment and Reporting System (STARS) to meet No Child Left Behind (NCLB) requirements. STARS was a decentralized local assessment system that measured academic content standards in Reading, Mathematics, and Science. The state reviewed every local assessment system for compliance and technical quality. NDE provided guidance and support for Nebraska educators by training them to develop and use classroom-based assessments. For accreditation, districts were also required to administer national norm-referenced tests. As a component of STARS, NDE administered one writing assessment annually in Grades 4, 8, and 11. NDE also provided an alternate assessment for students severely challenged by cognitive disabilities.

Nebraska Revised Statute 79-760.03<sup>1</sup> passed by the 2008 Nebraska Legislature requires a statewide assessment of the Nebraska academic content standards for Reading, Mathematics, Science, and Writing in Nebraska's K–12 public schools. The new assessment system was named the Nebraska State Accountability (NeSA). NeSA replaced previous school-based assessments for purposes of local, state, and federal accountability and were phased in beginning in the 2009–2010 school year.

Through the 2015–2016 academic year, assessments in Reading and Mathematics were administered in Grades 3–8 and 11; Science was administered in Grades 5, 8, and 11; and Writing was administered in Grades 4, 8, and 11. The 2015–2016 year was the final administration of the NeSA Reading, Mathematics, and Science tests in Grade 11. Nebraska adopted the ACT for high school testing in 2016–2017. NeSA-ELA tests were also implemented in Spring 2017, replacing NeSA Reading.

NSCAS replaced the NeSA assessments beginning in 2017–2018. Spring 2019 was the second administration of the NSCAS Summative ELA and Mathematics assessments that were administered adaptively, whereas Science continued to be administered as a fixed-form assessment. The new NSCAS Science assessment aligned to the three-dimensional NCCRS-S in Grades 5 and 8 intended to encompass the new content standards and technologies was piloted in March 2019, with a full-scale field test scheduled for Spring 2020 and an operational launch in Spring 2021.

## 1.3. Schedule of Major Events

Table 1.1 presents the major events regarding the development, administration, and reporting of the 2018–2019 NSCAS Summative assessments, including the new Science assessment. As shown in the table, NDE involves educators throughout the development process to produce customized items and provide an invaluable professional development opportunity, including ALD and item development meetings, item review meetings, alignment studies, and standard settings. For example, educators participated in standard setting and cut score review in 2018 and the Mathematics alignment study in 2019. They have also been involved with the development of the new NSCAS Science assessment that was piloted in March 2019.

---

<sup>1</sup> <https://nebraskalegislature.gov/laws/statutes.php?statute=79-760.03>



**Table 1.1. Schedule of Major Events**

Event	Date(s)
ELA passage review	April 19–20, 2018
Item writer workshop (IWW) in ELA and Mathematics	June 4–7, 2018
Content and bias review in ELA and Mathematics	July 17–20, 2018
Fall 2018 regional workshop	October 9–12, 2018
Summative test administration training	February 18–22, 2019
Science test development workshop	July 17–20, 2018
Science task review committee	September 11–12, 2018
Technical Advisory Committee (TAC) meeting	March 22, 2019
Science pilot testing window and cognitive labs	March 4–15, 2019
Spring 2019 operational testing window	March 18 – April 26, 2019
Make-up testing window	April 29 – May 3, 2019
Science achievement level descriptor (ALD) workshop	May 1–2, 2019
Mathematics alignment study	July 29 – August 8, 2019
NDE and districts review preliminary data and submit updates	August 12, 2019
Data review with NDE for ELA and Mathematics	September 5–6, 2019
Delivery of online reports	September 12, 2019
Delivery of printed Individual Student Reports (ISRs)	starting October 2, 2019

#### 1.4. Building a Validity Argument

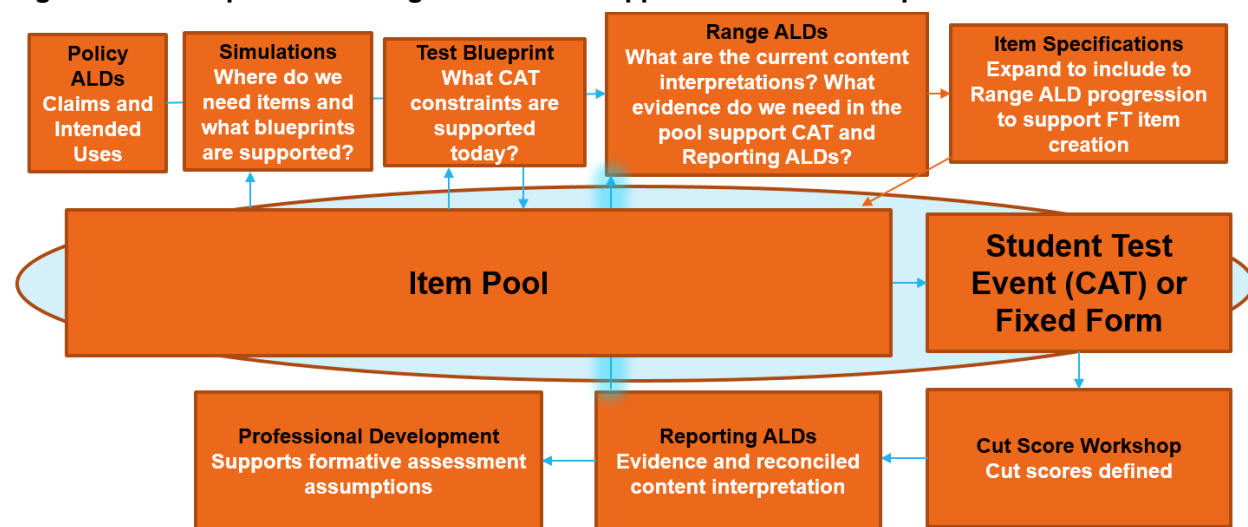
The NSCAS Summative assessments have been developed based on a principled approach to test design that centers around achievement level descriptors (ALDs) and conceptualizing test score use as part of a broader solution to achieve important outcomes for test users. The evidence needed to draw a conclusion about where a student is in their learning of content is made explicit in the ALDs and items are developed according to those evidence pieces (Huff, Warner, & Schweid, 2016; Egan, Schneider, & Ferrara, 2012; Schneider & Johnson, 2018). This approach builds validity evidence into the design from the very beginning of the process, which is especially important when the assessments are intended to support interpretations regarding how student learning grows more sophisticated over time (Pellegrino, DiBello, & Goldman, 2016). The purposes of a test design centered in ALDs include the following:

- To show how students increase in their reasoning with specific content across achievement levels to support collecting purposeful evidence of what mastery of college and career readiness means
- To support teachers in making more accurate inferences about what students know and can do

ALDs demonstrate how skills become more sophisticated as achievement and performance increase (Schneider et al., 2013). Such skill advancement is often related to increases in content difficulty and reasoning complexity and a reduction in the supports required for students to demonstrate what they know within a task or item. This use of ALDs helps teachers interpret the student work evidence to better identify where a student is in their learning and what they need next. Using a principled test design process supports teachers in better understanding that a single standard has easier and more difficult representations and that the goal of instruction is to support the development of cognitive skills in addition to content-based skills.

Figure 1.1 presents the balanced approach NDE took in the development process of the NSCAS Summative assessments. Policy ALDs are high-level expectations of student achievement within each achievement level across grades. Range ALDs are within-standard learning progressions that describe the knowledge and skills students at each achievement level should be able to demonstrate. They describe the current stage of learning within the standard and explicate observable evidence of achievement, demonstrating how skills change and become more sophisticated across achievement levels for each standard. Reporting ALDs are finalized versions of the Range ALDs supported by evidence from the test scale that were created after the final cut scores were adopted. Content interpretations were finalized after the standard setting and, as the highlighted blue arrow shows, are used to support item specifications to ensure a stable, comparable construct over time.

**Figure 1.1. Principled Test Design Process to Support Test Score Interpretations and Uses**



With a principled approach to test design, ALDs may be viewed as the score interpretation, or the construct interpretive argument described by Kane (2013). For ALDs to be the foundation of test score interpretation, they should reflect more complex knowledge, skills, and abilities (KSAs) as the achievement levels increase (Schneider, Huff, Egan, Gaines, & Ferrara, 2013). As such, NDE developed ALDs to articulate the following:

- The observable evidence teachers and item developers should elicit to draw conclusions about a student’s current level of performance
- What that evidence looks like when students are in different stages of development represented by different achievement levels
- How the student is expected to grow in reasoning and content skill acquisition across achievement levels within and across grades

Using ALDs, the NSCAS item bank has been aligned to the standards, represents the intended blueprint, and provides supports for students at all levels of proficiency within on-grade content. ALDs were developed in an iterative manner based on feedback from educators (Plake, Huff, & Reshetar, 2010), with the final ALDs providing the interpretive argument regarding what test scores mean. By developing ALDs this way, Nebraska is communicating how standards are interpreted for assessment purposes, how tasks can align to a standard but not be of sufficient difficulty and depth to represent mastery, and what growth on the test score continuum represents.

#### *1.4.1. Intended Purposes and Uses of Test Results*

Building a validity argument begins with identifying the purposes of the assessment and the intended uses of its test scores. The following are purposes of the NSCAS assessments:

1. To measure and report Nebraska students' depth of achievement regarding Nebraska's academic content standards
2. To report if student achievement is sufficient academic proficiency in ELA and Mathematics to be on track for achieving college readiness.
3. To measure students' annual progress toward college and career readiness.
4. To inform teachers how student thinking differs along different areas of the scale as represented by the ALDs as information to support instructional planning.
5. To assess students' construct relevant achievement in ELA, Mathematics, and Science for all students and subgroups of students.

Ultimate use of test scores is determined by Nebraska educators. However, some intended uses of the NSCAS test results include the following:

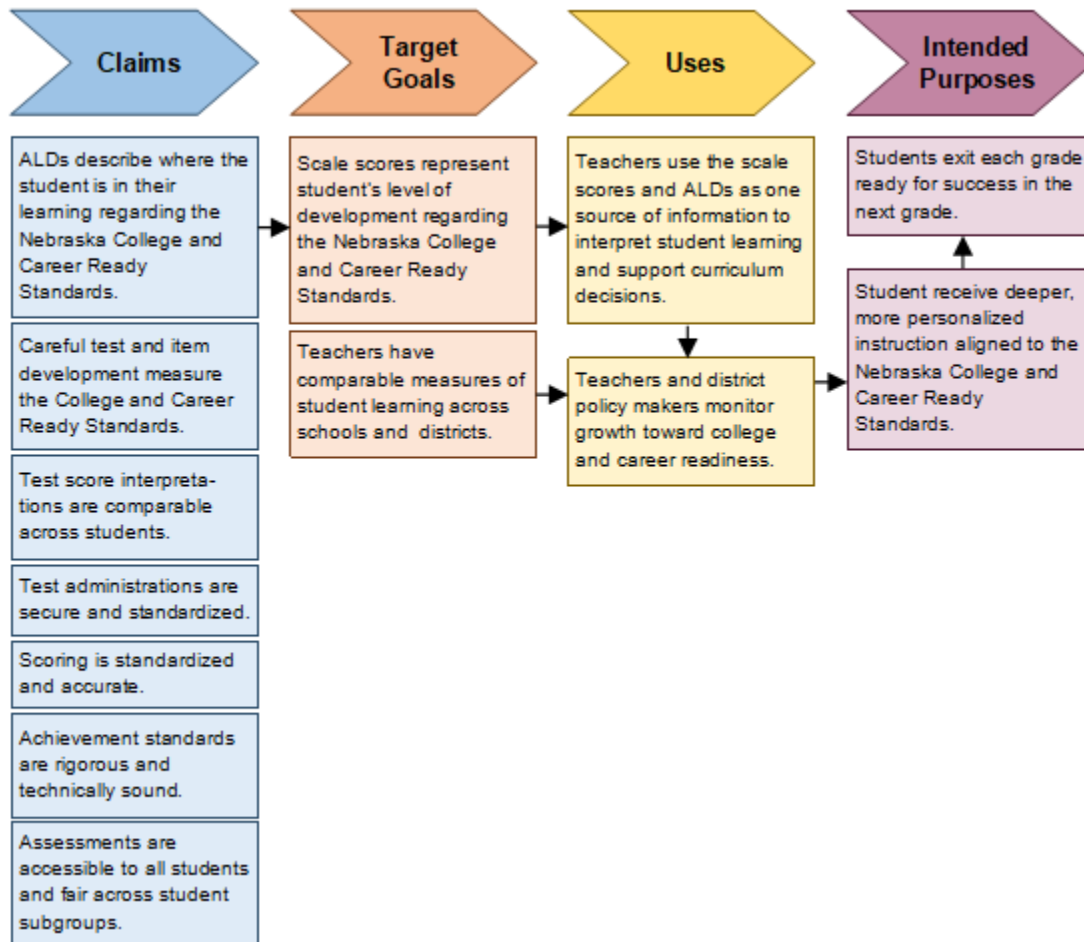
- To supplement teachers' observations and classroom assessment data and to improve the decisions teachers make about sequencing instructional goals, designing instructional materials, and selecting instructional approaches for groups and individuals
- To identify individuals for summer school and other remediation programs
- To gauge and improve the quality of education at the class, school, system, and state levels throughout Nebraska
- To assess the performance of a teacher, school, or system in conjunction with other sources of information

#### *1.4.2. Theory of Action*

A theory of action is a tool that connects test users and their needs to decisions made during test design and development. In other words, it connects the design of the assessment, such as decisions about what evidence to collect and how to provide that evidence, to the claims that test score interpretation and use contribute to a positive solution to the broader problem for the test user. Figure 1.2 presents the theory of action for the NSCAS system. The ultimate intended purpose of NSCAS is to have students exiting each grade ready for success in the next grade. Evidence to determine if the assessment system is supporting its intended purposes across time may include the following:

1. Does Nebraska have increases in percentages of students who are becoming on track for college and career readiness?
2. Are students who are at or above On Track in one year likely to be On Track or above the following year?
3. Are students who are at or above On Track across time likely to be identified as On Track on an assessment of college or career readiness when scores are matched?

Figure 1.2. NSCAS Theory of Action

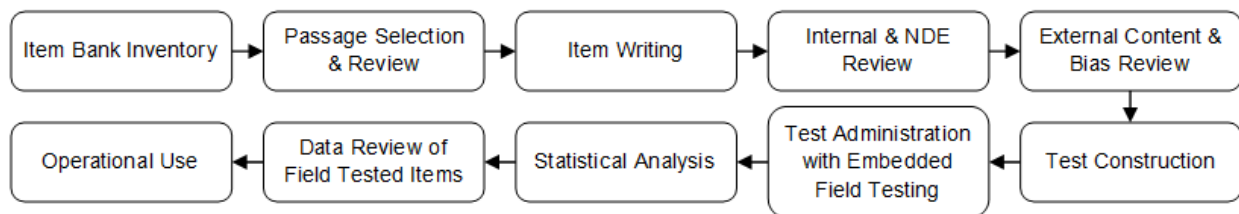


## Section 2: Test Design and Development

This section describes the test design and development processes for the 2019 NSCAS Summative assessments. Content development for the new three-dimensional Science assessment began in Summer 2018 with the pilot occurring in March 2019, but the Spring 2019 operational Science assessment for Grades 5 and 8 continued to be administered in fixed form using existing items. Information regarding the Science pilot development and administration is provided in a separate report (NWEA, 2019c), and a description of the current Science item development process is included in previous NeSA technical reports (e.g., DRC, 2017).

As Nebraska transitioned to an adaptive administration for ELA and Mathematics in 2017–2018, the need to build a large, robust item bank was a key requirement, and the development of new scales had to be accomplished concurrently with thinking about the development of ALDs. Development to support building of a bank to sufficiently support adaptive testing continued for 2018–2019 to have enough content available to populate field test slots in Spring 2019. Items were written by educators in an item writer workshop (IWW) and by independent contractors. Once initial item development was completed, all items were taken to content and bias review meetings with Nebraska educators. Items that survived these meetings were considered for the field test pool. Figure 2.1 outlines the general steps taken to develop the passages and items.

**Figure 2.1. Test Development Process**



### 2.1. Test Designs

Table 2.1 summarizes the different versions of the NSCAS Summative assessments available for 2019. Table 2.2 presents the number of items and points possible on each online and paper-pencil test form. Science had one form per grade, and all Science items were multiple choice, so the points possible was a fixed number. The paper-pencil forms served as accommodated versions that contained only operational items and were slightly longer than the adaptive assessments to support comparable levels of test score precision.

**Table 2.1. Available NSCAS Summative Assessments in 2019**

Content Area	Grade(s)	Available Assessments*				
		Online	PP	Spanish Online	Spanish PP	Breach
ELA	3–8	Adaptive (48 total per grade, 41 OP + 7 FT/VL)	One form per grade (48 OP)	Fixed (translation of PP form)	Same form as Spanish online	2018 PP form
Mathematics	3–8	Adaptive (48 total per grade, 41 OP + 7 FT/VL)	One form per grade (48 OP)	Fixed (translation of PP form)	Same form as Spanish online	2018 PP form
Science	5	Fixed (same form as PP)	One form (50 OP)	Fixed (translation of online form)	Same form as Spanish online	2018 PP form
	8	Fixed (same form as PP)	One form (60 OP)	Fixed (translation of online form)	Same form as Spanish online	2018 PP form

\*OP = operational. PP = paper-pencil. FT = field test. VL = vertical linking.

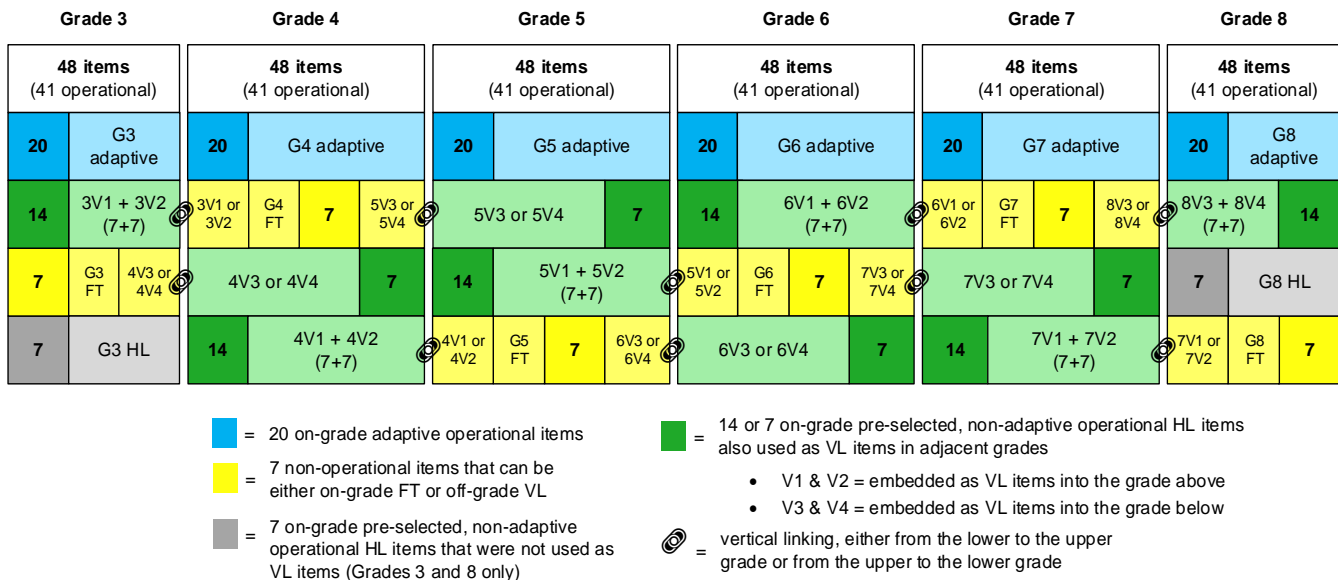
**Table 2.2. Number of Items and Points Per Test**

Content Area	Grade	Online						Paper-Pencil	
		Operational		FT/VL*		Total		#Items	#Points
		#Items	#Points	#Items	#Points	#Items	#Points		
ELA	3	41	47–51	7	7–10	48	54–61	48	59
	4	41	48–50	7	7–10	48	55–60	48	58
	5	41	51–54	7	7–10	48	58–64	48	52
	6	41	49–54	7	7–10	48	56–64	48	56
	7	41	50	7	7–10	48	57–60	48	53
	8	41	52–57	7	7–10	48	59–67	48	51
Mathematics	3	41	45	7	7–8	48	52–53	48	50
	4	41	45	7	7–8	48	52–53	48	50
	5	41	45	7	7–8	48	52–53	48	50
	6	41	45	7	7–8	48	52–53	48	52
	7	41	45	7	7–8	48	52–53	48	50
	8	41	45	7	7–8	48	52–53	48	50
Science	5	50	50	–	–	50	50	50	50
	8	60	60	–	–	60	60	60	60

\*FT/VL = field test/vertical linking. Items in this slot are either FT or VT items.

Figure 2.2 illustrates the online adaptive test design for the NSCAS ELA and Mathematics assessments using both horizontal linking (HL) and vertical linking (VL) anchor items. All students saw a total of 48 items (41 operational + 7 non-operational). Of the 41 operational items, 21 of them were non-adaptive pre-selected HL anchors. The remaining 20 operational items were selected adaptively based on student ability level. Thus, the test design is best classified as a multi-staged adaptive assessment in which students first receive the fixed anchor sets that act as a locator with which to begin adaptive selection for the second portion of the test. Each student also saw one set of 7 non-operational items that were either on-grade field test or off-grade VL items.


**Figure 2.2. Adaptive Test Design with Horizontal and Vertical Linking**



Horizontal linking occurred within the same grade to establish the scale across the different sets of items that students received. As shown in Table 2.3, each student saw a total of 21 HL items during their test administration. HL items were divided into Form 1 (i.e., horizontal anchor core), Form 2a (i.e., horizontal anchor Set A), and Form 2b (i.e., horizontal anchor Set B). All students in Grades 4–7 got Form 1 with 14 core items, while 50% got Set A and the other half got Set B (14 + 7 = 21). Students in Grades 3 and 8 received 7 core items and both Set A and Set B (7 + 7 + 7 = 21). Each HL item set had 7 items and was labeled as V1, V2, V3, V4, or HL in Figure 2.2. Items from the V1 and V2 sets were embedded as VL items in the grade above, whereas items from the V3 and V4 sets were embedded as VL items in the grade below. All VL items therefore also served as HL items in adjacent grades. The 7 HL core items specific to Grades 3 and 8 (as shown in gray boxes in Figure 2.2) were not used as VL items.

**Table 2.3. Horizontal Linking Configuration**

Grade	Horizontal Form 1 (core)			Horizontal Form 2a (Set A)			Horizontal Form 2b (Set B)			Total #HL Items Per Student
	Item Set(s)	#Items	%N	Item Set	#Items	%N	Item Set	#Items	%N	
3	HL	7	100%	V1	7	100%	V2	7	100%	21
4	V1+V2	14	100%	V3	7	50%	V4	7	50%	21
5	V1+V2	14	100%	V3	7	50%	V4	7	50%	21
6	V1+V2	14	100%	V3	7	50%	V4	7	50%	21
7	V1+V2	14	100%	V3	7	50%	V4	7	50%	21
8	HL	7	100%	V3	7	100%	V4	7	100%	21

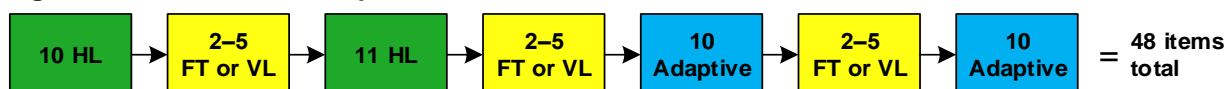
Vertical linking connected adjacent grades in a chain pattern (e.g., Grades 3/4, Grades 4/5, etc.). The adjacent grades (e.g., a Grade 3 student and a Grade 4 student) took the same set of anchor items to put the grades on the same scale, as shown by  in Figure 2.2. Students received either 7 non-operational off-grade VL items or 7 non-operational on-grade field test items during testing. For example, if Student A got a set of VL items, they did not receive any field test items. If Student B got field test items, they did not receive any VL items. Students in Grades 4–7 got one of four VL sets (either V1, V2, V3, or V4). Students in Grades 3 and 8 got one of two VL sets (either V3 or V4 for Grade 3 and either V1 or V2 for Grade 8). Each grade and content area assessment had about 200 field test slots for a total of approximately 2,400 field test items. To verify the vertical scales, VL items were embedded into field test slots in each grade. The design was originally intended to have a minimum of 1,250 student responses for each VL anchor and a minimum of 750 student responses for each field test item. In 2019, the minimum student responses for each VL anchor was changed from 1,250 to 1,000 to allow more field test items.

For Grades 4–7, the first 21 operational items were administered as 7 HL items from either Set A or Set B followed by 14 HL core items. For Grades 3 and 8, the first 21 operational items were administered as 7 items from Set A, 7 items from Set B, and then 7 core items. The 22nd operational item was then adaptively selected based on student responses to operational items 1–20; the 23rd operational item was adaptively selected based on the previous 1–21 operational items; etc. The “n-1” approach was applied, where the (n+1)th item was selected based on (n-1) items so that item selection and rendering could be quick.



As shown in Figure 2.3, the full sequence of items started with 10 HL items, followed by 2–5 field test or VL items, 11 more HL items, 2–5 field test or VL items, 10 adaptive operational items, 2–5 field test or VL items, and 10 more adaptive operational items. However, the item sequence was implemented as “preferred position” to allow the constraint-based engine to accommodate various constraints. The preferred position for the field test/VL item blocks was set to start at the 11th, 24th, and 37th position, but the actual sequence could be different. In addition, ELA field test and VL items, due to passages, were grouped to have 4–5 items and therefore only had two blocks of field test/VL items instead of three. The locations of the item blocks could also vary from one assessment to the next.

**Figure 2.3. General Item Sequence for ELA and Mathematics**



## 2.2. Academic Content Standards

As stated in Nebraska Revised Statute 79-760.01<sup>2</sup> that was effective as of August 30, 2015<sup>3</sup>:

*“The State Board of Education shall adopt measurable academic content standards for at least the grade levels required for statewide assessment pursuant to section 79-760.03. The standards shall cover the subject areas of reading, writing, mathematics, science, and social studies. The standards adopted shall be sufficiently clear and measurable to be used for testing student performance with respect to mastery of the content described in the state standards. The State Board of Education shall develop a plan to review and update standards for each subject area every seven years. The state board plan shall include a review of commonly accepted standards adopted by school districts.”*

On September 5, 2014, the Nebraska State Board of Education adopted Nebraska’s College and Career Ready Standards for ELA. On September 4, 2015, the Nebraska State Board of Education adopted Nebraska’s College and Career Ready Standards for Mathematics. On September 8, 2017, the Nebraska State Board of Education approved the NCCRS-S. These were implemented in the Spring 2019 pilot administration, although the operational NSCAS Science assessments continued to be aligned to the 2010 Science standards.

## 2.3. Table of Specifications (TOS)

The 2018–2019 NSCAS Summative blueprints are embedded in the NSCAS Table of Specifications (TOS) that indicate the range of test items included for each standards indicator in each content area. The adaptive test was constrained to make sure each student received items within the identified ranges. The 2018–2019 test fixed-forms and adaptive forms were not an exact match to the TOS given the attributes of available items in the item bank. Future forms will adhere more closely to the TOS as more items are available. Appendix A presents the TOS for each content area. The TOS for Science is different from ELA and Mathematics in that the total number of items is provided at the grade-level standard rather than at the indicator level. This decision was made based on input received from Science content experts from across the state. All indicators under a tested grade-level standard may be present on the Science test.

<sup>2</sup> <https://nebraskalegislature.gov/laws/statutes.php?statute=79-760.01>

<sup>3</sup> <https://www.education.ne.gov/contentareastandards/>



## 2.4. Item Types

Table 2.4 presents the item types available for the online ELA and Mathematics adaptive tests. The paper-pencil tests included multiple-choice, multiselect, and composite items. The operational fixed-form NSCAS Science assessment included multiple-choice items only.

**Table 2.4. Item Types for Online ELA and Mathematics**

Item Type	Description
Multiple-Choice (Choice)	Students select one response from multiple options.
Multiselect (Choice Multiple)	Students select two or more responses from multiple options. Some multiselect items are also two-point items for which students can earn partial credit.
Hot Text	Students select a response from within a piece of text or a table of information (e.g., word, section of a passage, number, symbol, or equation), which highlights the selected text. Some hot text items are also two-point items for which students can earn partial credit.
Text Entry	Students input answers using a keyboard.
Composite	Students interact with multiple interaction types included within a single item. Students may receive partial credit for composite items.
Drag & Drop	Students select an option or options in an area called the toolbar and move or “drag” these options (e.g., words, phrases, symbols, numbers, or graphic elements) to designated containers on the screen. Drag-and-drop items can include a click and click functionality in which students select the option and select the container it goes into instead of physically dragging it.
Gap Match	A type of drag-and-drop item in which students select one or more answer options from the item toolbox and populate a defined area, or "gap."
Graphic Gap Match	A type of drag-and-drop item in which students move one or more answer options from the toolbox and populate a defined area, or "gap," that has been embedded within an image in the item response area.

## 2.5. Depth of Knowledge (DOK)

With a principled approach to test design based on ALDs, increases in cognitive processing complexity (i.e., DOK levels) are intended to be embedded into evidence statements across achievement levels in a cogent way and to interact with content. In this way, the features of cognitive processing, content difficulty, and context interact to affect item difficulty. A principled approach to test design is intended to support the validity of inferences about the student’s stage of learning and the content validity of the assessment as a measure of student achievement. Under such a score interpretation model, construction of test blueprints should eventually not treat DOK as a separate blueprint constraint. Instead, DOK should be present as evidence embedded in to a descriptor for an achievement level that supports interpretations regarding the stage of thinking sophistication the student is at during the time of the test event. The items found within each achievement level should match the ALDs. The degree of alignment of items to the assessment, a component of the evidence gathered to support a validity framework, should focus on the degree of concurrence in the DOK and content alignment of items within an achievement level to the associated ALDs.

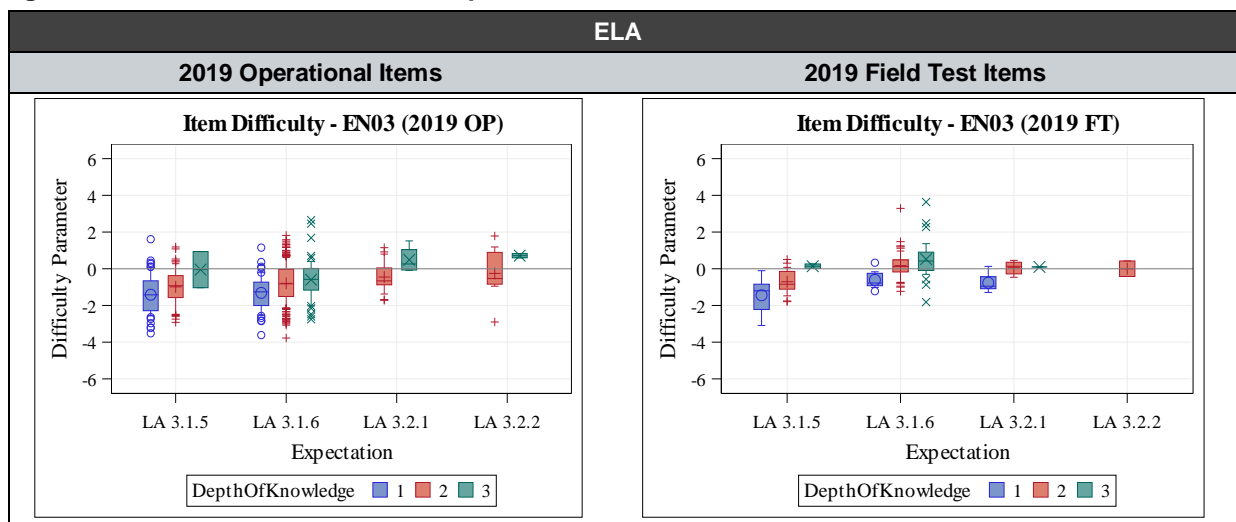
To ensure that the NSCAS assessments include a deep pool of items that span a full range of cognitive levels and skills, each item was evaluated and tagged with one of the following DOK levels (Webb, 1997). DOK Level 4: Extended Thinking items are not included because the tests do not contain any extended-response items or performance tasks.

- DOK 1: Recall & Reproduction
- DOK 2: Skill & Concepts
- DOK 3: Strategic Thinking

Items at DOK 2 and 3 require inferential thinking. DOK 3 items typically demand that students analyze and synthesize concepts from various parts of a text or from the text as a whole. ELA passages demonstrate varying degrees of complexity to support students at all levels of achievement. Because the NSCAS ELA and Mathematics tests were adaptive, the overall distribution of DOK for any given test event varied based on individual student achievement and other factors. In February 2018, the state adopted the policy that Developing items could be at or below cognitive level of the standards, On Track items could be at the cognitive level of the standards, and CCR Benchmark items could be at or above the cognitive level of the standards. This policy decision influenced the development of the ALDs and the review of field test items.

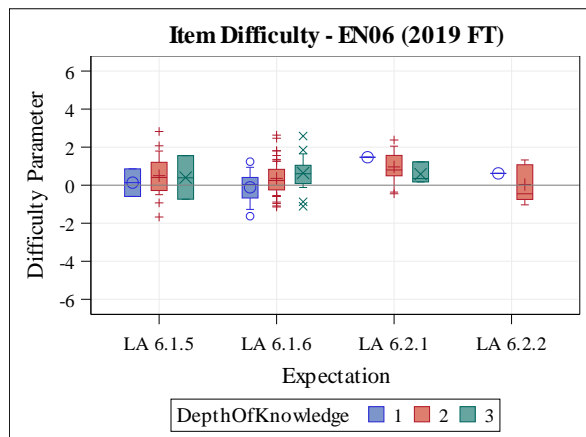
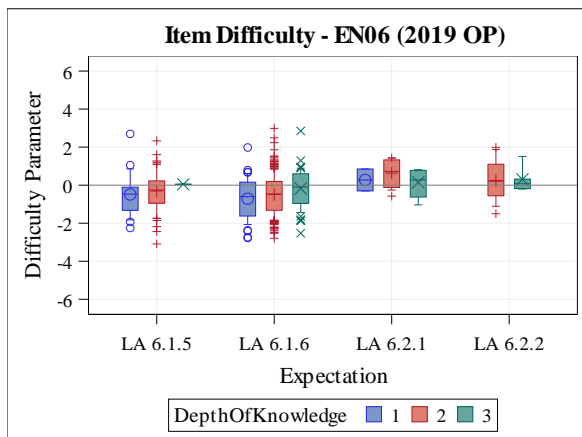
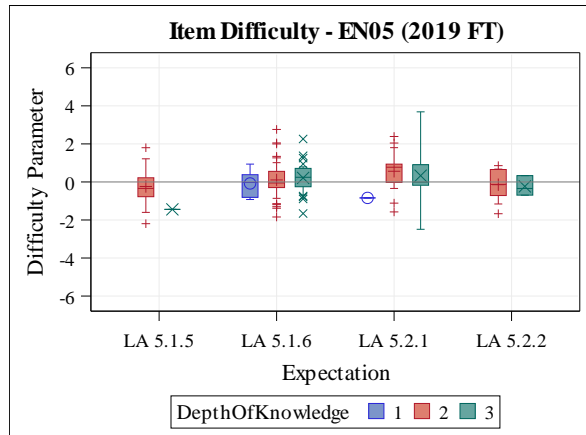
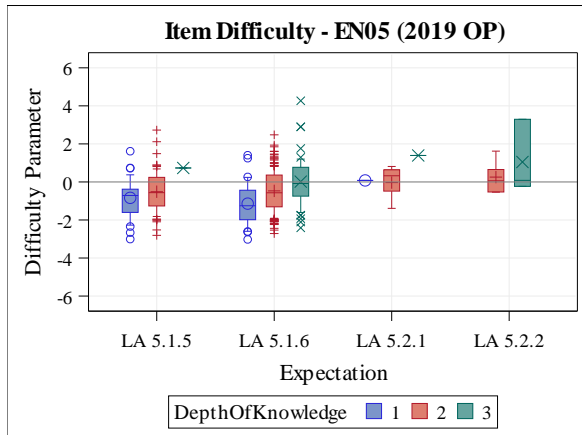
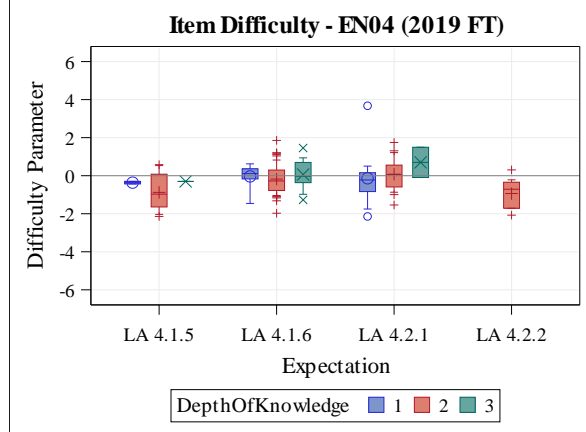
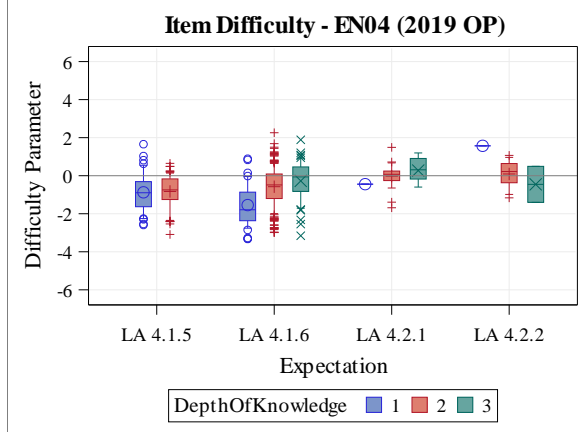
Figure 2.4 and Figure 2.5 present boxplots of item DOK levels based on the state’s interpretation of DOK for the 2019 operational item pool and 2019 field test items. ELA items were largely successfully developed to match intended content and cognitive complexity, including higher-order thinking as shown by the trends of items increasing in difficulty on average for the same indicator as they increase in DOK with the range of item difficulties not being unreasonably restricted depending on the level of cognitive complexity. A different trend is seen Mathematics. The state considers items that measure procedural knowledge in isolation as DOK 1 and items that measure procedural knowledge in a practical real-world context as an increase in depth of knowledge (DOK 2). The data trends for these DOK levels are largely similar item difficulties not being unreasonably restricted for DOK 1 and DOK 2 items with opportunities to develop DOK 3 items in standards in the future based on the state policy decision in February 2018. As the Range ALDs will be used in the future to elicit item content, it is expected that trend data based on a priori ALD level classifications will produce expected trends.

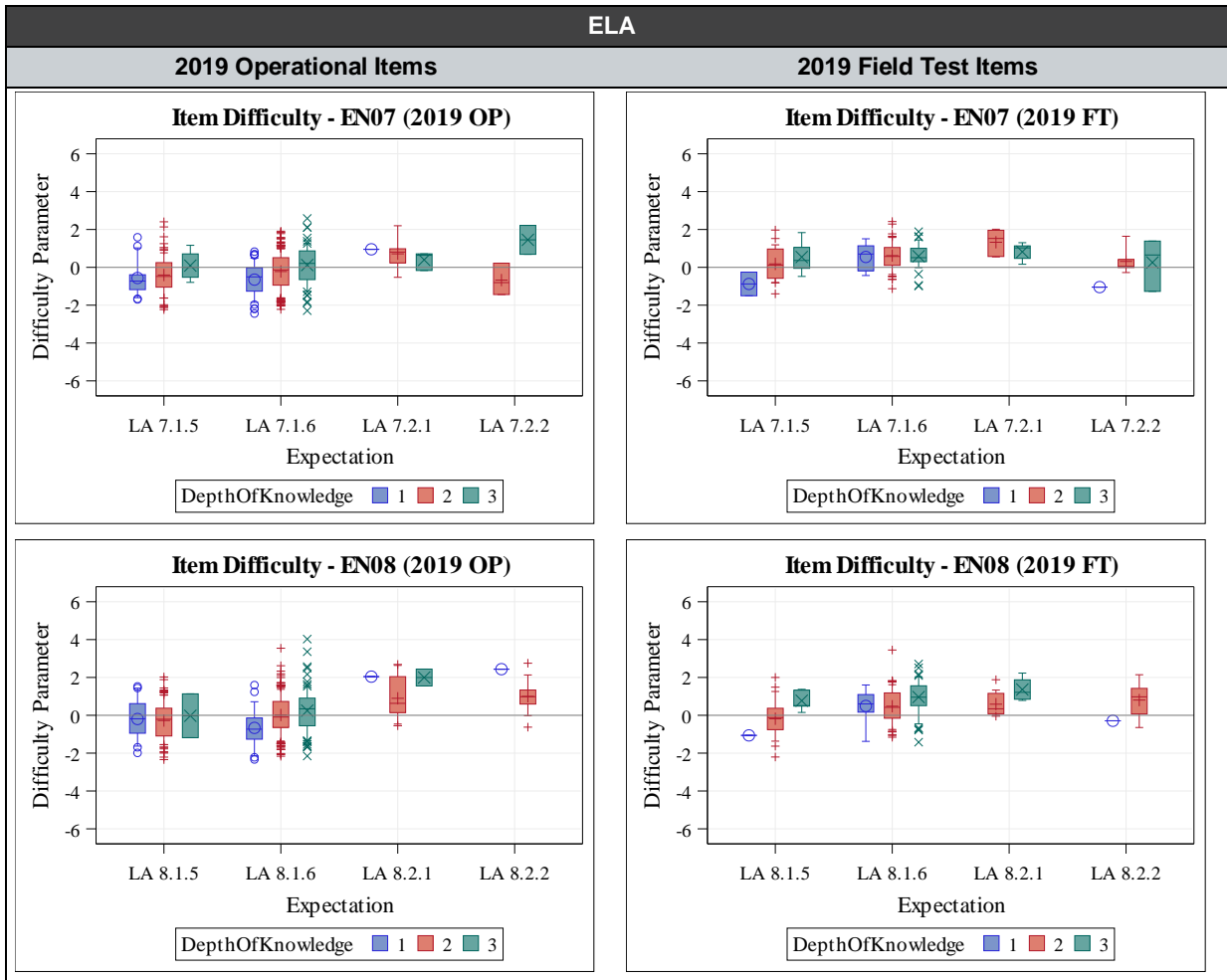
**Figure 2.4. DOK Box Plots for 2019 Operational and Field Test Items—ELA**



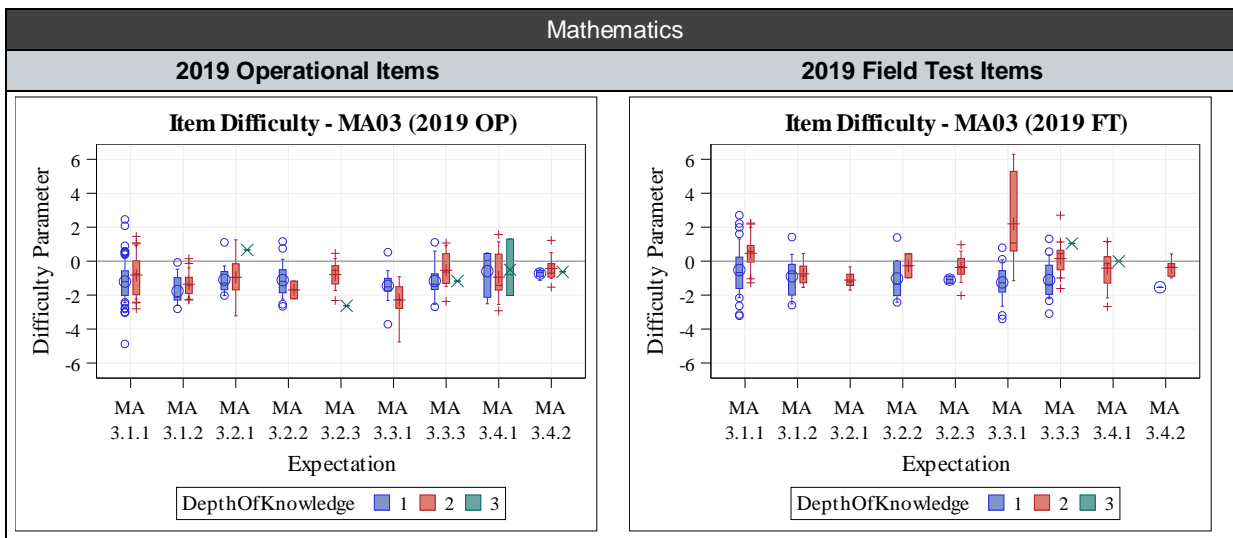
2019 Operational Items

2019 Field Test Items



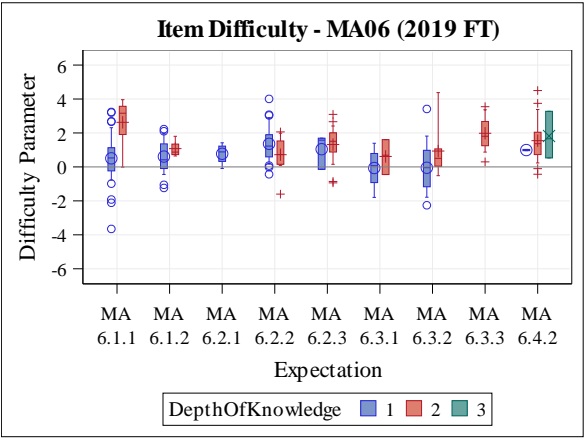
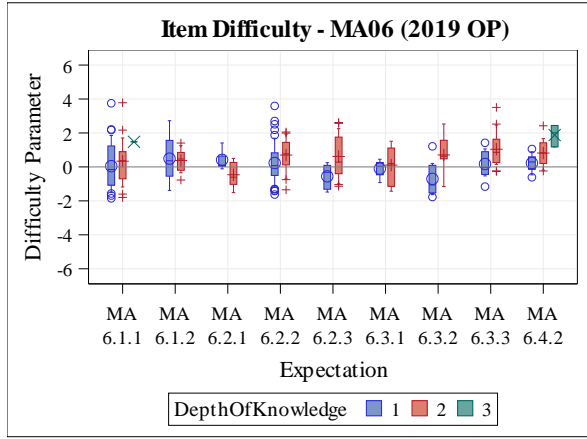
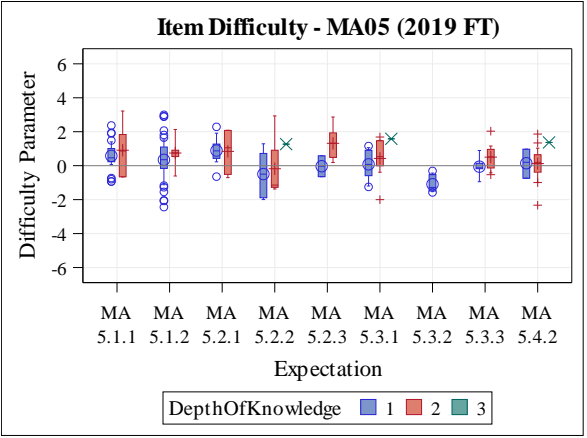
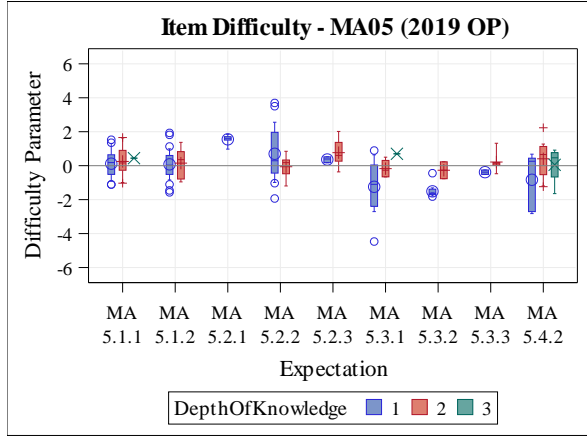
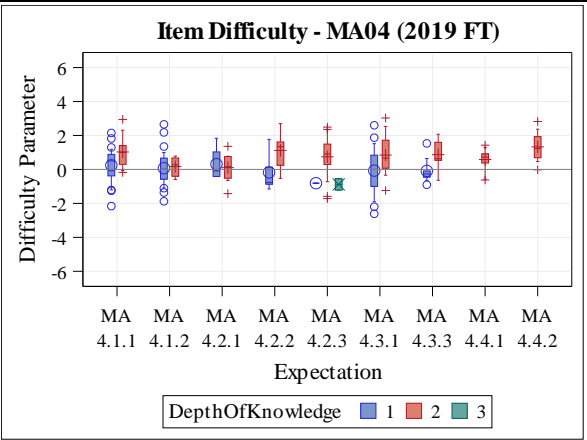
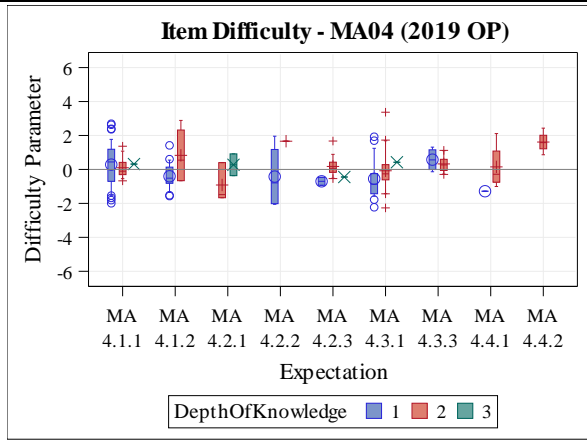


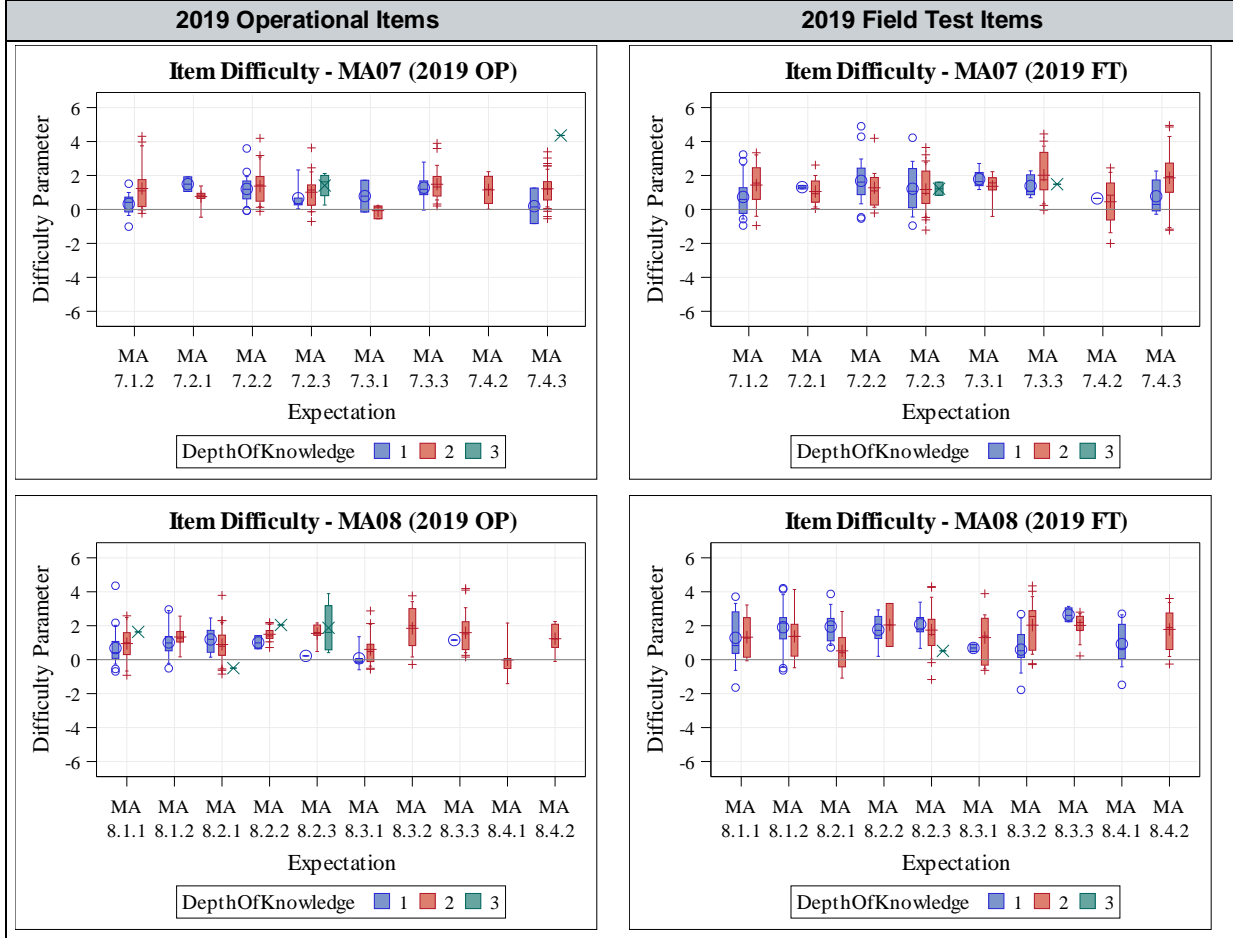
**Figure 2.5. DOK Box Plots for 2019 Operational and Field Test Items—Mathematics**



2019 Operational Items

2019 Field Test Items





## 2.6. ALD Development

The NSCAS ALDs were developed based on the following ALD development stages proposed by Egan, Schneider, and Ferrara (2012) to correspond with the closely linked uses of ALDs in test development and score reporting. ALD development using this model is consistent with a construct-centered approach to assessment design (Messick, 1994).

1. Policy ALDs: High-level expectations of student achievement within each achievement level across grades, often defined by the state
2. Range ALDs: Detailed descriptions of each achievement level by grade that show students' increasing ability to apply practices and concepts
3. Reporting ALDs: Reflect student performance based on the final approved cut scores

### 2.6.1. Policy ALDs

The following Policy ALDs were developed to communicate the vision of what a test score is intended to represent, or where a student is in their learning regarding the content standards. When carefully crafted, Policy ALDs can be viewed as the assessment claim because they set the tone for how the content and cognitive demand is intended to be articulated along the test scale. The Nebraska Policy ALDs guide the establishment of the intended policy outcomes NDE desires for Nebraska students.

- Developing learners do not yet demonstrate proficiency in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards.
- On Track learners demonstrate proficiency in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards.
- CCR Benchmark learners demonstrate advanced proficiency in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards.

### 2.6.2. Range ALDs

Range ALDs provide the intended content-based interpretations of what test scores within an achievement level represent and explicate observable evidence of achievement, demonstrating how the skill changes and becomes more sophisticated across achievement levels for each standard and achievement level on an assessment. Teachers can use the Range ALDs to determine how students with different scores within different achievement levels may differ in their abilities. Range ALDs for ELA and Mathematics were developed in 2018 and refined during the July 2018 standard setting and cut score review meetings. Range ALDs are being generated for the new Science assessment aligned to the NCCRS-S, beginning with an ALD workshop in May 2019 (NWEA, 2019d). These Science ALDs are still in draft form.

To develop the ELA Range ALDs, educators at the July 2018 cut score review meeting used the ALDs from the original standard setting to develop a first draft. After the cut score review, NWEA reviewed the draft ALDs, editing for consistency of language and clarity in a second draft and considering the final approved cut scores. Next, NWEA worked across grades to ensure a logical vertical progression and consistent language between the grades. Once a coherent and cohesive third draft was created, it was sent to NDE for review. NWEA implemented NDE's feedback and sent the resulting fourth draft back to NDE for an additional review. NDE signed off on this document, creating the current version of the ELA ALDs available online at <https://www.education.ne.gov/assessment/nscas-general-summative-assessment/nscas-english-language-arts-ela/>.

To develop the Mathematics Range ALDs, an educator committee was convened in April 2018 to review a first draft. NWEA and NDE then engaged in an extensive revision process that involved several iterations of rework. The draft ALDs were brought to the July 2018 standard setting meeting where they were reviewed and refined by educators based on the cut scores. After receiving the final approved cut scores, NWEA reconciled the ALDs based on item content, participant recommendations, and the final cut scores consistent with recommended practice (Egan et al., 2012). Those edits were used to inform changes throughout the ALDs. These updates were shared with NDE for feedback. After receiving NDE's feedback, NWEA made the requested edits or responded to the posted questions. The files were then formatted and submitted to NDE. The final Mathematics ALDs are available online at <https://www.education.ne.gov/assessment/nscas-general-summative-assessment/nscas-mathematics/>.

Figure 2.6 presents an example of the Range ALDs for ELA Grade 3. The progression descriptor (i.e., Developing, On Track, and CCR Benchmark) describes where a student is in their learning regarding the standard. Within a single expectation (e.g., LA 3.1.5.a) can be ranges of content- and thinking-skill difficulty that describe different stages of reasoning.

Figure 2.6. Range ALD Example: NSCAS Summative ELA Grade 3

ALD	Indicator No.	Indicator Text	Developing	On Track	CCR Benchmark
			With a range of texts with text complexity commonly found in Grade 3, a student performing in Developing can likely	With a range of texts with text complexity commonly found in Grade 3, a student performing in On Track can likely	With a range of texts with text complexity commonly found at the intersection of Grade 3 and Grade 4, a student performing in CCR Benchmark can likely
<b>Reading Vocabulary</b>					
	LA 3.1	<b>Reading:</b> Students will learn and apply reading skills and strategies to comprehend text.			
	LA 3.1.5	<b>Vocabulary:</b> Students will build and use conversational, academic, and content-specific grade-level vocabulary.			
	LA 3.1.5.a	Determine meaning of words through the knowledge of word structure elements, known words, and word patterns (e.g., contractions, plurals, possessives, parts of speech, syllables, affixes, base and root words, abbreviations).	Identify basic word structure elements and word patterns to determine meaning of words (e.g., plurals, parts of speech, syllables).	Apply knowledge of word structure elements, known words and word patterns to determine meaning of words (e.g., contractions, plurals, possessives, parts of speech, syllables, affixes, base and root words, abbreviations).	Analyze complex word structure elements, known words and word patterns to determine meaning of words (e.g., contractions, plurals, possessives, parts of speech, syllables, affixes, base and root words, abbreviations).
	LA 3.1.5.b	Apply context clues (e.g., word, phrase, and sentence clues) and text features to help infer meaning of unknown words.	Apply explicit context clues (e.g., word and phrase) and/or text features to help understand meaning of unknown words.	Apply context clues (e.g., word, phrase, and sentence clues) and text features to help infer meaning of unknown words.	Apply implicit context clues (e.g., word, phrase, and sentence clues) and text features to infer meaning of unknown, complex words.
	LA 3.1.5.c	Acquire new academic and content-specific grade-level vocabulary, relate to prior knowledge, and apply in new situations.	Acquire grade-level vocabulary and relate to prior knowledge.	Acquire new academic and content-specific grade-level vocabulary, and relate to prior knowledge, and apply in new situations.	Acquire and use new academic and content-specific vocabulary, relate to prior knowledge, and apply accurately in new situations.

Source: <https://www.education.ne.gov/assessment/nscas-general-summative-assessment/nscas-english-language-arts-ela/>



The Nebraska standards are organized so that each expectation level represents a specific skill or building block for problem solving. This could be a learning progression, but these indicators are in separate expectation levels. Therefore, how each indicator may be expected to increase in sophistication needs to be defined to support defining the test score interpretations across achievement levels. Because the indicators are separate for these types of steps, the ALDs focus on other differentiating factors within each indicator to represent the progression of student knowledge and understanding of the specified skill. The ALDs also strive to preserve differentiation between the skills as they progress across grades. The following example shows where content limits, or conscious decisions about how content should increase in difficulty within an indicator, are used to differentiate items aligned with different achievement levels within an indicator, as well as across grades:

- Standard MA 3.1.1.b in Grade 3 Mathematics is about comparing whole numbers through the hundred thousands.
- The corresponding standard at Grade 2 compares two three-digit numbers.
- The lower level of Grade 3 continues the progression of the skill with comparing one three-digit number to a number between 1,000 and 100,000.
- The middle-level ALD then progresses to two numbers between 1,000, and 100,000.

The ALDs also differentiate between achievement levels through the presentation of information to the student or what supports are provided. In some cases, visual models are required at the lower level but not at the higher levels (provided the standard does not require visual models). The higher-level ALDs aim to require analysis of ELA and Mathematics to better assess conceptual understanding and higher levels of cognitive processing while also staying true to the indicator. The definition of content across achievement levels in this way is critical to supporting the development of content aligned to the state indicators and expectations at the levels of specificity denoted by state's test blueprints in terms of numbers of items per indicator. All items under this framework align to the indicators, and the explicit manipulation of item features to support changes in item difficulty is consistent with the Range ALD development framework in which content difficulty, cognitive processing demands, and contextual features such as scaffolding, visuals, and relationships with other standards are explicitly built into the ALDs (Egan et al., 2012). While this approach is helpful in a fixed-form context, it is critical to item development for an adaptive assessment.

### *2.6.3. Reporting ALDs*

Reporting ALDs are provided at the overall score level and are optimally created after final cut scores are adopted following the standard setting procedure. Reporting ALDs represent the reconciliation of the Range ALDs with the final cut scores. The Range ALDs reflect a state's initial expectation for student performance within an achievement level, whereas the Reporting ALDs reflect actual student performance based on the final approved cut scores. The Reporting ALDs define the appropriate inferences stakeholders may make based on the student's test score in relation to the final approved cut scores. Teachers are optimally given supportive information regarding how to interpret them to support formative practice.

## 2.7. ELA Passage Development

Table 2.5 presents the number of passages developed for the NSCAS Summative ELA assessments by passage type (literary vs. informational) and passage source (commissioned vs. public domain), including the development targets. As shown in the table, the targets were met with a total of 180 passages being developed, of which 64% were commissioned and 36% were from the public domain. All passages were reviewed during educator review meetings.

**Table 2.5. ELA Passage Targets and Development by Passage Type and Source**

Grade	Targets	#Passages				Total
		Passage Type		Passage Source		
		Literary	Informational	Commissioned	Public Domain	
3	30	10	20	24	6	30
4	30	12	18	21	9	30
5	30	12	18	21	9	30
6	30	15	15	21	9	30
7	30	10	20	15	15	30
8	30	14	16	12	18	30
Total	180	63	117	114	66	<b>180</b>

### 2.7.1. Passage Specifications

Passage specifications were developed prior to the start of passage development for ELA. Passages were not newly developed in any other content area. The document capture specifications such as what types of passages would be found or developed as well as the following passage considerations:

- Grade-level appropriateness
- Readability
- Word Count
- Accuracy of facts within the passage
- Bias, Sensitivity, and Fairness

### 2.7.2. Readability Measures

NWEA used both qualitative and quantitative measures during passage development. Qualitative aspects of a passage were critical when identifying reading material for the NSCAS ELA Assessments. Factors to consider included the following. The NWEA Text Complexity Qualitative Analysis Rubric was completed for each passage submitted for consideration.

- Text structure
- Levels of meaning
- Language features
- Demands on the reader
- Purpose
- Bias and sensitivity concerns
- ALD placement

The quantitative measures of a passage were also considered as a factor for all passages. Lexiles were used as the readability measure for this content development work. For pieces of text such as poems that perform poorly when Lexiles are run, Flesch-Kincaid was run as a secondary measure. Table 2.6 presents the acceptable Lexile ranges for each grade, as well as the total word count per passage. The passages selected for a grade spanned a range of acceptable readabilities. The word count must be reasonable for the task and, within the acceptable word count ranges, provide enough richness to support robust item sets.

**Table 2.6. Lexile and Word Count Ranges**

Grade	Lexile Range	Word Count
3	450L – 790L	200–700
4	745L – 980L	200–900
5	745L – 980L	300–1000
6	925L –1155L	400–1100
7	925L –1155L	400–1100
8	925L –1155L	400–1200

## 2.8. Item Development

Item development for 2018–2019 occurred for ELA, Mathematics, and Science. The adaptive and paper-pencil item pools are the same and therefore follow the same development processes. For ELA and Mathematics, an in-person IWW generated 60% of the development for this cycle. Independent contractors were then used to offset gaps in the item bank (i.e., about 40% of development) to ensure that enough items were developed to fulfill the item development requirements. Development of the new three-dimensional Science assessment aligned to the NCCRS-S began in July 2018 when a group of educators developed tasks and prompts for the March 2019 pilot test. For more information, refer to the Science pilot technical report (NWEA, 2019c). The tasks are currently in the final development stages. In the interim, existing items were used for the NSCAS Summative Science assessments in Grades 5 and 8.

### 2.8.1. Item Specifications

Each item on the NSCAS Summative assessments should align to one standard and should follow best practices for creating test items. The ALDs provide detailed information regarding each standard and how to assess student knowledge at different levels for each standard. Items should meet the level specified for each standard. Following the best practices, including style, helps ensure that items are accurately measuring student knowledge at each level by focusing the items on construct relevant information and presentation. The item specifications incorporates information from each source into a single file to provide a high-level overview for creating Nebraska summative test items.

There is a separate item specifications document for each content area. Item specifications for both ELA and Mathematics capture aspects such as the following and are reviewed at the start of each new development cycle to ensure accuracy. Mathematics also used draft ALDs to guide development. Item specifications for the new science assessment were based heavily on mathematics and are being updated collaboratively with NDE throughout the development process.

- General item writing guidelines in terms of overall content, item stems, item responses, style, and scoring rules
- Specific guidelines for using TEIs
- Specific standard information for Grades 3–8

### 2.8.2. Development Targets

Table 2.7 presents the ELA and Mathematics item development targets. TEIs are any item type that is not an MC item and can be worth 1 or 2 points. The item development plan included the development of 2,374 items across both content areas. Table 2.8 and Table 2.9 then present a breakdown of the targets for each content area. The ELA item bank had a notable shortage of writing items, most likely influenced by the Text Depend Analysis items that will not be used due to human handscoring requirements. Therefore, NWEA focused heavily on these items, which are not passage-dependent. After the Mathematics item bank realignment was complete, a review was done in 2018 prior to development. The item development plan is based on this review. Grades had different development targets based on the needs of each grade.

**Table 2.7. Overall Item Development Targets—ELA and Mathematics**

Grade	Overall Item Targets		
	MC	TEI	Total
<b>ELA</b>			
3	134	90	224
4	134	90	224
5	134	90	224
6	138	86	224
7	138	96	224
8	138	86	224
<b>Mathematics</b>			
3	59	88	147
4	76	107	183
5	64	94	158
6	76	115	191
7	63	102	165
8	72	114	186
Total	1,226	1,158	<b>2,374</b>

**Table 2.8. Item Development Targets—ELA**

Grade	Item Targets								
	Reading			Writing			Overall		
	MC	TEI	Total	MC	TEI	Total	MC	TEI	Total
3	102	66	168	32	24	56	134	90	224
4	102	66	168	32	24	56	134	90	224
5	102	66	168	32	24	56	134	90	224
6	106	62	168	32	24	56	138	86	224
7	106	62	168	32	24	56	138	86	224
8	106	62	168	32	24	56	138	86	224
Total	624	384	1,008	192	144	336	816	528	<b>1,344</b>

**Table 2.9. Item Development Targets—Mathematics**

Grade	Item Targets				Overall*
	MC	TEI		Total	
		1-pt.	2-pt.		
3	59	58	30	88	147
4	76	71	36	107	183
5	64	62	32	94	158
6	76	78	37	115	191
7	63	68	34	102	165
8	72	76	38	114	186
Total	410	413	207	620	<b>1,030</b>

\*The overall target is 1,020 with 10 additional items to allow for attrition with contract writers.

### 2.8.3. Item Writer Workshop (IWW)

The IWW from June 4–7, 2018, provided a professional development opportunity to educators and allowed them to be a part of the item development process. Table 2.10 presents the number of participants in each panel who were recruited and selected by NDE. The expertise of Nebraska teachers was critical to the item writing process. Nebraska educators wrote test items that were featured on the assessments. This ensured content that seems familiar to students as they take the tests; they will not see unfamiliar wording or approaches that might negatively impact performance.

**Table 2.10. IWW Panel Composition—ELA and Mathematics**

Panel	#Panelists
ELA 3–4	17
ELA 5–6	18
ELA 7–8	17
Math 3	9
Math 4	8
Math 5	8
Math 6	8
Math 7	9
Math 8	8
Total	85

During the IWW, educators were trained on how to write high-quality items aligned to the state standards for their content area. Participants met in smaller groups by grade level for training on the systems needed to enter items, as well as an orientation on their assignments. In this training, delivered collaboratively by NDE and NWEA, participants learned to write items that met the following criteria:

- Are properly aligned
- Ask clear and meaningful questions and use clear, concise wording
- Use technology as a logical enhancement to the item (rather than technology for technology's sake)
- Target content appropriate for the grade level (and ALD in math)
- Avoid stereotypes and topics that may cause discomfort to test takers
- Are accessible and adhere to universal design

A general session was held to train educators on the basics of item writing. A second, subject-specific training was completed with each group to dive into ELA and Mathematics issues. Once trained in both general and content-specific information, participants received training on the item management system. The participants then chose a standard and an item type to complete their assignment. This process was repeated until all required assignments were completed to meet the IWW targets. Throughout this process, educators partnered and shared their expertise as they wrote multiple-choice items and TEIs. NWEA and NDE staff circulated in break-out rooms to answer questions and provide guidance to participants. After the initial draft of an item was submitted, the participants and NWEA staff collaborated and engaged in brief group editing sessions that encouraged discussion and the continuing development of item-writing skills.

#### 2.8.4. Item Development Results

All newly developed items underwent a rigorous internal review. All items survived internal review of content and bias/fairness. The items were then reviewed by Nebraska educators during external item content and bias reviews. Table 2.11 and Table 2.12 present the number of newly developed items taken to the external content and bias reviews. Appendix B presents the number of items by standard taken to committee for both ELA and Mathematics. Table 2.13 then provides the difference between the item development targets and the actual number of items that were fully developed. The difference was added to the Summer 2018 item development targets.

**Table 2.11. Item Development Results—ELA**

Grade	#Items		
	MC	TEI	Total
3	151	132	283
4	134	114	248
5	128	128	256
6	129	99	228
7	109	97	206
8	152	126	278
Total	803	696	<b>1,499</b>

**Table 2.12. Item Development Results—Mathematics**

Grade	#Items				Overall*
	MC	TEI		Total	
		1-pt.	2-pt.		
3	59	58	29	87	146
4	76	70	35	105	181
5	64	61	32	93	157
6	75	77	36	112	188
7	63	68	34	102	165
8	72	76	38	114	186
Total	409	411	203	614	<b>1,023</b>

\*Over target by three due to not losing as many items from contractors as expected.

**Table 2.13. Item Development Targets vs. Number of Items Developed**

Grade	Target #Items	#Items Developed	Difference to be Added to the Summer 2018 Development
<b>ELA</b>			
3	224	283	0
4	224	248	0
5	224	256	0
6	224	228	0
7	224	206	0
8	224	278	0
<b>Mathematics</b>			
3	147	146	1
4	183	181	2
5	158	157	1
6	191	188	3
7	165	165	0
8	186	186	0

### 2.8.5. External Content and Bias Review

Nebraska educators gathered together from July 17–20, 2018, for two concurrent meetings: one to review items for content validity and one to review items for any possible sources of bias and sensitivity issues. While Nebraska educators served as the originators of a significant percentage of items, educator involvement in item reviews provided another opportunity to make sure that the material was appropriate and to provide a valuable professional development opportunity for participants. Participants received training, delivered collaboratively by NDE and NWEA, at the beginning of each review session and were provided checklists to refer to during the reviews.

Participants in item content review learned to review items for qualities such as the following:

- Proper alignment and cognitive complexity
- Clear and concise wording
- Presence of a correct answer

Participants in item bias review learned to review items for qualities such as the following:

- Diversity of background and cultural representation
- Avoidance of stereotypes
- Avoidance of topics that may cause discomfort to test takers
- Stimuli and item accessibility, and adherence to universal design

NWEA and NDE staff answered questions from participants during the workshop and helped to make sure that the review sessions remained productive and engaging for all attendees. Both groups reached consensus on each item and made one of the following decisions. Only items that were accepted during both reviews are eligible for field testing.

- Accept the item as is
- Accept the item with proposed modifications
- Reject the item

Table 2.14 presents the panel compositions for both the bias and content review meetings. Table 2.15 presents the number of items accepted, modified, or rejected results at the external content and bias review meeting. For ELA, 94.4% of items were either accepted or accepted with modifications, and 5.6% of items were rejected. For Mathematics, 99.6% of items were either accepted or accepted with modifications, and 0.4% of items were rejected.

**Table 2.14. Item Review Meeting Panel Composition**

Item Review Meeting	Panel	#Panelists
Bias Review	ELA 3–4	4
	ELA 5–6	5
	ELA 7–8	3
	Math 3–4	5
	Math 5–6	5
	Math 7–8	5
	<b>Total</b>	<b>24</b>
Content Review	ELA 3	5
	ELA 4	5
	ELA 5	4
	ELA 6	5
	ELA 7	5
	ELA 8	4
	Math 3	4
	Math 4	4
	Math 5	4
	Math 6	4
	Math 7	5
	Math 8	5
	<b>Total</b>	<b>50</b>
<b>Grand Total</b>		<b>74</b>

**Table 2.15. External Item Review Results**

Grade	#Items			Total
	Accepted	Modified	Rejected	
<b>ELA</b>				
3	152	111	20	283
4	196	37	15	248
5	208	45	3	256
6	168	56	4	228
7	76	115	15	206
8	163	88	27	278
Total	963	452	84	<b>1,499</b>



Grade	#Items			
	Accepted	Modified	Rejected	Total
<b>Mathematics</b>				
3	105	41	0	146
4	102	77	2	181
5	46	109	2	157
6	106	82	0	188
7	85	80	0	165
8	122	64	0	186
Total	566	453	4	<b>1,023</b>

### 2.8.6. Item Retirement

Newly developed items that do not survive the review process are not added to the item pool, and field tested items are removed from the pool if they do not pass data review. Operational items are removed (i.e., retired) based on content and psychometric reviews of items flagged based on their item statistics and a set of flagging criteria after each administration. Items with significant parameter changes based on the Robust Z statistic of +/-1.645 critical value are also removed. See Sections 6.2.3. and 6.5.5 for more information. There is no limit to how many times an item can be used operationally. Items may also be re-field tested if deemed necessary (e.g., if an item changed grades based on a new set of standards).

## 2.9. Content Alignment

To fully represent the constructs being assessed by NSCAS to determine if students are ready for college and careers, solid content alignment was critical. This was covered in several ways, including adherence to specifications, common interpretations of the standards, and an agreed-upon approach for cognitive complexity across all item types.

### 2.9.1. Alignment and Adaptive Testing

Within an adaptive testing context, the documentation of content blueprint features and percentages present in the item pool become one evaluation tool used to frame alignment discussions. Both item pool structure and constraints used to establish the administration of items during test events support the definition of the construct for alignment purposes. Full test blueprints must be supportable for students in each achievement level. Therefore, an ideal item pool has similar percentages of items within each indicator by achievement level cell.

As ALDs were developed based on theories of how student thinking grows within the state's structure of state standards, and the evidence needed to support that conclusion, the characteristics of items depend on the student's stage of reasoning. As ALDs describe increases in student thinking and reasoning, test developers have a rationale regarding why a percentage of particular item types (e.g., technology-enhanced items) and Depth of Knowledge (DOK) levels are necessary in the item bank, as well as the percentage of items that should be developed to particular levels of cognitive complexity within an item bank. Those decisions are driven based on the construct-based evidence that should be collected and included in item specifications. These decisions are made within each indicator by achievement level cell.

Students who are in earlier stages of reasoning can be forced into harder cognitive levels with harder content when computer adaptive constraints force all students to receive a certain percentage of items at a particular DOK level. A fundamental development practice for the Range ALDs (Egan et al., 2012) is that DOK levels follow the indicator progression. While DOK may increase across achievement levels, the DOK level should not automatically increase with the achievement level increase. What may be required from a learning theory perspective is that students have support accessing the standards, such as with visual supports demarcating a manipulation of an item context feature. They then may access the standards without the visual aids, followed by accessing the standards at a higher DOK level. Thus, if the item development is purposeful to the progression, DOK specifications are not required as a constraint conditional that items are measuring what the ALDs say they are.

When item development is purposeful to a clearly defined construct, dictating a certain percentage of items at a particular DOK level will unintentionally route a student to items that provide less information about their current stage of thinking and reasoning with the content. Thus, from a student and item bank evaluation perspective, alignment processes must consider the specific item demands of the ALDs within an achievement level and ask independent judges if items align to a specific ALD within an achievement level. This can be done during external content reviews with educators. Next, with the documented ALD matching of each item, the relationships among the achievement level categorizations, the item difficulty, and the degree of alignment can be used as evidence of alignment from a content validity perspective.

### *2.9.2. 2019 Mathematics Alignment Study*

NDE held an alignment study for the NSCAS Mathematics assessment from July 29 to August 8, 2019, based on Webb's DOK framework (1997, 1999, 2007) to examine the extent to which the NSCAS item pools represent Nebraska's College and Career Ready Standards for Mathematics and test interpretations as represented by the NSCAS Mathematics TOS. The workshop was conducted virtually. The results of the study contribute to the validity evidence to support the use of the NSCAS as a measure of the academic content standards. The study was a collaborative effort of NDE personnel, NWEA, EdMetric, and Nebraska educators. NWEA provided content via their Item Review Platform, Nebraska educators participated actively as panelists, and EdMetric facilitated and trained panelists in the process of examining test items and content to determine alignment ratings. The following questions guided this research:

- To what extent do the item pools represent the full range of the assessable Nebraska content standards?
- To what extent do the item pools measure student knowledge at the same level of complexity expected by the Nebraska content standards?

The results indicated that the NSCAS Mathematics assessment showed adequate alignment in terms of categorical concurrence, cognitive complexity (DOK), and both range and balance of knowledge. The degree of alignment varied across grade levels. The results further showed that further item development is needed for some reporting categories and additional DOK 3 items should be developed. Based on evidence from study results, the NSCAS item pools cover the full range of assessable Nebraska content standards, since the test events cover the full range of assessment standards and therefore the pools cover this range. The results of this study provide strong evidence that the item pools measure student knowledge at the same level of complexity expected by the NSCAS TOS for almost all grades for the NSCAS assessments. For full details and results of this alignment, please refer to alignment study report (EdMetric, 2019).

## 2.10. Universal Design

Ensuring that assessments are accessible to students with a variety of needs, including those with disabilities, is a critical part of item development. With a strong foundation in Universal Design for Learning (UDL), the assessments become engaging and accessible for all students. The NWEA content team ensures that each item is created with the principles of UDL in mind. These principles provide a framework for developing flexible items to support many kinds of learners and maximize options for assessments provide multiple means of representation, action and expression, and engagement. Applying UDL principles to assessments helps to reduce barriers and minimize irrelevant information from the items, so the assessment can show what each student knows.

## 2.11. Sensitivity and Fairness

NWEA takes seriously the task of creating items that are free from bias and sensitivity issues and is fair to all students, as defined below. Items are revised to eliminate bias, sensitivity, and fairness issues—or rejected when an issue cannot be remedied through the revision process.

- **Bias:** Item content, unrelated to the concept or skill being assessed, that may unfairly influence a student’s performance, or an item construct that does not have equivalent meaning for all students.
- **Sensitivity:** The experience of taking a test differs from the classroom experience in that students do not have the opportunity to discuss the material with a teacher or their peers. Sensitive content risks drawing students out of the testing experience by provoking negative emotional responses.
- **Fairness:** Equitable treatment of all students during the assessment process. To make a test fair, test developers must work to eliminate any barriers that prevent students from understanding and interacting with item content in a manner that accurately demonstrates what they know or are able to do.

A successful item is free of bias and sensitivity issues and is accessible to all students. An item should NOT:

- Distract, upset, or confuse in any way
- Contain inappropriate or offensive topics
- Require construct-irrelevant knowledge or specialized knowledge
- Favor students from certain language communities
- Favor students from certain cultural backgrounds
- Favor students based on gender
- Favor students based on social economic issues
- Employ idiomatic or regional phrases and expressions
- Stereotype certain groups of people or behaviors
- Favor students from certain geographic regions
- Favor students who have no visual impairments
- Use height, weight, test scores, or homework scores as content or data in an item

There is not a hard and fast “list” of material that is potentially distracting or upsetting, but some topics are seldom appropriate for K–12 assessments, such as sexuality, illegal substances, illegal activities, excessive violence, discriminatory descriptions, death, grieving, catastrophes, animal neglect or abuse, and loss of a family member.

## 2.12. Test Construction

The adaptive tests were produced by selecting the item pools, building the test models that configured the engine and provided the constraints, running simulations, approving the results, and conducting user acceptance testing (UAT). The fixed forms were created based on the TOS and fixed-form construction specifications that included the following statistical guidelines:

- Absolute test characteristic curve (TCC) difference  $<.05$
- A max of three items with differential item functioning (DIF) flag of C- or C+
- A max of three items with item-total correlation flag
- A max of three items with omit rate  $> 5\%$
- A max of three items with item-total correlation for a distractor  $> 0.05$
- A max of three items with  $p$ -value  $< 0.2$  or  $> 0.9$
- A max of three items with  $p$ -value for answer key is  $<$  distractor  $p$ -value
- No items with answer key item-total correlation  $<$  item-total correlation for a distractor
- No items with negative item-total correlation

The content team selected the items based on the TOS and specifications for each grade and content area, including the following. Item selection was an iterative process between the psychometrics and content teams before being sent to NDE for review and approval.

- Number of items per standard indicator
- Number of items at each level of cognitive complexity
- The balance between dichotomous and polytomous items
- The balance between multiple-choice and technology-enhanced items

## 2.13. Data Review

Data review is the process of reviewing field tested items for quality and appropriateness based on the results of statistical analysis of student responses. The review of content alignment and statistics of the Spring 2019 NSCAS Summative ELA and Mathematics field tested items occurred virtually in September 2019 between NDE and NWEA. Table 2.16 and Table 2.17 present the data review flagging criteria for multiple-choice and non-multiple-choice items, respectively. Items were flagged based on these criteria and brought to the data review meeting,<sup>4</sup> although items with a negative item-total correlation or polytomous items without a second step parameter were marked Do Not Use (DNU) and not included. Participants were provided a spreadsheet with the statistics for each item, as well as a data review “cheat sheet” provided in Appendix C. Only flagged items were brought to the data review meeting.

**Table 2.16. Data Review Flagging Criteria—Multiple-Choice Items**

Statistic	Criterion	Indication
DIF of gender or ethnicity	C+ or C-	potential bias toward a certain group of students
item fit statistics	$< 0.7$ or $> 1.3$	poor fit
$p$ -value	$< 0.20$ or $> 0.9$	very difficult item
item-total correlation	$< 0.20$	poorly discriminating item
item-total correlation for distractors	$> 0.05$	poorly discriminating item
omit rate	$> 5\%$	unclear or very difficult item

<sup>4</sup> The details of field testing item analyses are included in Section 5 of this technical report.

**Table 2.17. Data Review Flagging Criteria—Non-Multiple-Choice Items**

Statistic	Criterion	Indication
DIF of gender or ethnicity	C+ or C-	potential bias toward a certain group of students
item fit statistics	< 0.7 or > 1.3	poor fit
step parameters	Step 1 > Step 2	not a good separation of students into different stages of learning
Item-total correlation	< 0.1	poorly discriminating item
Item-total correlation for score of 0	> 0.0	poorly discriminating item
item-total correlation for score of 1 < item-total correlation for score of 0	–	poorly discriminating item
item-total correlation for score of 2	< 0.1	poorly discriminating item
item-total correlation for score of 2 < item-total correlation for score of 1	–	poorly discriminating item
low student count for each score	=0	no one got a certain score (e.g., no student got a score of 2)

Table 2.18 presents the data review results, including the number of field test items included in the pool, the number of field test items administered during the 2019 testing window, the number of field test items not accepted or labeled as DNU, and the number of accepted field test items.

**Table 2.18. Data Review Results**

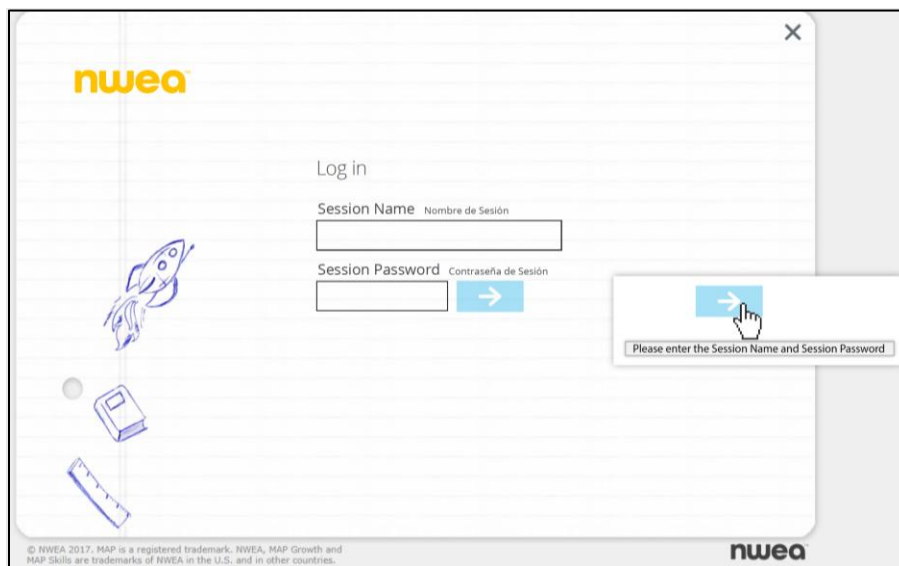
Grade	#FT Items in the Pool	#FT Items Administered in 2019	#Rejected/DNU Items	#Accepted Items
<b>ELA</b>				
3	194	180	25	155
4	191	191	44	147
5	185	185	20	165
6	193	192	28	164
7	195	195	53	142
8	189	184	20	164
<b>Mathematics</b>				
3	231	231	3	228
4	231	231	1	230
5	231	231	0	231
6	231	231	1	230
7	231	231	2	229
8	231	231	4	227

### Section 3: Test Administration and Security

The Spring 2019 NSCAS Summative testing window was from March 18 to April 26, 2019, and the make-up testing window was from April 29 to May 3, 2019. Due to natural disasters in the state, the testing window was extended through May 17 for a select number of districts. The ELA and Mathematics tests each had 48 items and were adaptive. The Science operational tests each had 50–60 items and were fixed form. All tests were untimed. Testing sessions were structured as a single session, although students could complete the tests in more than one sitting by pausing the test. Students were not able to go back to previous items.

The tests were administered online via the NWEA Comprehensive Assessment Platform (CAP) test management system, a roles-based platform that allowed users to roster students, set up test sessions, and administer the assessment. Figure 3.1 presents the student CAP login screen. CAP works with the NWEA secure lockdown testing browser to administer the assessments, which is required for summative testing. Paper-pencil versions were also available as an accommodation. Each district was required to return either a paper-pencil answer sheet or an online record for all Grades 3–8 students enrolled in the district.

**Figure 3.1. CAP Student Login Screen**



The NSCAS Summative administration supported student testing on Windows® PC, Macintosh®, iPads, and Chromebooks that met the following specifications. Touch screens were not supported, and Chromebook tablets were only supported if the student was using an external keyboard. iPad mini® devices were not recommended.

- Windows 7, 8.1, or 10
- Mac OS X® v10.12 to 10.15
- iOS 11 to 12 and iPadOS 13.1.2 or higher recommended
- Google Chrome™ OS 65 or higher

### 3.1. User Roles and Responsibilities

Table 3.1 summarizes the user roles and responsibilities for the NSCAS test administration.

**Table 3.1. User Roles and Responsibilities**

User	Roles and Responsibilities
District Assessment Contacts	Responsible for coordinating the testing activities of all schools within their districts. Responsibilities included but were not limited to coordinating the test schedules of the schools within the district and setting up test sessions.
School Assessment Coordinators	Served as single points of contact at the schools for the District Assessment Contacts and were responsible for coordinating the testing activities within their schools. Responsibilities included but were not limited to secure handling of test materials such as test tickets and coordination of proctors. A School Assessment Coordinator and District Assessment Contact might be the same person depending on the district's decisions.
Proctors	Responsible for administering the tests to students.

District Assessment Contacts were responsible for scheduling the test for all schools within the district and coordinating the distribution and collection of test materials, as well as any specific training that the District felt was needed. It was recommended that District Assessment Contacts conduct an orientation session for School Assessment Coordinators to review and/or discuss:

- District test schedule
- General information in the Test Administration Manual (TAM)
- Procedures for distribution and collection of test materials
- Procedures for maintaining security, outlined in the TAM and the NSCAS Security Manual
- Proctor orientation

School Assessment Coordinators were responsible for providing secure test materials to proctors and conducting proctor orientations, reviewing topics such as:

- Test schedule
- Administration preparation
- Students with special needs
- Testing conditions
- Security

### 3.2. Administration Training

In addition to district- and school-held trainings, NWEA, in collaboration with NDE, held two trainings for district leaders in advance of testing. The Fall 2018 regional workshops were half-day, in-person workshops held across multiple regions of the state from October 8–12, 2018. Information on the spring summative administration including test sessions, accessibility, and student rostering was presented. The summative test administration workshops from February 18–22, 2019, were two-hour virtual sessions that provided important information on the NSCAS assessments. Table 3.2 presents the locations and number of participants based on the registration numbers for the Fall 2018 regional workshop, and Table 3.3 presents the dates and number of participants based on the registration numbers for the summative test administration workshop. Appendix D presents the PowerPoint training presentations for each training.

**Table 3.2. Fall 2018 Regional Workshop Locations and Participation**

Location	#Participants
Scottsbluff – Gering Civic Center	26
Kearney – Younes Hospitality	66
West Point – Nielsen Community Center	32
Lincoln – The Cornhusker Marriott	57
Omaha – DC Centre	30

**Table 3.3. Summative Test Administration Workshop Dates and Participation**

Date	#Participants
February 18, 2019	42
February 19, 2019	46
February 20, 2019	20
February 21, 2019	27
February 22, 2019	27

### 3.3. Item Type Samplers

Item Type Samplers were available online and in PDF paper-pencil formats for all content areas and grades and were available on the NSCAS Assessment Portal at <https://community.nwea.org/community/nebraska/practice-tests>. The username and password for the item samplers were available in the Item Type Sampler manual (username = ne, password = sampler). Large print and Braille versions were also created and available for order when requested through the Educational Data Systems (EDS) ordering system for paper materials.

The Item Type Samplers were not adaptive and had the same 20 items for each respective grade in a content area. They were also untimed, although the estimated test-taking time for each was 40 minutes. Unlike the actual summative assessments, progress on the item sampler was not saved. If a student did not complete the test in one sitting, they had to take the entire test again if they restarted it. A score was not generated at the end of the test, but keys were made available.

The Item Type Sampler Manual was provided on the NSCAS Assessment Portal with information on the item sampler, how to access it, and recommended proctor scripts. The purpose of the item samplers was to allow students to experience the types of items, tools (e.g., calculator), and item aids (e.g., highlighter) available on the actual summative assessments. They also allowed other stakeholders such as parents and administrators to experience the summative assessment environment. For the best student experience, it was recommended that students view the Online Student Tutorial located on the NSCAS Assessment Portal to learn about the available tools and their uses before taking the item samplers. Text-to-speech was available for all practice tests, but it was recommended that it only be enabled for students with a documented need on an Individualized Education Plan (IEP) or 504 Plan to be consistent with the requirements for use on the NSCAS Summative assessment.



### 3.4. Accommodations and Accessibility Features

Table 3.4 presents the accessibility supports available for the Spring 2019 NSCAS test administration, including the embedded and non-embedded accommodations and universal features. More information and guidance about these supports can be found in the NSCAS Summative & Alternate Accessibility Manual (NDE, 2018).

- Accommodations are changes in procedures or materials that ensure equitable access to instructional and assessment content and generate valid assessment results for students who need them. Embedded accommodations (e.g., text-to-speech) are provided digitally through instructional or assessment technology, while non-embedded accommodations (e.g., computation supports) are provided locally. Accommodations are available for students for whom there is a documented need on an IEP or 504 Plan.
- Universal features are accessibility supports that are embedded and provided digitally through instructional or assessment technology (e.g., answer choice eliminator), or nonembedded and provided non-digitally at the local level (e.g., scratch paper). Universal features are available to all students as they access instructional or assessment content.

Supports such as linguistic supports and aids for English language learners (ELLs) were also available to students, either universally or according to need (i.e., IEP or 504 Plan). A complete list of linguistic supports is included in the NSCAS Summative & Alternate Accessibility Manual.

**Table 3.4. Accommodations and Universal Features**

Support	Description
<b>Embedded Accommodations</b>	
Text-to-speech	A student can use this feature to hear audio of the item content.
<b>Non-Embedded Accommodations</b>	
Paper-pencil	A student takes the assessment on paper instead of online.
Computation supports	For students who need additional supports for math computations (e.g. abacus, calculation device, number line, addition/multiplication charts, etc.)
Assistive technology	Includes such supports as typing on customized keyboards, assistance with using a mouse, mouth or head stick or other pointing devices, sticky keys, touch screen, and trackball, speech-to-text conversion, or voice recognition
Audio amplification device	Hearing impaired student uses an amplification device (e.g., FM system, audio trainer)
Braille	A raised-dot code that individuals read with the fingertips. Graphic material is presented in a raised format.
Braille writer or notetaker	A blind student uses a braille writer or note-taker with the grammar checker, internet, and file-storing functions turned off.
Flexible scheduling	The number of items per session can be flexibly defined based on the student's need.
Large print test booklet	A large print form of the test provided to the student with a visual impairment. A student may respond directly into test booklet. Test administrator transfers answers onto answer document.
Project online test	An online test is projected onto a large screen or wall. Student must use alternate supervised location that does not allow others to view test content.

<b>Support</b>	<b>Description</b>
Primary mode of communication	Student uses communication device, pointing or other mode of communication to communicate answers.
Read aloud	Only for students who have a documented need for paper-pencil. The student will have those parts of the test that have audio support in the computer-based version read by a qualified human reader in English.
Response assistance	Student responds directly into test booklet. Test administrator transfers answers onto answer sheet.
Scribe	The student dictates their responses to an experienced educator who records verbatim what the student dictates.
Sign interpretation	An educational sign language interpreter signs the test directions, content and test items to the student. ELA passages may not be signed. The student may also dictate responses by signing.
Specialized presentation of test	Examples include colored paper, tactile graphics, color overlay, magnification device, and color of background.
Voice feedback	Student uses an acoustical voice feedback device (e.g., WhisperPhone)
<b>Embedded Universal Features</b>	
Answer choice eliminator	Used to cross out answer choices that do not appear to be correct.
Flexible scheduling	Districts and schools have flexibility to schedule each content test. Each test is only a single session and can be scheduled for one or multiple days.
Highlighter	Used for marking desired text, items, or response options with a color.
Keyboard navigation	The student can navigate throughout test content by using a keyboard (e.g., arrow keys). This feature may differ depending on the testing platform or device.
Line reader/line guide	Used as a guide when reading text.
Math tools	These digital tools (e.g., ruler, protractor, calculator) are used for tasks related to math items. They are available only with the specific items for which one or more of these tools would be appropriate.
Notepad	Used as virtual scratch paper to make notes or record responses.
Zoom (item-level)	The student can enlarge the size of text and graphics on a given screen. This feature allows students to view material in magnified form on an as-needed basis. The student may enlarge test content at least fourfold. The system allows magnifying features to work in conjunction with other accessibility features and accommodations provided.
<b>Non-Embedded Universal Features</b>	
Alternate location	Student takes test at home or in a care facility (e.g., hospital) with direct supervision. For facilities without internet, a paper-pencil test will be allowed.
Directions	Test administrator rereads, simplifies or clarifies directions aloud for student as needed.
Color contrast	Background color can be adjusted based on student's need.

Support	Description
Cultural considerations	The student receives a paper-pencil form due to specific belief or practice that objects to the use of technology. This student does not use technology for any instructional related activities. Districts must contact NDE to request this accessibility feature.
Noise buffer/headphones	The student uses noise buffers to minimize distraction or filter external noise during testing.
Redirection	Test administrator directs/redirects student focus on test as needed.
Scratch paper	The student uses blank scratch paper, blank graph paper, or an individual erasable whiteboard to make notes or record responses.
Setting	The student is provided a distraction-free space or alternate, supervised location (e.g., study carrel, front of classroom, alternate room).
Student reads test aloud	The student quietly reads the test content aloud to self. This feature must be administered in a setting that is not distracting to other students.

### 3.5. User Acceptance Testing (UAT)

User acceptance testing (UAT) is conducted each year to test the most common configurations in use in Nebraska on each device based on the following criteria:

- Content
- Item type functionality (e.g., make sure only correct answer can be selected for a multiple-choice item)
- Universal features/item aids and tools (e.g., highlighter, eraser, answer eliminator)
- Item-specific features (e.g., ruler, protractor)
- Accessibility features (e.g., TTS)
- New features/enhancements

Over the course of three days, 39 testers participated in UAT in 2019. Each were assigned 2–4 tests. Testers are typically NWEA staff who are at least somewhat familiar with how functionality is supposed to interact. In addition to a training and kick-off on the process and a checklist of tasks, technical product managers are present at the kick-off meeting to describe the UAT process overall, expected enhancements to functionality, and known issues. Use cases describing each item feature and other support documentation are provided to testers to review prior to UAT. Testers should spend 1–2 hours reviewing existing documentation prior to performing testing. They are also encouraged to explore the item type sampler beforehand.

To conduct UAT, testers are assigned tests on a particular device and location (e.g., work desk, at home) and spend approximately 30–40 minutes per test. Bugs are reported and tracked manually. Daily triage meetings take place with key NWEA staff to review all new entries that have been reported and to update the status for known issues. During the UAT process, testers review live, secure NSCAS tests. Test security is taken very seriously, and testers are not allowed to share, copy, record, or take photos of the items they review. This is considered a serious breach in test security.

### 3.6. Student Participation

All students with disabilities were expected to participate in the NSCAS. No student, including students with disabilities, could be excluded from the state assessment and accountability system. All students were required to have access to grade-level content, instruction, and assessment. Students with disabilities may have been included in state assessment and accountability in the following ways:

- Students were tested on the NSCAS Summative assessments without accommodations.
- Students were tested on the NSCAS Summative assessments with approved accommodations specified in the student's IEP. Accommodations provided to students must have been specified in the student's IEP and used during instruction throughout the year. Accommodations may have required paper-pencil testing.
- Students could be tested with the NSCAS Alternate assessment if they qualified for these assessments. Only students with the most significant cognitive disabilities (typically less than 1% of students) could take these tests. The NSCAS Alternate test was distributed and administered by DRC.

Use of non-approved accommodations may have invalidated the student's score. Non-approved accommodations used in state testing resulted in both a zero score and no participation credit. Accommodations provided adjustments and adaptations to the testing process that do not change the expectation, grade level, construct, or content being measured. Accommodations should have only been used if they are appropriate for the student and used during instruction throughout the year. In contrast, modifications are adjustments or changes in the test that affect test expectations, grade level, construct, or content being measured. Modifications were not acceptable in the NSCAS assessments.

#### 3.6.1. Paper-Pencil Participation Criteria

Students participating in the paper-pencil administration had to meet one of the following criteria:

- Student has medical condition that does not allow the use of computer screens
- Student requires Braille/Large Print
- Facility does not allow internet access
- Student requires written translations of languages other than Spanish
- Cultural considerations
- Student needs test in both English and another language side-by-side (Mathematics and Science only)
- Student is an English Learner with limited prior access to technology

#### 3.6.2. Participation of English Language Learners (ELLs)

According to the Elementary and Secondary Education Act (ESEA), ELLs are students who have a native language other than English, OR who came from an environment where a language other than English has had a significant impact on their level of English proficiency, AND whose difficulties in speaking, reading, writing, or understanding the English language may be sufficient to deny the individual (i) the ability to meet the state's proficient level of achievement on state assessments, (ii) the ability to successfully achieve in classrooms where the language of instruction is English, or (iii) the opportunity to participate fully in society. (For full text of the definition, please see Public Law 107-110, Title IX, Part A, Sec. 9101, (25) of the No Child Left Behind Act of 2001.)

Each district with ELL students should have a written operational definition used for determining services and meeting Office of Civil Rights requirements. Both state and federal laws require the inclusion of all students in the state testing process. ELL students must be tested on the NSCAS Summative. Districts should have reviewed the following guidelines before summative testing:

- In determining appropriate linguistic supports for students in the NSCAS system, districts should use the NSCAS Summative & Alternate Accessibility Manual (NDE, 2018).
- Districts must be aware of the difference between linguistic supports (accommodations for ELLs) and modifications.
- For students learning the English language, linguistic supports are changes to testing procedures, testing materials, or the testing situation that allow the students meaningful participation in the assessment. Effective linguistic supports for ELL students address their unique linguistic and socio-cultural needs. Linguistic supports for ELL students may be determined appropriate without prior use during instruction throughout the year.
- Modifications are adjustments or changes in the test or testing process that change the test expectation, grade level, construct, or content being measured. Modifications are not acceptable in the NSCAS assessments.

### *3.6.3. Participation of Recently Arrived Limited English Proficient (RAEL) Students*

Recently Arrived Limited English Proficient (RAEL) students are defined by the U.S. Department of Education as students with limited English proficiency who attended schools in the United States for fewer than 12 months. The phrase “schools in the United States” includes only schools in the 50 states and the District of Columbia. It does NOT include Puerto Rico. Districts must assess all RAEL students on all NSCAS assessments each year based on the grade level of the student using linguistic supports.

### **3.7. Test Security**

In a centralized testing process, it is critical that equity of opportunity, standardization of procedures, and fairness to students is maintained. Therefore, NDE asked that all school districts review the NSCAS Security Procedures provided in the TAM. Breaches in security are taken very seriously, and it was emphasized that they must be quickly identified and reported to NDE’s Statewide Assessment Office. Districts were encouraged to maintain a set of policies that includes a reference to Nebraska’s NSCAS Security Manual. A sample district testing and security policy was included in Nebraska’s Standards, Assessment, and Accountability Updates posted on NDE’s website. Whether districts use this sample, the procedures offered by the State School Boards Association, or policies drafted by other law firms, local district policy should address the NSCAS Security Manual. NDE encouraged all districts with questions to contact their own local school attorney for customization of such a policy.

As part of NDE’s security policy, the principal of each school participating in the NSCAS Summative assessments were required to complete and sign a Building Principal Security Agreement and return it to the Statewide Assessment Office by October 12, 2018. District Assessment Contacts were required to complete and sign the District Assessment Contact Confidentiality of Information Agreement and return it to the Statewide Assessment Office by October 12, 2018. School districts were bound to hold all certificated staff members in school districts accountable for following the Regulations and Standards for Professional Practice Criteria as outlined in Rule 27. The NSCAS Security Manual was intended to outline clear practices for appropriate security.

### *3.7.1. Online Test Security*

Students had access to the testing environment through the NWEA lockdown browser, a secure testing browser that disabled access to all external programs to allow secure testing. There was a series of authentication steps to allow students to access the test through the lockdown browser, including individual student test tickets with log-in information, and proctor permission being granted by the proctor.

Student test tickets, generated after test session creation by a School Assessment Coordinator or District Administrator Contact, contained student-level password information for accessing the tests and were kept secure. Proctors were given the student test tickets prior to test administration, allowing them ample time to review and organize the tickets for distribution before the test begins. Once a test session was started, only the student taking the test as allowed to view the student's screen. No one could view or copy test content while a student was testing. Test tickets were to be securely destroyed following the end of testing.

During testing there was a 15-minute inactivity setting that engaged after 15 minutes of no activity in the testing environment. Activity consisted of any mouse movement. This was a measure to maintain the test security should the student step away from the desk.

### *3.7.2. Paper-Pencil Test Security*

For paper-pencil testing, districts were responsible for the secure handling of all physical test materials, including test booklets and answer sheets. Districts were instructed that, in between test sessions, all materials must be kept in a predetermined, locked, secure storage area, and never be left unattended. At the end of the test window, all materials were to be returned to NWEA for scoring and/or destruction. Any materials not returned to NWEA due to concerns such as biohazard contamination or fire were to be securely destroyed. All items securely destroyed at the district level were to be recorded on the Local Destruction Reporting Form sent to districts in their packet of test materials. This form was to be returned with the remainder of the materials. Instructions and guidance on test security best practices were included in the Test Administration Manual for Paper and Pencil and districts were expected to adhere to them.

#### 3.7.2.1. Physical Warehouse Security

All EDS personnel—including subcontractors, vendors, and temporary workers who have access to secure test materials—were required to agree to keep the test materials secure and sign security forms that state the understanding of the secure nature of test items and the confidentiality of student information. Access to the document-processing warehouse was by rolling gates, which were always locked except when opened to allow pickup or receipt of test materials. A secure chain-link fence with a barbwire top surrounds the document-processing facility. A verified electronic security system monitored access to the offices and warehouse areas 24 hours a day, seven days a week. All visitors entering the facility were required to sign in at the front desk and obtain an entry badge that allowed them access to the facility. The following additional security procedures were maintained for the NSCAS Summative program:

- Test materials received from the printing subcontractors were stored in a secure warehouse facility prior to packaging and shipping to districts.
- All boxes and pallets placed in the secure warehouse for long-term storage were recorded electronically so that they could be retrieved at any time. Scanned (used) answer documents were stored in labeled “scan” boxes on labeled pallets in the same

warehouse. The scan box and pallet numbers were scanned into a database for retrieval as needed. Documents are stored until the second week of January following the test administration or until NDE provides express written consent to destroy them.

#### 3.7.2.2. Secure Destruction of Test Materials

EDS will manage the secure destruction of test materials during the first two weeks of January 2020. Using the information from the long-term storage database, EDS will retrieve the documents and systematically destroy them through a secure shredding process. The shredding company uses a high-capacity mobile onsite document destruction vehicle that provides the most advanced document destruction technology in the industry. The shred trucks, equipped with a 20-inch monitors so EDS staff may monitor the documents going into and being expelled in a pulverized state, provide the quickest, most complete, and most confidential destruction of sensitive documents. Every sensitive document is pulverized using a *hammermill* process that creates the smallest pieces in the document destruction industry.

After the test materials destruction process is complete, the shredding company provides a certificate of destruction that will remain on file at EDS. The long-term storage database will be updated to reflect that the materials have been destroyed. During the first two weeks of January 2020 and upon written approval, EDS will also delete the answer document images from the server hard drive and all backup drives. The deletion process will securely erase the data to ensure that the images cannot be retrieved through data restorative means. EDS will provide NDE with archives of all data files prior to deletion, upon request.

#### 3.7.2.3. Shipping Security

Hardcopies of the prepress test materials for proof approval were provided to NWEA via traceable courier and tracked to ensure arrival. All proofs arrived with no incident. For district shipments, EDS used the secure and trackable UPS ground and two-day shipping services to send materials to and receive materials from districts. The system interfaced with the in-house UPS shipping system, thus making certain that deliveries were made to accurate and correct addresses. Address verification was used to ensure that the materials were shipped to known UPS addresses before shipping. To ensure correct deliveries to all sites, all boxes belonging to a school or district were numbered and labeled with unique barcode numbers tracked in the system. Every box was assigned a unique UPS tracking number and the numbers were uploaded to the Materials Tracking module allowing EDS, districts, NWEA, and NDE to track all shipments and diagnose problems early. One-hundred percent of shipments containing test documents were tracked and monitored to and from sites. EDS resolved all shipping issues in a timely manner and no material reships were required.

#### 3.7.2.4. Electronic Security of Test Materials and Data

All computer systems that store test materials, test results, and other secure files required password access. During the test material printing processes, electronic files were transferred via a server accessed by Secure File Transfer Protocol (SFTP). Access to the site was password controlled and on an as-needed basis. Transmission to and from the site was via an encrypted protocol. Transfer of student data between NWEA and EDS followed secure procedures. Data files were exchanged through an SFTP site and the secure application program interface (API). During use, the data files resided on secure EDS servers with controlled access.

### 3.7.3. Caveon Test Security

#### 3.7.3.1. Monitoring for Disclosure of Test Content

Caveon Web Patrol investigated NSCAS Summative assessments online with the primary goals of detecting, reporting, and eliminating, where possible, exposures and infringing content from the individual assessments. During the administration windows, Caveon Core was used as a secure incident reporting and encrypted materials storage platform for NWEA or NDE. Live test items provided to Caveon Web Patrol by NWEA were protected by placing them securely on a non-networked air-gapped computer. Access to those live items was only authorized to be used by Caveon's Executive Web Patrol Manager. Live items were never used for searching but only for verification in the case of potential infringements. Use of materials, other than live test items, were also limited to only Caveon Web Patrol employees assigned to this project. Each employee signed non-disclosure agreements before engaging in work for NWEA and NDE and was trained in how to protect their security online using anonymous email addresses, Virtual Private Networks, and prescribed processes for accessing, transferring, and handling of secure client files and associated information. Once infringing content was found and verified, it was reported to NDE through the notification tools built into Caveon Core. A secondary notification by email message was sent from the Web Patrol Director of Operations or Executive Web Patrol Manager as a means of redundancy to ensure that NWEA and NDE were made aware of the potential infringements in a timely manner.

#### 3.7.3.2. Monitoring for Potential Test Security Violations

Caveon data forensics analyses were performed to discover anomalous results that may be indicative of potential test security violations. These analyses provided information regarding where and when test security incidents may have occurred, by whom, and their effects on the testing program. Table 3.5 summarizes the statistical analyses performed. The data forensics analyses were conducted to identify potential test security violations relating to individual students, schools, and items on the exams.

**Table 3.5. Statistical Analysis and Potential Incidents**

<b>Statistical Analysis</b>	<b>Potential Incident</b>
Response Times	Responding to items inconsistently regarding time or supplying answers in unusually short lengths of time can indicate pre-knowledge of test content or unsanctioned aid given to students while taking the test (i.e., test coaching).
Person-fit (Aberrance) Statistics	When students respond in a manner that is inconsistent with the student population supportive evidence of pre-knowledge or test coaching may be present.
Scored Differences	Non-scored items are typically being field tested and are usually newly created. Large performance differences between the operational and non-scored items suggest that students may have access to the test content prior to the exam.
Item Performance Changes	Performance shifts, indicating the items have become easier during the test administration window, provide evidence that the item might have been disclosed to the students.
Exposed Differences	Item exposure (i.e., administrations to individual students) vary in CAT pools. When student performance is higher on frequently exposed items than on the other items, there is a possibility that some or a few students had access to some of the test content prior to the exam.



Statistical Analysis	Potential Incident
Linking Difference	Linking items are administered to nearly all students. Due to their greater exposure, these items have a greater risk of being compromised. This statistic compares each student's performance on linking items against their performance on the non-linking items to determine if any students potentially had pre-knowledge of the linking items.
Volatile Scores	Tests are detected by this statistic when the test score is inconsistently high or low in comparison with prior test scores. Unusual gains or losses can be indicative of pre-knowledge of test content and/or coaching during the prior or current year.
M4 Similarity	Exams that use fixed forms (e.g., Science) were analyzed for excessive agreement between pairs of students. These statistics can identify where answer copying by students, sharing of test responses between students, or large-scale collusion may have occurred.
Identical Test	When students receive the same items (i.e., because they were administered the same form), it is possible they may have identical responses to the items. This is more likely when they use the same disclosed answer key. When this happens, students will often have very high scores on the exam.
Perfect Test	A concentration of perfect scores at a school, which are very unusual, may indicate the presence of a test security incident.
Synchronicity	When students answered questions at or near the same time of day, there is a possibility that they were guided or paced through the exam. This analysis detects potential incidents when this occurred.

As provided in the data forensics report from Caveon (Mulkey, Maynes, & Scott, 2019), data for 326,553 test instances administered at 828 schools in 246 districts were analyzed. The most significant findings are as follows:

- Twenty-four test instances were flagged for extreme similarity.<sup>5</sup> These 24 test instances formed 12 pairs of extremely similar tests. The observed similarity is extremely improbable under the assumption of independent test taking.
- High detection rates by the Volatile Scores Statistic, accompanied by increased performance for detected test instances, may be evidence of a security violation for a few schools.
- High detection rates by the Synchronicity Statistic,<sup>6</sup> accompanied by increased performance for detected test instances, may be evidence of a security violation for a few schools.
- At the school level, one school-subject-grade group had a high detection rate by the Linking Difference Statistic. For this group, the Linking Difference anomalies were not associated with improved performance.

Overall, the assessments appeared to have been administered securely.

<sup>5</sup> Similarity analysis was conducted for only Science 5 and 8 because those were the only assessments with fixed forms; all other NSCAS forms were administered adaptively.

<sup>6</sup> Synchronicity analysis was conducted for only Science 5 and 8 because those were the only assessments with fixed forms; all other NSCAS forms were administered adaptively.

### 3.8. Partner Support

NWEA’s Partner Support Services team provided implementation and technical support throughout the 2018–2019 school year for the NSCAS Summative assessments. This team provided resources to support Nebraska and its educators, assisting with generating roster files, configuration of the assessment program, accessing online reports, and general questions with the use of the online assessment system. NWEA provided phone, email, and chat support to schools and educators from 8:00 a.m. to 5:00 p.m. Central Time (CT) Monday through Friday, and 7:00 a.m. to 5:00 p.m. CT during the testing windows, as described in Table 3.6. Table 3.7 presents the number of cases presented to the Partner Support team by case type for the entire 2018–2019 school year from August 2018 to June 2019 for the NSCAS tests. More than half of the cases were related to testing (i.e., administration questions).

**Table 3.6. Partner Support Communication Options**

<b>Phone Support</b>	NWEA used Voice Over Internet Protocol (VOIP) phone systems to allow callers to quickly reach the first available representative. VOIP also provided remote access capabilities for our staff, enabling Partner Support team members to provide seamless service even during times of inclement weather or office closure. Reports from our phone system and customer relationship management tool, as well as call monitoring tools, were used in monitoring quality and in the determination of additional training needs.
<b>Email Support</b>	Emailed support requests are also handled quickly and efficiently. It was our goal to respond to all emails within twenty-four hours from time of receipt. Emails received within NWEA business hours are responded to on the same business day.
<b>Chat Support</b>	Chat is a convenient method of contacting support for in-the-moment questions or for use in the rare occurrence of a phone service disruption.

**Table 3.7. Number of NSCAS Cases to Partner Support in 2018–2019**

Case Type	#Cases	% of Total Cases
Student Mobility	2	0.2
Reports	162	14.0
Navigation	69	5.9
Setup and Management	186	16.0
Testing	741	63.9
<b>Total</b>	<b>1,160</b>	<b>100.0</b>

NWEA monitored all service activities through daily, weekly, and monthly reports and made adjustments as needed to ensure appropriate coverage for Nebraska support needs during peak use times, such as prior to and throughout the testing windows. All Tier 1 and Tier 2 support staff members were required at hire to undergo a two-week training program led by the NWEA Senior Support Specialist team and team trainers. The training program consisted of a combination of instructor-led and self-paced eLearning courses, covering all relevant team policies and procedures, including security requirements of handling student data, product expertise, and troubleshooting requirements. In addition, several days of “phone shadowing” were built into the program to ensure that each new staff member had the opportunity to participate in calls with veteran staff monitoring prior to working independently. Senior Support Specialists were responsible for continually updating training program content to ensure that all support team staff members were knowledgeable of current policies. In addition, the project managers and product training resources were dedicated to NDE’s program to train the support staff on Nebraska-specific policies. This equated to roughly 90 hours of training (80 for initial training and 10 for state-specific training and the testing platform).

## Section 4: Scoring and Reporting

The online ELA and Mathematics assessments were administered adaptively via NWEA's constraint-based engine, whereas the Science assessments, all paper-pencil tests, and all Spanish versions were administered as fixed-form. Specifically, the ELA and Mathematics tests were minimally adaptive because the item pool and test design did not allow for item-by-item adaptive decisions to be made for every student. Therefore, students saw the same items until a time when the item bank could support a more individualized administration. In the fixed-form test, every student received the same items. All tests were scored with maximum likelihood estimation (MLE) scoring.

### 4.1. Scoring Rules

An attemptedness rule is the minimum number of items a student must attempt during testing to be included in psychometric analyses and/or receive a numeric score. Table 4.1 presents the attemptedness rules for scoring.

**Table 4.1. Attemptedness Rules for Scoring**

#OP Items Attempted	Include in Psychometric Analyses?	Receive Scale Score?	Receive Achievement Level?
0	No	Yes, LOSS	Yes, lowest level
1–9	No	Yes, LOSS +1	Yes, lowest level
10+	Yes	Yes, calculated MLE scores	Yes

The attemptedness rule was decided based on the results of the standard error of measurement (SEM) that became relatively stable after 10 operational items from the simulation data and the finding of a small number of 2017 students who attempted less than 10 items. Regarding scoring, NWEA ran analyses using a subpopulation of the 2017 students and found that the number of not-reached items increased the amount of estimation error, suggesting larger estimation error with the penalty function (i.e., to score those not-reached items as wrong). However, scoring consistency were also considered for fixed forms (e.g. Science). Thus, NDE made the following scoring rules in consultation with the state and district coordinators:

1. Students who took the adaptive assessment (i.e., ELA and Mathematics online adaptive forms) received straight MLE scoring (i.e., regular MLE scoring with no penalty) regardless of the test completion status. Students who took the Spanish online assessment also received straight MLE scoring.
2. Except for the Spanish online form, MLE scoring with penalty was applied to fixed forms (i.e., Science online and paper-pencil, Spanish paper-pencil, and ELA and Mathematics paper-pencil), treating omit and multi-marks as incorrect.
3. Sub-scores were provided for students who attempt a minimum of 10 items overall and four items within each specific reporting category.

### 4.2. Paper-Pencil Scoring

#### 4.2.1. Scanning of Answer Sheets by EDS

EDS scanned and imaged all paper-pencil student responses and captured student demographic information provided on the answer sheets. Answer sheets were scanned using high-speed optical mark reading (OMR) NCS 5000i scanners. The scanning, editing, and scoring processes

were performed after most answer sheets were returned by districts. Answer sheets were scanned and edited in accordance with the NSCAS data processing specifications created by EDS and NWEA. The editing processes included steps to check the spelling of the student name (i.e., that the scanner picked up all the bubbled letters and that there were no multiple marks, no embedded blanks, and no initial blanks in the name) and that the scanner picked up all the bubbled digits in the NSCAS Identifier. Since some answer sheets contained preprinted precode information from the roster file, the student demographics provided via the roster file were merged into the scan file so that all demographics and scan marks were included in one file. This merge was completed on the precode barcode ID number printed on the answer sheet. Checks were performed to eliminate duplicate barcode numbers during each step of the merging process. Finally, EDS created a Scan Export File for each grade to merge the scan data with the NWEA response “choice” conversion data. The resulting data (student demographics, scan marks, and choice conversation values) was transferred in JSON format to NWEA via an API.

#### 4.2.1.1. Quality Control of Scanning and Scoring

Before scanning began, a complete deck of controlled data, the “test deck,” was created and scanned. The test deck documents were created by bubbling the answer sheets based on the test deck control file, which contained various combinations of demographic information and answer responses for all grades and all content areas. To test that the scanners and programs were functioning correctly, the test deck scan file was compared to the test deck control file to ensure that the outputs matched.

Next, a complete check of the scanning system was performed. Intensity levels of all scanners were constantly monitored by running diagnostic sheets through each scanner before and during the scanning of each batch of answer sheets. Scanners were recalibrated if discrepancies were found. Documents received in poor condition (e.g., torn, folded, or stained) that could not be fed through the scanners were transferred to a new scannable document to ensure proper scoring of student responses. Editing and resolution procedures were followed to resolve demographic information issues on the answer sheets (e.g., multiple marks, poor erasures, or incomplete data). Multiple iterations of error listings were prepared to verify the correction of all errors and to correct any errors introduced during the editing process.

Scanner operators performed ongoing maintenance checks designed to ensure that the scanners read reliably. After two hours of scanning, operators cleaned and dusted all open areas with continuous-stream compressed air and performed a quick check. If the quick check failed, the read heads were calibrated. Additionally, calibration occurred at a minimum of every four hours of scanning, and an Image Calibration Log was completed and checked by the lead operator. A software utility program notified the scanner operator of a buildup of dust, erasure fragments, or other irregularities that affect the quality of the images. This utility notified the scanner operators of an issue in time to prevent data errors. A user exit program checked whether the scanner read heads were registering values in coordinates that should be blank and alerted the operator that the read heads needed cleaning. In addition, cleaning of the rollers, read-head de-skew tests, and barcode-reader tests were performed periodically.

A final check was made of the actual counts of student answer sheets scanned compared to the expected counts from the Group Identification Sheet (GIS) and School Group List (SGL). All discrepancies for both scannable and non-scannable and/or missing test materials were investigated and resolved.

#### 4.2.1.2. Quality Control of Image Editing

The test deck was used to test all possible errors in the edit specifications. This set of test documents was used to verify that all images from the answer sheets were saved correctly for the NSCAS program (e.g., images of the barcode and student name sections of the answer sheet), including the following checks:

- Verifying that the image-editing program correctly indexes scanned images to the correct student and that fields needing editing are completely captured as an image
- Verifying that the number of images in a given scan file (for the grades in the file) is accurate prior to loading the file into the image-editing program for scoring

#### 4.2.1.3. Quality Control of Answer Document Processing and Scoring

Before the processing and scoring system was used operationally, a complete test deck of controlled data was run through the scanning, routing, and merging programs, resulting in the production of complete student records and reports. The following quality checks were made immediately after scanning:

- The scanning process is checked to ensure that the scanner was properly calibrated.
- Data that can be captured from answer sheets but was not bubbled properly into the scannable grids are edited and verified.
- The number of scanned student records, the quantity bubbled on the scanned GIS, and the quantity written on the SGL are compared to ascertain that all documents assigned to a scan file are contained in the scan file.
- The system is programmed to confirm that students are correctly coded as belonging to a valid school, district, and grade. Changes are made as necessary.
- All invalid or out-of-range lithocodes are reviewed and resolved.

If editors found discrepancies between scan counts and counts from the GIS and SGL, they investigated by counting the physical documents in the scan boxes. They also reviewed the GIS, SGL, and documents in the previous and subsequent group to be sure documents were not scanned out of order. All discrepant counts were verified and reconciled before the scan file was cleared for subsequent processing. Finally, steps were in place to process the scan and choice conversion processes on two different software platforms (parallel processing). The data was provided to NWEA only when the outputs from both processes matched.

#### 4.2.2. *Scoring by NWEA*

The paper-pencil scanned documents were converted to JSON and ingested by an NWEA API. The data was then matched to existing student records and new test events were created. The test events were designated with a non-tested code (NTC) of PPA. The test events went through the constraint-based engine and were scored based on the test models developed for PPA. The records were treated as the online records and went through the normal scoring process.

### 4.3. Score Reporting Methods

Student performance on the NSCAS Summative assessment was reported as a scale score and achievement level. Scale scores ranged from 2220 to 2890 for ELA, 1000 to 1550 for Mathematics, and 0 to 200 for Science, as shown in Table 4.2. In isolation, scale scores are difficult to interpret. In the interpretation of test results, it is not appropriate to compare scale scores across content areas. Each content area is scaled separately. Therefore, the scale scores for one content area cannot be compared to another content area.

**Table 4.2. Scale Score Ranges**

<b>Grade</b>	<b>Scale Score Ranges</b>		
<b>ELA</b>	<b>Developing</b>	<b>On Track</b>	<b>CCR Benchmark</b>
3	2220–2476	2477–2556	2557–2840
4	2250–2499	2500–2581	2582–2850
5	2280–2530	2531–2598	2599–2860
6	2290–2542	2543–2602	2603–2870
7	2300–2555	2556–2629	2630–2880
8	2310–2560	2561–2631	2632–2890
<b>Mathematics</b>	<b>Developing</b>	<b>On Track</b>	<b>CCR Benchmark</b>
3	1000–1189	1190–1285	1286–1470
4	1010–1221	1222–1316	1317–1500
5	1020–1235	1236–1330	1331–1510
6	1030–1243	1244–1341	1342–1530
7	1040–1246	1247–1345	1346–1540
8	1050–1263	1264–1364	1365–1550
<b>Science</b>	<b>Below the Standards</b>	<b>Meets the Standards</b>	<b>Exceeds the Standards</b>
5	0–84	85–134	135–200
8	0–84	85–134	135–200

An achievement level is a written description of the student’s overall performance and is used to help make the scale scores meaningful. There are three other important reasons for establishing achievement levels:

- Give meaning to the scale scores to help Nebraska students and parents use the results effectively
- Connect the scale scores on the tests to the ELA, Mathematics, and Science standards to assist Nebraska educators in supporting students to become college and career ready
- Meet the requirements of the U.S. Department of Education

The Nebraska State Board of Education defined three achievement levels for each content area, as shown in Table 4.3.

**Table 4.3. Achievement Level Descriptions**

Achievement Level	Description
<b>ELA &amp; Mathematics</b>	
Developing	Developing learners do not yet demonstrate proficiency in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards. These results provide evidence that the student may need additional support for academic success at the next grade level.
On Track	On Track learners demonstrate proficiency in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards. These results provide evidence that the student will likely be ready for academic success at the next grade level.
CCR Benchmark	CCR Benchmark learners demonstrate advanced proficiency in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards. These results provide evidence that the student will likely be ready for academic success at the next grade level.
<b>Science</b>	
Below the Standards	Overall student performance in science reflects <i>unsatisfactory</i> performance on the standards and <i>insufficient</i> understanding of the content at grade level. A student scoring at the Below the Standards level <i>inconsistently</i> draws on a broad range of scientific knowledge and skills in the areas of inquiry, physical, life, and Earth/space sciences.
Meets the Standards	Overall student performance in science reflects <i>satisfactory</i> performance on the standards and <i>sufficient</i> understanding of the content at grade level. A student scoring at the Meets the Standards level <i>generally</i> draws on a broad range of scientific knowledge and skills in the areas of inquiry, physical, life, and Earth/space sciences.
Exceeds the Standards	Overall student performance in science reflects <i>high</i> academic performance on the standards and a <i>thorough</i> understanding of the content at grade level. A student scoring at the Exceeds the Standards level <i>consistently</i> draws on a broad range of scientific knowledge and skills in the areas of inquiry, physical, life, and Earth/space sciences.

The reporting categories in Table 4.4 were used for scoring and reporting. Items were mapped to a reporting category based on the indicators.

**Table 4.4. Reporting Categories**

Content Area	Reporting Categories
ELA	<ul style="list-style-type: none"> <li>• Reading Vocabulary</li> <li>• Reading Comprehension</li> <li>• Writing Skills</li> </ul>
Mathematics	<ul style="list-style-type: none"> <li>• Number</li> <li>• Algebra</li> <li>• Geometry</li> <li>• Data</li> </ul>
Science	<ul style="list-style-type: none"> <li>• Inquiry, Nature of Science &amp; Tech</li> <li>• Physical Science</li> <li>• Life Science</li> <li>• Earth/Space Sciences</li> </ul>

#### 4.4. Report Summary

The following reports were produced for the 2019 NSCAS Summative test administration and made available in September and October 2019. Appendix E presents examples of each report. A Reports Interpretive Guide was developed to help district leaders understand, explain, and use the NSCAS results and was made available on the NSCAS Assessment Portal. A separate Individual Student Report (ISR) Reports Interpretive Guide was also created for parents in both English and Spanish. All reports were delivered online in CAP according to user role. Printed ISRs were also delivered to districts.

- Student-Level Reports
  - Individual Student Report (ISR)
  - Individual Student Report (ISR) with Non-Tested Code (NTC)
- School-Level Reports
  - School Roster
  - School Achievement Level Summary
- District-Level Reports
  - District Achievement Level Summary
- State-Level Reports
  - State Achievement Level Summary

ISRs showed a student’s performance on the NSCAS Summative tests. Reports were posted in PDF format online within CAP. School districts and state administrators could download them in English or Spanish from the CAP View Reports page. Two copies of each ISR in English were also printed, sorted by school, and delivered to each district. One copy was sent home with the student, and the second copy was to be filed in the student’s cumulative folder. During July and August 2019, districts had the option to choose to use their fall MAP® Growth™ roster file to indicate where ISRs for students should be shipped.

If a non-tested code (NTC) was applied to any content area, the student’s achievement level scores and proficiency by reporting category within the respective content area were reported as affected by the NTC, as defined in Table 4.5. If a student had an NTC of INV, PAR, SAE, or UTT assigned to their test, the automatically assigned score displayed with a score of one less than the lowest scale score for that grade and content.

**Table 4.5. Non-Tested Codes (NTCs)**

NTC	Achievement Level Received		Description
	ELA & Mathematics	Science	
<b>ALT</b> Alternate assessment	N/A	N/A	Student participated in the alternate assessment.
<b>EMW</b> Emergency Medical Waiver	No level	No level	Student was not tested because of an approved emergency medical waiver.



NTC	Achievement Level Received		Description
	ELA & Mathematics	Science	
<b>INV</b> Score invalidated by the state	Developing	Below the Standards	Student's test was invalidated; student received a score of one less than the lowest scale score for that grade and content area and an achievement level of Developing/Below the Standards.
<b>NLE</b> No longer enrolled	No level	No level	Student was not enrolled in the district at the time of testing.
<b>OTH</b> Other	No level	No level	Left blank. These tests are excluded on the School Roster, and the corresponding content area is omitted from the ISR.
<b>PAR</b> Parental refusal	Developing	Below the Standards	Student was not tested because of a written request from parent or guardian; student received a score of one less than the lowest scale score for that grade and content area and an achievement level of Developing/Below the Standards
<b>SAE</b> Student absent for entire test window	Developing	Below the Standards	Student was absent during the entire test window; student received a score of one less than the lowest scale score for that grade and content area and an achievement level of Developing/Below the Standards.
<b>UTT</b> District was unable to test student	Developing	Below the Standards	District unable to test student; student received a score of one less than the lowest scale score for that grade and content area and an achievement level of Developing/Below the Standards.

The School Roster report listed students who were required to take the NSCAS Summative tests and presented a report of their performance. The size of this document depended on the class size. The School Achievement Level Summary report presented a summary of performance and demographics for all students at a school by grade who were required to take the NSCAS Summative tests. The District Achievement Level Summary report was for internal district use only and was required for state and federal reporting purposes. It was available for district-level personnel to access through the Reports landing page within CAP. Information to protect small numbers of students was not suppressed. The State Achievement Level Summary report presented the average state performance based on demographics for the NSCAS Summative tests. It was available for state-level personnel to access through the Reports landing page within CAP.

## 4.5. Reporting Process

### 4.5.1. Online Reports

To access the online reports, users generated reports in the reports landing page based on their role, as shown in Figure 4.1. Users selected the report type (e.g., ISR, school performance summary, etc.) and criteria (e.g., district, school, and grade) before hitting the “Download Report” button. The user’s role interacted properly constrained users in the reports landing page to only access reports they were authorized to see. For example, teacher-level users would only be able to access student reports for students in their own classes. The reporting page was also protected by the same security measures that applied to every aspect of CAP.

**Figure 4.1. Reports Landing Page Example—District Assessment Contact**

The screenshot displays a web interface for selecting report criteria. At the top, it says "Select your report criteria:" with a "Clear Selections" link and a "GENERATE REPORT" button. Below this, there are two main sections: "Select your report type (select 1):" and "Select your report criteria (select 1 per dropdown):".

**Select your report type (select 1):**

- Student-Level Reports**
  - Individual Student Reports
- School-Level Reports**
  - School Performance Level Summary
  - Student Roster File

**Select your report criteria (select 1 per dropdown):**

- \*District:** A dropdown menu with "Search Districts" and a list containing "BELLEVUE PUBLIC SCHOOLS".
- \*School:** A dropdown menu with "Search Schools" and a list containing "AVERY ELEMENTARY SCHOOL", "BELLEAIRE ELEMENTARY SCHOOL", "BELLEVUE ELEMENTARY SCHOOL", and "BELLEVUE MISSION MIDDLE SCHOOL".
- \*Grade:** A dropdown menu with "Search Grade" and a list containing "All Grades", "3", "4", "5", and "6".

### 4.5.2. Printed ISRs

ISRs were the only reports that were printed and shipped. Education Strategy Consulting (ESC) developed the ISRs based on the NSCAS Reports Specifications and mockups. The reports were printed in greyscale, and the data was transferred vis SFTP. EDS compiled the individual ISRs into school-level packages with appropriate district, school, and grade headers. School-specific barcodes were added to the header sheets for packing quality control procedures. ISRs were printed using EDS’s in-house high-speed laser printers on plain white paper with blue headers. Once each school’s reports were printed, the package was shrink-wrapped and placed on the pick and pack line for packing into boxes.

The data transfer and compiler programs were tested on several sample reports provided by ESC. A final review was completed prior to live report printing with actual districts and students with report exceptions. During the pick and pack of reports, a barcode on the school package header sheet was scanned into the packing database, which was prepopulated with the expected barcodes per district. The report shipment could not be closed and shipped until the correct schools and quantity of packages were scanned into the database.

#### 4.5.3. Report Verification

The NSCAS report quality assurance (QA) process consisted of validating the data and reports using the scoring and reporting specifications, mockups, layouts, and scale score and cut information. The first step was to validate that the data were accurate and the appropriate rules were applied. PDF reports were then generated and validated. Specific schools were identified to validate the scoring and reporting rules. After the reports passed quality control, they were loaded to a staging environment to verify the report landing page and user access. Printed reports were also spot checked onsite at EDS prior to packaging of the ISRs. The quality control reviews completed by EDS included ensuring print quality, that all districts, schools, and students requiring reports received a printed report, that they were collated and shrink-wrapped per the reporting specifications, that the header pages were accurate and collated correctly, and that all ancillaries such as the packing list and cover letter were complete and accurate.

The objectives of report verification were to ensure that:

- The reports match NDE's expectations.
- The data on the report are accurate.
- The data on the report are presented per NDE's expectations.
- NDE and users can access the reports.

The following report sections were checked during the QA process:

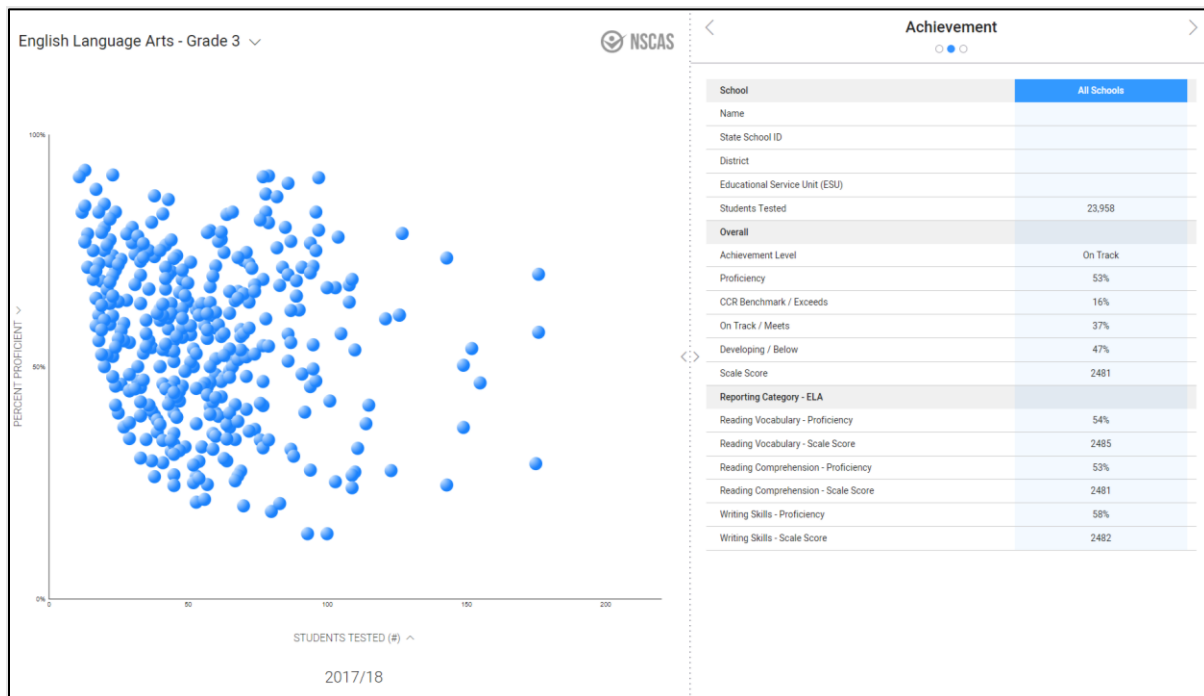
- Formatting
- Static text (text that does not change)
- Dynamic text (text that changes)
- Student data (demographic information)
- Score-related data (scale scores, average scale scores, achievement levels)
- Graphs the scored data
- Footnotes
- NTC behavior
- Not enough items behavior
- Accurate number of reports generated
- Sorting (sort order of the report)
- Naming conventions reports, files, and folders

#### 4.6. Matrix

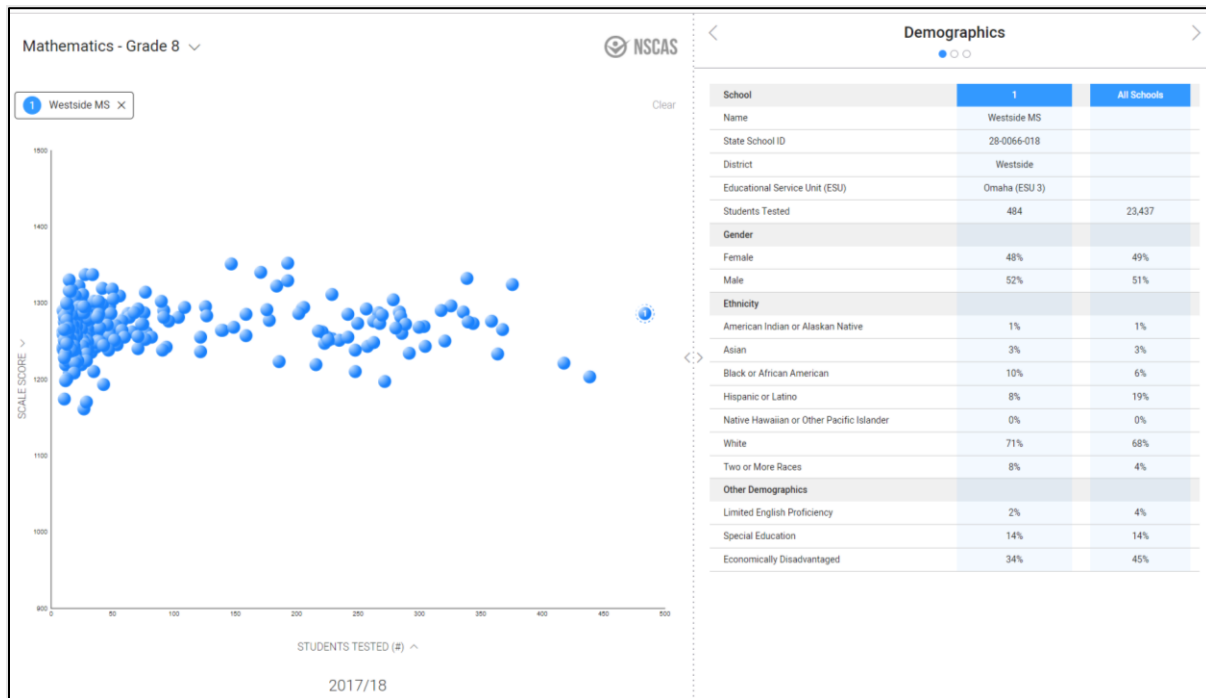
NWEA used ESC's tools to view web-based visualizations for the NSCAS assessments, including combinations of aggregate and disaggregate information of results by demographics and other filtering options. This web portal, referred to as the Matrix, allowed users to save and print specific plot and screen images from the interactive visualization. Users could interact with and explore many different levels of information to answer targeted questions about their district, school, or state. The main feature of this tool was an interactive scatterplot designed to display longitudinal data, as shown in Figure 4.2, Figure 4.3, and Figure 4.4. The X and Y axes were modifiable. Users could construct a spreadsheet from all the available variables within the visualization via the export function. This feature allowed for easy access to high-quality data that had gone through rigorous auditing. Users could then explore and sort data to meet their individual needs. Suppression rules were applied to the data for all users. For example, all data was suppressed for a school if the number of tested students was less than 10.

Districts and educational service units (ESUs) were provided direct access to the Matrix, and role-based filter conditions of the Matrix were available for state personnel and researchers who had a deep familiarity with the data. District Administrator Contacts and School Assessment Coordinators also had access. All user roles except ESUs accessed the Matrix through a hyperlink on the Reports Landing page in CAP. ESU representatives were given direct links to access the Matrix. The Matrix is password protected, and all users saw the same info and could download all data because suppression had been applied. ESC developed videos on the navigation aspects of the Matrix to help users learn how to best use the tool. In collaboration with NDE, ESC also developed professional development videos to help users understand how to interpret and apply the data.

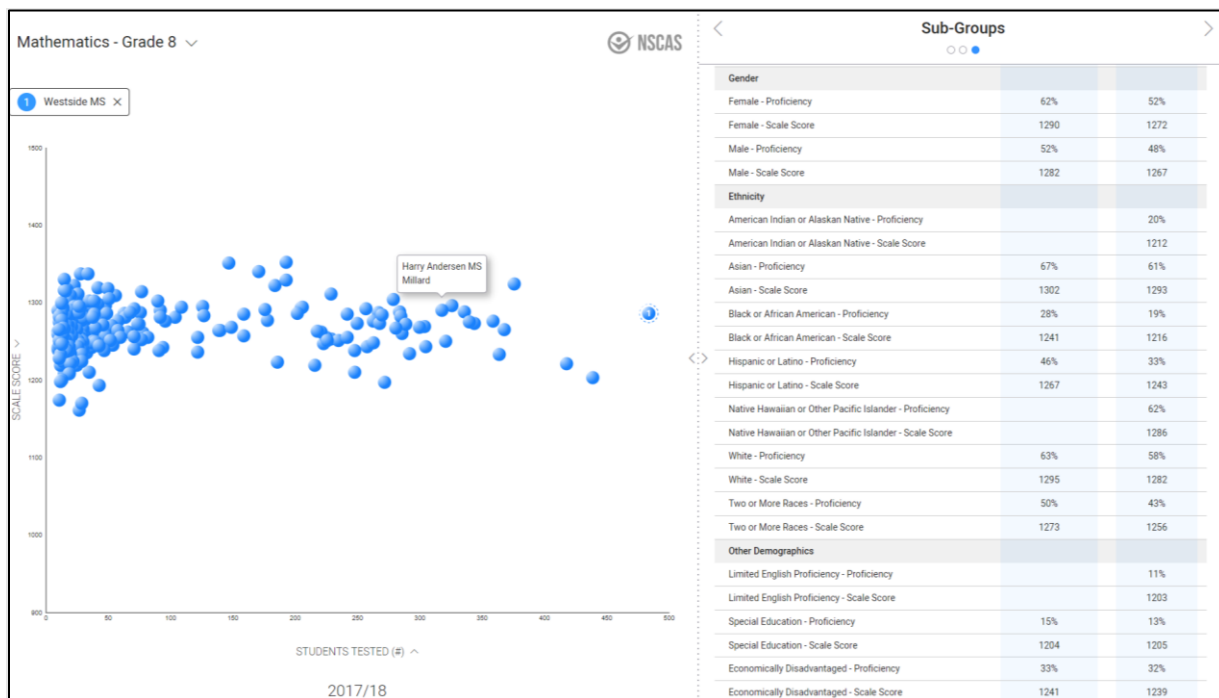
**Figure 4.2. Matrix Example: Percent Proficient**



**Figure 4.3. Matrix Example: Scale Score by Demographics**



**Figure 4.4. Matrix Example: Scale Score by Sub-Groups**

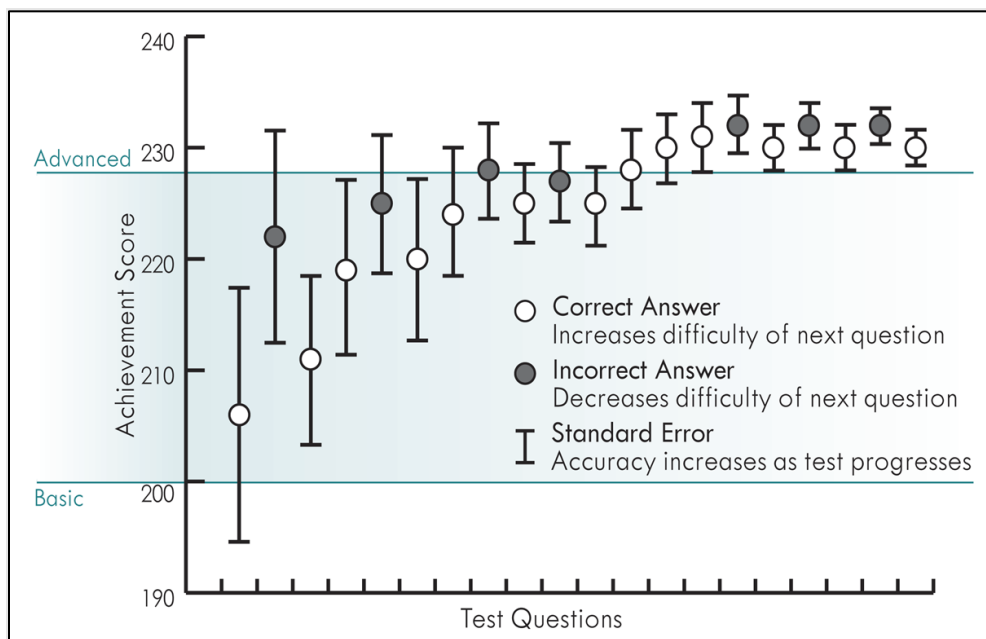


## Section 5: Constraint-Based Engine

### 5.1. Overview

An adaptive assessment administers items to match the ability level of the student. Students receive different items based on item difficulty and their ability levels. For example, students with lower ability levels (based on their answers to previous items) receive easier items compared to students with higher ability levels who receive harder items as the test progresses. A constraint is a rule given to the engine when selecting items. For example, the engine must meet the TOS when considering the next item. The adaptive engine uses the TOS and a student's momentary theta ( $\theta$ ) to drive item selection, as shown in Figure 5.1. Momentary theta is the ability estimate of the student that is recalculated and updated after answering each item.

**Figure 5.1. Adaptive Engine Overview**



Items were selected based on item difficulty. The goal of the constraint-based engine's item selection was to provide a test that meets "must-have" (hard) constraints and "nice-to-have" (soft) constraints. Examples of hard constraints are all item selection constraints, such as all levels of standards, field test items, and operational items. Examples of soft constraints are student population exposure goals and population exposure limits by anchor items.

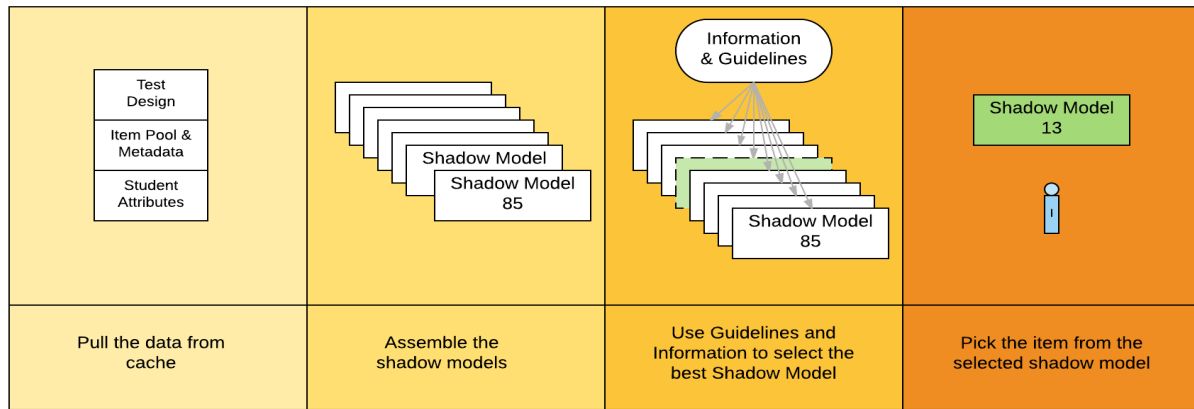
The adaptive engine has two stages of consideration as it selects the next item that conforms to the TOS while providing the maximum information about the student based on the student's momentary ability estimate:

1. Shadow test approach
2. A variation of the weighted penalty model

As shown in Figure 5.2, the shadow test approach (Van der Linden & Reese, 1998) selects items based on the required aspects of the TOS, and a new valid shadow model is selected upon each update to the student's momentary theta. In other words, this approach uses the student's answer to the last item to create shadow models that are waiting "in the shadows"

while the student answers the current item. When the student responds to the item, that answer is used to select the next correct shadow model. Because multiple shadow models can be drawn from an item pool, a variation of the weighted penalty model (Segall & Davey, 1995) then selects which shadow model is optimal based on additional content guidelines while ensuring the most representative sample for linking and field test items. The shadow model with the smallest penalty is selected when multiple shadow tests meet the required attributes of the test and have similar information.

**Figure 5.2. Shadow Test Approach**



## 5.2. Engine Simulations and Evaluation

Pre-administration engine simulations and a post-administration engine evaluation studies are important evidence, along with post-administration analyses, for confirming interpretation and test score use arguments regarding student proficiency with the state standards. Pre-administration simulations were conducted prior to the Spring 2019 operational testing window to evaluate the constraint-based engine’s item selection algorithm and estimation of student ability based on the TOS. The simulation tool used the operational constraint-based engine, thereby providing results with the same properties and functionality as what would be seen operationally. Detailed information regarding the simulation study can be found in the full report (NWEA, 2019a).

After the Spring 2019 testing window closed, a post-administration evaluation study was then conducted to determine whether the constraint-based engine performed as expected. The results included a blueprint constraint accuracy analysis to determine whether the constraint-based engine administered the assessments based on the TOS; item exposure rates to determine the number of items administered to students; score precision and reliability; and population explore for linking and field test items. Detailed information regarding all results of the post-administration evaluation study can be found in the full report (NWEA, 2019b).

Overall, the constraint-based engine performed as it should based on the blueprint (i.e., TOS) constraints. The reporting category points had a 100% match. The points at the indicator level are also matched to the blueprints. The constraint-based engine also showed a similar performance when estimating the students’ ability in terms of SEM and reliability. Item exposure rates were also acceptable given that the constraint-based engine used almost half of the items to administer the test and most used items had a 0–20% exposure rate.

### 5.2.1. Evaluation Criteria

Computational details of the precision ability estimation statistics (i.e., bias,  $p$ -value, and MSE) are as follows (CRESST, 2015):

$$bias = N^{-1} \sum_{i=1}^N (\theta_i - \hat{\theta}_i) \quad (5.1)$$

$$MSE = N^{-1} \sum_{i=1}^N (\theta_i - \hat{\theta}_i)^2 \quad (5.2)$$

where  $\theta_i$  is the true score, and  $\hat{\theta}_i$  is the estimated (observed) score. To calculate the variance of theta bias, the first-order Taylor series of the above equation is used as follows:

$$var(bias) = \sigma^2 * g'(\hat{\theta}_i)^2 = \frac{1}{N(N-1)} \sum_{i=1}^N (\theta_i - \hat{\theta}_i)^2 \quad (5.3)$$

where  $\hat{\theta}_i$  is an average of the estimated theta. Significance of the bias is then tested as follows:

$$Z = bias / \sqrt{var(bias)} \quad (5.4)$$

A  $p$ -value for the significance of the bias is reported from this z-test with a two-tailed test. The average standard error (SE) is computed as follows:

$$Mean(se) = \sqrt{N^{-1} \sum_{i=1}^N se(\hat{\theta}_i)^2} \quad (5.5)$$

where  $se(\hat{\theta}_i)^2$  is the standard error of the estimated  $\theta$  for individual  $i$ . To determine the number of students falling outside the 95% and 99% confidence interval coverage, a  $t$ -test was performed as follows:

$$t = \frac{\theta_i - \hat{\theta}_i}{se(\hat{\theta}_i)} \quad (5.6)$$

where  $\hat{\theta}_i$  is the ability estimate for individual  $i$ , and  $\theta_i$  is the true score for individual  $i$ . The percentage of students' estimated theta falling outside the coverage was determined by comparing the absolute value of the  $t$ -statistic to a critical value of 1.96 for 95% coverage and to 2.58 for the 99% coverage.

Traditional reliability coefficients from classical test theory consider individual items and depend on all test takers to take common items, whereas students receive different items in a CAT. Therefore, NWEA calculated the marginal reliability coefficient for the CAT administration. Samejima (1994) recommended the marginal reliability coefficient because it uses test information (e.g., variance of estimated theta and SEM) to estimate the reliability of student scores:

$$\text{Marginal Reliability} = \frac{var(\hat{\theta}) - \sigma^2}{var(\hat{\theta})} \quad (5.7)$$

where  $\sigma$  is defined as:

$$\sigma = E\{[I(\theta)]^{-1/2}\} \quad (5.8)$$



### 5.2.2. Blueprint Constraint Accuracy

Table 5.1 and Table 5.2 present the blueprint constraint results at the reporting category level for the pre-administration simulation study and the post-administration evaluation, respectively. The findings from the engine evaluation study appeared similar to those in the simulation study, as expected. For both studies and content areas, the number of items and points at the reporting category level resulted in a 100% match for all grades based on the blueprint. Results were also provided at the indicator level by passage type selection, DOK level, and item range requirements (NWEA, 2019a, 2019b). While most DOK levels also resulted in a 100% match, some indicators did not because the constraint-based engine used DOK level as a guideline or a “nice to have” given the limited number of items at a specified DOK level for some indicators. Passage type for ELA also resulted in a less-than 100% match for some indicators. Overall, the matching rate at the indicator level had increased for the engine evaluation study compared to the simulation results, with some minor decreased matching rates.

**Table 5.1. Blueprint Constraint by Reporting Category—Simulations**

Grade	Reporting Category	#Items			#Points		
		Min.	Max.	%Match	Min.	Max.	%Match
<b>ELA</b>							
3	Reading Vocabulary	8	9	100.0	9	10	100.0
	Reading Comprehension	22	24	100.0	26	28	100.0
	Writing Skills	8	8	100.0	12	13	100.0
4	Reading Vocabulary	9	9	100.0	9	10	100.0
	Reading Comprehension	24	24	100.0	28	28	100.0
	Writing Skills	8	8	100.0	11	12	100.0
5	Reading Vocabulary	8	9	100.0	9	10	100.0
	Reading Comprehension	22	24	100.0	28	30	100.0
	Writing Skills	10	10	100.0	14	14	100.0
6	Reading Vocabulary	8	9	100.0	9	10	100.0
	Reading Comprehension	22	23	100.0	27	29	100.0
	Writing Skills	9	10	100.0	13	16	100.0
7	Reading Vocabulary	9	9	100.0	10	10	100.0
	Reading Comprehension	22	22	100.0	28	28	100.0
	Writing Skills	10	10	100.0	12	12	100.0
8	Reading Vocabulary	9	9	100.0	9	10	100.0
	Reading Comprehension	21	24	100.0	28	31	100.0
	Writing Skills	11	11	100.0	15	16	100.0
<b>Mathematics</b>							
3	Number	16	16	100.0	17	18	100.0
	Algebra	6	6	100.0	7	8	100.0
	Geometry	11	11	100.0	12	12	100.0
	Data	8	8	100.0	9	9	100.0

Grade	Reporting Category	#Items			#Points		
		Min.	Max.	%Match	Min.	Max.	%Match
4	Number	17	18	100.0	18	19	100.0
	Algebra	10	11	100.0	11	12	100.0
	Geometry	8	9	100.0	9	10	100.0
	Data	6	7	100.0	7	8	100.0
5	Number	16	17	100.0	17	18	100.0
	Algebra	10	10	100.0	11	11	100.0
	Geometry	8	9	100.0	9	10	100.0
	Data	5	6	100.0	6	7	100.0
6	Number	11	12	100.0	12	13	100.0
	Algebra	14	15	100.0	15	16	100.0
	Geometry	8	9	100.0	9	10	100.0
	Data	7	8	100.0	8	9	100.0
7	Number	9	9	100.0	10	10	100.0
	Algebra	14	15	100.0	15	16	100.0
	Geometry	8	9	100.0	9	10	100.0
	Data	9	10	100.0	10	11	100.0
8	Number	10	11	100.0	11	12	100.0
	Algebra	13	14	100.0	14	15	100.0
	Geometry	12	13	100.0	13	14	100.0
	Data	5	5	100.0	6	6	100.0

**Table 5.2. Blueprint Constraint by Reporting Category—Engine Evaluation**

Grade	Reporting Category	#Items			#Points		
		Min.	Max.	%Match	Min.	Max.	%Match
<b>ELA</b>							
3	Reading Vocabulary	8	9	100.0	9	10	100.0
	Reading Comprehension	22	24	100.0	26	28	100.0
	Writing Skills	8	8	100.0	12	13	100.0
4	Reading Vocabulary	9	9	100.0	9	10	100.0
	Reading Comprehension	24	24	100.0	28	28	100.0
	Writing Skills	8	8	100.0	11	12	100.0
5	Reading Vocabulary	8	9	100.0	9	10	100.0
	Reading Comprehension	22	24	100.0	28	30	100.0
	Writing Skills	10	10	100.0	14	14	100.0
6	Reading Vocabulary	8	9	100.0	9	10	100.0
	Reading Comprehension	22	23	100.0	27	29	100.0
	Writing Skills	9	10	100.0	13	16	100.0
7	Reading Vocabulary	9	9	100.0	10	10	100.0
	Reading Comprehension	22	22	100.0	28	28	100.0
	Writing Skills	10	10	100.0	12	12	100.0
8	Reading Vocabulary	9	9	100.0	9	10	100.0
	Reading Comprehension	21	24	100.0	28	31	100.0
	Writing Skills	11	11	100.0	15	16	100.0

Grade	Reporting Category	#Items			#Points		
		Min.	Max.	%Match	Min.	Max.	%Match
<b>Mathematics</b>							
3	Number	16	16	100.0	17	18	100.0
	Algebra	6	6	100.0	7	8	100.0
	Geometry	11	11	100.0	12	12	100.0
	Data	8	8	100.0	9	9	100.0
4	Number	17	18	100.0	18	19	100.0
	Algebra	10	11	100.0	11	12	100.0
	Geometry	8	9	100.0	9	10	100.0
	Data	6	7	100.0	7	8	100.0
5	Number	16	17	100.0	17	18	100.0
	Algebra	10	10	100.0	11	11	100.0
	Geometry	8	9	100.0	9	10	100.0
	Data	5	6	100.0	6	7	100.0
6	Number	11	12	100.0	12	13	100.0
	Algebra	14	15	100.0	15	16	100.0
	Geometry	8	9	100.0	9	10	100.0
	Data	7	8	100.0	8	9	100.0
7	Number	9	9	100.0	10	10	100.0
	Algebra	14	15	100.0	15	16	100.0
	Geometry	8	9	100.0	9	10	100.0
	Data	9	10	100.0	10	11	100.0
8	Number	10	11	100.0	11	12	100.0
	Algebra	13	14	100.0	14	15	100.0
	Geometry	12	13	100.0	13	14	100.0
	Data	5	5	100.0	6	6	100.0

### 5.2.3. Item Exposure Rates

Table 5.3 and Table 5.4 present the item exposure rates from the engine simulation study and post-administration engine evaluation study, respectively. Because students received different items based on blueprint constraints and their ability during the adaptive administration, it is ideal to have a low exposure rate. The exposure rate for each item was calculated as the percentage of students who received that item. For example, if Item 1 was administered to 500 out of 1,000 students, the exposure rate would be 50%. In Table 5.3, “Total” is the total number of items in the operational item pool. “Fixed” is the number of horizontal linking items. “CAT” indicates the pool of adaptive items that students may be administered during the test. “Unused” shows the percentage of unused items that were never administered to students. In Table 5.4, “Total” is the total number of items in the operational item pool except the vertical linking and field test items. “Administered” indicates the number of adaptive items administered to students during the test. “Unused” items were never administered to students during the Spring 2019 administration.

The patterns of exposure rate for the engine evaluation study are very similar to the simulation results. For both studies, all horizontal linking items were also part of the item exposure rate calculation. Horizontal Form 1 (i.e., the core form) given to all students had a 100% exposure rate and is therefore included in the 81–100% exposure rate bin, and the horizontal linking Set A and Set B each had an approximately 50% exposure rate for Grades 4–7 and are therefore included in the 41–60% exposure rate bin.

For the engine evaluation study, most items across grades and content areas had a 0–20% exposure rate, as expected. Compared to the simulation results, the unused percentage of adaptive items slightly decreased, most likely because the operational test had more students than the simulation study. Mathematics Grade 5 had the lowest percentage of unused items because of the limited number of items targeting On Track students. ELA Grades 3 and 8 had the highest percentage of unused items. One possible reason is that ELA Grade 3 only 15 items were adaptively selected in 2019 to provide a paired passage, while all the other grades had 20 adaptively selected items. For ELA Grade 8, a possible reason is the large proportion of 2-point items. To meet the blueprint, ELA Grade 8 needs at least six polytomous items for Reading Comprehension and four polytomous items for Writing Skills. This requirement was confounded with DOK requirements and selected anchor items, plus all Reading Comprehension items were associated with passages. This may have resulted in the high percentage of unused items.

**Table 5.3. Item Exposure Rates—Simulations**

Grade	#Items				Exposure Rate											
					0–20%		21–40%		41–60%		61–80%		81–100%		100%	
	Total	Fixed	CAT	Unused %	N	%	N	%	N	%	N	%	N	%	N	%
<b>ELA</b>																
3	513	26*	487	75.4	70	55.6	17	13.5	11	8.7	1	0.8	27	21.4	26	20.6
4	500	28	472	40.8	240	81.1	15	5.1	25	8.5	–	–	16	5.4	14	4.7
5	436	28	408	51.1	156	73.2	16	7.5	18	8.5	6	2.8	17	8.0	15	7.0
6	453	28	425	54.3	149	72.0	16	7.7	21	10.1	5	2.4	16	7.7	14	6.8
7	472	28	444	44.3	203	77.2	20	7.6	20	7.6	2	0.8	18	6.8	16	6.1
8	481	21	460	77.5	58	53.7	10	9.3	7	6.5	3	2.8	30	27.8	24	22.2
<b>Mathematics</b>																
3	336	21	315	35.7	172	79.6	11	5.1	6	2.8	4	1.9	23	10.7	21	9.7
4	212	28	184	24.5	101	63.1	18	11.3	21	13.1	6	3.8	14	8.8	14	8.8
5	224	28	196	8.9	149	73.0	16	7.8	20	9.8	4	2.0	15	7.4	14	6.9
6	322	28	294	23.9	193	78.8	19	7.8	18	7.4	1	0.4	14	5.7	14	5.7
7	255	28	227	26.3	124	66.0	24	12.8	22	11.7	2	1.1	16	8.5	14	7.4
8	230	21	209	40.4	85	62.0	21	15.3	6	4.4	–	–	25	18.3	24	17.5

\*ELA Grade 3 has an additional user-defined set for paired passages. The set has five items.

**Table 5.4. Item Exposure Rates—Engine Evaluation**

Grade	#Items			Exposure Rate											
				0–20%		21–40%		41–60%		61–80%		81–100%		100%	
	Total	Administered	%Unused	N	%	N	%	N	%	N	%	N	%	N	%
<b>ELA</b>															
3	513	129	74.9	73	56.6	16	12.4	11	8.5	2	1.6	27	20.9	26	20.2
4	500	339	32.2	280	82.6	19	5.6	24	7.1	2	0.6	14	4.1	14	4.1
5	436	237	45.6	180	76.0	17	7.2	17	7.2	6	2.5	17	7.2	15	6.3
6	453	252	44.4	193	76.6	17	6.8	20	7.9	7	2.8	15	6.0	14	5.6
7	472	312	33.9	251	80.5	18	5.8	23	7.4	1	0.3	19	6.1	16	5.1
8	481	112	76.7	64	57.1	8	7.1	6	5.4	4	3.6	30	26.8	24	21.4
<b>Mathematics</b>															
3	336	225	33.0	180	80.0	12	5.3	6	2.7	4	1.8	23	10.2	21	9.3
4	212	161	24.1	102	63.4	16	9.9	22	13.7	6	3.7	15	9.3	14	8.7
5	224	207	7.6	153	73.9	14	6.8	21	10.1	4	1.9	15	7.3	14	6.8
6	322	249	22.7	194	77.9	21	8.4	19	7.6	1	0.4	14	5.6	14	5.6
7	255	203	20.4	138	68.0	24	11.8	22	10.8	4	2.0	15	7.4	14	6.9
8	230	137	40.4	85	62.0	20	14.6	5	3.7	2	1.5	25	18.3	24	17.5

**5.2.4. Score Precision and Reliability**

The pre-administration evaluation using simulations provided precision ability estimations that showed how well the constraint-based engine recovered students’ true ability based on the item pool. Both the pre- and post-administration studies included the standard deviation of estimated theta, mean SEM, SEM by deciles, and marginal reliability.

The following indexes were used to examine the functionality of the constraint-based engine during the pre-administration simulations:

- Precision of ability estimation (how well the engine recovered students’ true ability based on the item pool):
  - Bias: Shows the difference between true and final estimated theta.
  - *P*-value for the z-test: Determines if the difference of bias between the true and final estimated theta is statistically different. If the *p*-value is larger than 0.05, there is no statistical difference of bias between the true and final estimated theta.
  - Mean standard error (MSE): Provides the square of the bias statistic. While bias shows the difference between true and final estimated theta, MSE shows the magnitude of the difference.
  - 95% and 99% coverage: Shows the percentage of students who fall outside of that range in terms of theta.

Table 5.5 presents the results of the precision ability estimation from the pre-administration simulations. The mean biases across all students are small, ranging from -0.03 to 0.02 for both ELA and Mathematics. The *p*-value supports the null-hypothesis that there is not a significant difference between the simulated students’ true and final estimated thetas. The MSE is also relatively small, showing that the constraint-based engine typically recovered a value near the student’s true theta.

**Table 5.5. Mean Bias of the Ability Estimation (True - Estimated)—Simulations**

Grade	Bias		P-Value for Z-Test	MSE	95% Coverage	99% Coverage
	Mean	SE				
<b>ELA</b>						
3	-0.02	0.01	0.60	0.11	4.40	0.30
4	0.01	0.01	0.87	0.11	5.30	0.70
5	-0.01	0.01	0.75	0.09	4.00	0.60
6	0.00	0.01	0.93	0.09	5.60	0.90
7	-0.01	0.01	0.73	0.10	4.10	0.50
8	-0.02	0.01	0.59	0.10	5.20	1.00
<b>Mathematics</b>						
3	-0.02	0.01	0.56	0.13	4.80	0.60
4	-0.01	0.01	0.83	0.13	4.90	1.30
5	-0.03	0.01	0.46	0.12	4.80	0.70
6	-0.02	0.01	0.70	0.12	4.50	1.10
7	0.02	0.01	0.62	0.12	4.1	0.60
8	0.02	0.01	0.53	0.12	4.20	0.90

Table 5.6 and Table 5.7 present the score precision and reliability estimates for the simulation and engine evaluation studies, respectively, including the average number of items administered, the standard deviation (SD) of the estimated theta, the mean SEM, the RMSE, and a marginal reliability coefficient. For both studies, the SD, mean SEM, and RMSE are relatively small. The marginal reliability for the simulations ranges from 0.90 to 0.92 for ELA and 0.86 to 0.88 for Mathematics, whereas it ranges from 0.89 to 0.90 across grades for ELA and is 0.92 for all grades in Mathematics. These results indicate that, overall, the score precision is relatively good.

**Table 5.6. Score Precision and Reliability—Simulations**

Grade	Average #Items	SD of Estimated Theta	Mean SEM	RMSE	Reliability
<b>ELA</b>					
3	41	1.17	0.33	0.33	0.92
4	41	1.02	0.32	0.32	0.90
5	41	0.98	0.31	0.31	0.90
6	41	1.00	0.30	0.30	0.91
7	41	1.12	0.32	0.33	0.91
8	41	0.97	0.31	0.31	0.90
<b>Mathematics</b>					
3	41	1.29	0.35	0.36	0.92
4	41	1.31	0.35	0.36	0.93
5	41	1.25	0.34	0.35	0.92
6	41	1.35	0.34	0.35	0.93
7	41	1.27	0.34	0.34	0.93
8	41	1.31	0.35	0.36	0.93

**Table 5.7. Score Precision and Reliability—Engine Evaluation**

Grade	Average #Items	SD of Estimated Theta	Mean SEM	RMSE	Reliability
<b>ELA</b>					
3	41	1.00	0.32	0.32	0.90
4	41	1.01	0.32	0.32	0.90
5	41	0.94	0.30	0.31	0.89
6	41	0.94	0.29	0.30	0.90
7	41	1.00	0.31	0.32	0.90
8	41	0.94	0.30	0.31	0.89
<b>Mathematics</b>					
3	41	1.31	0.35	0.36	0.92
4	41	1.22	0.35	0.35	0.92
5	41	1.28	0.35	0.36	0.92
6	41	1.24	0.34	0.34	0.92
7	41	1.20	0.34	0.34	0.92
8	41	1.29	0.35	0.36	0.92

Table 5.8 and Table 5.9 present the average SEM by decile of the true overall proficiency score, including the overall student ability distribution, for both the simulation and evaluation studies, respectively. A decile is similar to a percentile rank, with 10 ranks related to the 10th, 20th...90th, 100th percentile ranks. For both studies, the average SEM is similar across deciles except Decile 1 and Decile 10 that have a higher standard error compared to the other deciles. Overall, the SEM is in acceptable ranges.

**Table 5.8. SEM by Deciles—Simulations**

Grade	Proficiency Score Distribution										Overall
	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	
<b>ELA</b>											
3	0.37	0.31	0.29	0.29	0.29	0.30	0.30	0.32	0.34	0.44	0.33
4	0.36	0.31	0.29	0.29	0.29	0.29	0.30	0.32	0.34	0.39	0.32
5	0.36	0.31	0.29	0.28	0.28	0.28	0.28	0.29	0.31	0.37	0.31
6	0.34	0.29	0.28	0.28	0.27	0.27	0.28	0.29	0.31	0.38	0.30
7	0.35	0.30	0.29	0.29	0.29	0.29	0.30	0.32	0.35	0.44	0.32
8	0.33	0.30	0.29	0.29	0.29	0.29	0.29	0.30	0.32	0.37	0.31
<b>Mathematics</b>											
3	0.35	0.32	0.31	0.31	0.32	0.33	0.34	0.36	0.39	0.50	0.35
4	0.37	0.33	0.32	0.32	0.32	0.33	0.34	0.35	0.37	0.46	0.35
5	0.34	0.31	0.31	0.31	0.31	0.31	0.33	0.34	0.37	0.48	0.34
6	0.36	0.32	0.31	0.31	0.32	0.32	0.33	0.34	0.37	0.45	0.34
7	0.41	0.33	0.32	0.32	0.31	0.31	0.32	0.33	0.35	0.40	0.34
8	0.41	0.34	0.33	0.32	0.32	0.33	0.33	0.34	0.36	0.44	0.35

**Table 5.9. SEM by Deciles—Engine Evaluation**

Grade	Proficiency Score Distribution										Overall
	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	
<b>ELA</b>											
3	0.34	0.30	0.29	0.29	0.29	0.29	0.30	0.31	0.33	0.39	0.32
4	0.36	0.31	0.29	0.29	0.28	0.29	0.30	0.32	0.34	0.40	0.32
5	0.35	0.31	0.30	0.28	0.28	0.28	0.28	0.29	0.31	0.37	0.30
6	0.34	0.30	0.28	0.27	0.27	0.27	0.27	0.29	0.30	0.35	0.29
7	0.35	0.30	0.29	0.29	0.29	0.29	0.30	0.31	0.33	0.37	0.31
8	0.35	0.31	0.29	0.29	0.29	0.29	0.29	0.29	0.30	0.35	0.30
<b>Mathematics</b>											
3	0.35	0.32	0.31	0.31	0.32	0.33	0.34	0.36	0.39	0.50	0.35
4	0.36	0.33	0.32	0.32	0.32	0.33	0.34	0.35	0.37	0.44	0.35
5	0.33	0.31	0.31	0.30	0.31	0.31	0.32	0.34	0.38	0.53	0.35
6	0.36	0.32	0.31	0.31	0.31	0.32	0.33	0.34	0.36	0.43	0.34
7	0.38	0.34	0.32	0.32	0.31	0.31	0.31	0.32	0.34	0.40	0.34
8	0.38	0.34	0.32	0.32	0.32	0.32	0.33	0.34	0.37	0.47	0.35



## Section 6: Psychometric Analyses

During the Spring 2019 testing window, the pre-equated item parameter estimates were used to score student responses and select the next items to administer for the adaptive portions of the NSCAS Summative ELA and Mathematics assessments. After the testing window was closed, the following post-administration analyses were conducted to calibrate the items for ELA, Mathematics, and Science (e.g., to construct the ELA and Mathematics vertical scales). The purpose of conducting these analyses is to establish the psychometric quality of the items used in the assessments, which will bolster the arguments regarding the validity of the interpretations and uses of the test scores.

- Classical item analyses
- Differential item functioning (DIF)
- Item response theory (IRT) calibration
- Vertical scaling for ELA and Mathematics
- Post-equating check for Science

### 6.1. Number of Student Included in the Analyses

Table 6.1 presents the number of students included in the post-administration analyses presented in this section (i.e., classical analyses, DIF, IRT calibration, equating, and scaling). As in the 2018 technical report, only online test-takers who attempted at least 10 operational items were used. The results from these students are referred to as the “analyses data.” It is typically ideal to use 100% of the student data, including both online and paper-pencil tests. However, NDE decided to use only online tests due to the goal of completing the standard setting by the end of July 2018 and because the number of paper-pencil test-takers was less than 100 for each grade.

**Table 6.1. Number of Students Included in the Psychometric Analyses**

Content Area	Grade	Test ID	N
ELA	3	4220	23,414
	4	4221	23,940
	5	4222	23,942
	6	4223	22,354
	7	4224	23,476
	8	4225	23,071
Mathematics	3	4226	23,373
	4	4227	23,897
	5	4228	23,897
	6	4229	22,308
	7	4230	23,421
	8	4231	23,028
Science	5	4304	23,900
	8	4305	23,026

## 6.2. Classical Item Analyses

This section summarizes the  $p$ -values and item-total correlations for operational and field test items. Appendix F provides the classical item-level statistics. Off-grade vertical linking items are included in the operational tables. Omit rates across all content areas and grades were close to 0, which is to be expected since students were required to answer each item before moving on to the next one. Additionally, item statistics obtained from less than 100 students were not calibrated and therefore not used for calibration and subsequent analyses. For such items, item parameters on the old scale were transformed on to the new scale and used for student scoring.

### 6.2.1. Item Difficulty ( $P$ -Value)

Item difficulty is measured by the  $p$ -value that shows the proportion of students who answered an item correctly and is bounded by 0 and 1. Generally, a high  $p$ -value indicates that an item is easy (i.e., high proportion of students answered it correctly), whereas a low  $p$ -value indicates that an item is hard. For example, a  $p$ -value of 0.79 indicates that 79% of students answered the item correctly. For polytomous items, the  $p$ -value is the average item score (i.e., the sum of student scores on an item divided by the total number of students who responded to the item) divided by the number of possible score points on the item.

Table 6.2 and Table 6.3 present the summary statistics for the  $p$ -values across all operational and field test items, respectively, including the number of items by  $p$ -value range (i.e., less than or equal to a  $p$ -value of 0.1, 0.2, etc.). These data were calculated for items with and without a representative sample (i.e., horizontal linking and field test items vs. adaptive items, respectively). Items without a representative sample are those administered during the adaptive stage of the assessment, and the expected  $p$ -value is typically between 0.4 and 0.6 for these items. Appendix G provides the summary  $p$ -value statistics by item type. Typically, test developers target  $p$ -values in the range of 0.3 to 0.8. The average  $p$ -values range for the 2019 NSCAS Summative assessments range from 0.4 to 0.6 across content areas and grades.

**Table 6.2. Summary  $P$ -Values—Operational Items**

Grade	#Items	Mean	SD	Min.	Max.	#Items by $P$ -Value Range									
						≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	> 0.9
<b>ELA</b>															
3	143	0.495	0.164	0	1	1	0	12	31	35	26	24	10	3	1
4	366	0.472	0.193	0	1	14	14	25	70	101	63	39	20	11	9
5	265	0.470	0.178	0	1	13	6	11	49	77	56	33	13	3	4
6	280	0.462	0.206	0	1	12	10	31	59	58	53	27	15	5	10
7	338	0.443	0.211	0	1	23	13	36	59	99	51	20	16	10	11
8	124	0.506	0.171	0	0.949	2	3	6	17	39	21	20	10	3	3
<b>Mathematics</b>															
3	240	0.495	0.139	0.160	1	0	2	18	29	87	58	27	12	4	3
4	189	0.507	0.147	0.012	0.941	1	2	4	32	63	48	16	13	9	1
5	234	0.515	0.138	0	0.883	1	1	6	31	80	59	29	19	8	0
6	277	0.499	0.133	0	1	1	3	9	34	106	83	18	12	9	2
7	229	0.493	0.141	0	1	1	3	11	40	69	66	23	10	4	2
8	151	0.502	0.123	0.113	0.821	0	1	5	18	55	46	12	12	2	0
<b>Science</b>															
5	50	0.636	0.165	0.338	0.908	0	0	0	4	9	9	8	9	10	1
8	60	0.604	0.165	0.193	0.902	0	1	2	5	6	17	10	12	6	1

**Table 6.3. Summary P-Values—Field Test Items**

Grade	#Items	Mean	SD	Min.	Max.	#Items by P-Value Range									
						≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	> 0.9
<b>ELA</b>															
3	180	0.443	0.161	0.012	0.921	5	2	29	40	50	19	27	6	1	1
4	191	0.573	0.162	0.094	0.891	1	2	8	18	32	41	50	21	18	0
5	185	0.536	0.177	0.107	0.935	0	8	11	17	43	33	40	20	11	2
6	192	0.512	0.174	0.135	0.880	0	9	14	34	33	39	35	21	7	0
7	195	0.520	0.169	0.074	0.919	1	7	14	25	37	43	41	21	5	1
8	184	0.529	0.194	0.088	0.934	1	5	17	38	18	34	32	22	16	1
<b>Mathematics</b>															
3	231	0.553	0.199	0.004	0.944	3	7	18	29	30	38	47	34	21	4
4	231	0.516	0.182	0.105	0.904	0	11	20	31	49	43	40	25	11	1
5	231	0.573	0.173	0.116	0.934	0	6	8	24	37	56	42	31	22	5
6	231	0.496	0.207	0.052	0.982	5	16	20	40	35	38	40	20	13	4
7	231	0.440	0.207	0.038	0.874	11	21	33	42	33	35	25	25	6	0
8	231	0.439	0.211	0.078	0.927	8	21	49	35	29	25	33	20	9	2

**6.2.2. Item Discrimination (Item-Total Correlation)**

Item-total correlation describes the relationship between performance on a specific item and performance on the entire test based on the overall test score. Students who do well on a test are expected to select the right answer to any given item, and students who do poorly are expected to select the wrong answer. This means that for a highly discriminating item, students who get the item correct will have a higher average test score than students who get the item incorrect. The item-total correlation coefficient ranges between -1.0 and +1.0. An item with a high positive item-total correlation discriminates between low-performing and high-performing students better than an item with an item-total correlation near zero. A negative item-total correlation indicates that lower-performing students did better on that item than higher-performing students. However, a very difficult item (or a very easy item) would have little variance in student responses, meaning most students respond incorrectly (or correctly). The resulting item-total correlation is typically low since both groups have the same score.

Table 6.4 and Table 6.5 present the summary statistics for the item-total correlations across all operational and field items, respectively. Appendix H provides the results by item type. Instead of using the number-correct score, the estimated final theta score was used to compute the item-total correlations because number-correct scores would not provide much insight into student performance on an adaptive test since, in theory, all students get 50% correct on an adaptive assessment. The results appear out of bounds from traditional metrics, but this is because the NSCAS ELA and Mathematics tests were adaptive. Due to adaptive selection of items, some items were administered to a small number of students. The relatively higher number of ELA items in the ≤ 0.2 range are mostly obtained from n-counts less than 100 (and were therefore not calibrated in 2019). Therefore, the means of the correlations are reasonable and the number of items with less than 0.2 are relatively small.

**Table 6.4. Summary Item-Total Correlations—Operational Items**

Grade	#Items	Mean	SD	Min.	Max.	#Items by Item-Total Correlation Range						
						≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	> 0.6
<b>ELA</b>												
3	143	0.325	0.175	-0.866	1.000	7	10	43	43	25	11	4
4	366	0.315	0.172	-0.869	0.859	34	29	81	131	53	24	14
5	265	0.321	0.187	-1.000	1.000	25	19	58	87	49	17	10
6	280	0.314	0.197	-1.000	1.000	33	20	69	78	53	15	12
7	338	0.283	0.217	-1.000	1.000	48	39	80	94	49	12	16
8	124	0.367	0.135	0.000	0.818	2	8	30	34	33	13	4
<b>Mathematics</b>												
3	240	0.364	0.117	-0.374	1.000	3	3	41	122	51	16	4
4	189	0.355	0.094	0.040	0.663	1	6	44	86	37	13	2
5	234	0.360	0.089	0.000	0.602	1	5	49	108	54	16	1
6	277	0.355	0.120	-0.832	0.694	6	6	51	130	64	17	3
7	229	0.336	0.160	-0.621	0.899	10	9	52	90	47	17	4
8	151	0.360	0.093	-0.018	0.680	2	1	31	78	27	10	2
<b>Science</b>												
5	50	0.380	0.073	0.230	0.525	0	0	8	22	17	3	0
8	60	0.377	0.077	0.158	0.540	0	1	10	21	27	1	0

**Table 6.5. Summary Item-Total Correlations—Field Test Items**

Grade	#Items	Mean	SD	Min.	Max.	#Items by Item-Total Correlation Range						
						≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	> 0.6
<b>ELA</b>												
3	180	0.329	0.128	-0.011	0.609	8	27	35	55	45	9	1
4	191	0.350	0.125	-0.159	0.567	9	16	26	63	61	16	0
5	185	0.323	0.131	-0.111	0.581	13	19	39	58	44	12	0
6	192	0.324	0.134	-0.014	0.649	15	20	38	56	49	12	2
7	195	0.338	0.136	-0.070	0.598	14	17	30	64	53	17	0
8	184	0.322	0.138	-0.129	0.638	14	15	40	63	36	15	1
<b>Mathematics</b>												
3	231	0.411	0.114	-0.001	0.742	4	5	26	66	87	33	10
4	231	0.419	0.096	0.105	0.627	0	6	24	51	108	37	5
5	231	0.420	0.107	0.051	0.643	1	5	29	60	75	58	3
6	231	0.391	0.117	0.023	0.665	4	11	34	67	72	40	3
7	231	0.400	0.125	-0.009	0.640	6	11	25	54	88	40	7
8	231	0.405	0.117	-0.054	0.647	3	11	21	63	86	40	7

### 6.2.3. Item Suppression

Based on the item analysis conducted using the Spring 2019 results and the flagging criteria presented in Table 6.6 and Table 6.7 for multiple-choice and partial-credit (i.e., non-multiple-choice) items, a total of 599 MC and 35 non-MC items from the adaptive forms across both content areas were flagged. After removing items with n-counts less than 100 (statistics for items with  $N < 100$  are considered to be unstable), 264 MC items and 26 non-MC items from the adaptive assessments were identified for content and psychometric review. Due to very low student n-counts, the following items with blank or negative item-total correlations were identified for content and psychometric review for fixed forms (i.e., English paper-pencil, Spanish online, and Spanish paper-pencil): 35 MC items and four non-MC items for English paper-pencil and 73 MC items and five non-MC items for Spanish online.

**Table 6.6. Flagging Criteria for MC Items**

Flag Type*	Criterion	Indication
low $p$ -value	$< 0.20$	very difficult item
high $p$ -value	$> 0.90$	very easy item
low item-total	$< 0.20$	poorly discriminating item
omit rate	$> 5\%$	unclear or very difficult item
high item-total for a distractor	$> 0.05$	poorly discriminating item
the key not being the most popular answer choice	$p$ -value of the key $<$ $p$ -value of a distractor	possible miskey

\*item-total = item-total correlation

**Table 6.7. Flagging Criteria for Partial-Credit Items**

Flag Type*	Criterion
low item-total	$< 0.10$
high item-total for a score of 0	$> 0$
item-total for a score of 1 is less than item-total for a score of 0	score of 1 item-total $<$ score of 0 item-total
low item-total for a score of 0	$< 0.10$
item-total for a score of 2 is less than item-total for a score of 1	score of 2 item-total $<$ score of 1 item-total
low student count for each score	$< 0$

\*item-total = item-total correlation. All flags in this table indicate poor discrimination.

After the content and psychometric team reviewed these flagged items, NWEA recommended suppressing seven items (four ELA and three Mathematics items) from the 2019 scoring and removing 12 items (nine ELA and three Mathematics items) from the future item pool, as shown in Table 6.8. NWEA also recommended removing one fixed-form item from the 2019 scoring, as indicated in the table (ELA Grade 6 Item 11191450). All recommendations were approved by NDE, so these suppressed items were not included for all subsequent analyses and score reporting. There was no suppression for Science.

**Table 6.8. Suppressed Items**

Grade	Item Code	Item Role*	Item Type	Standard	Max. #Points	Key	NWEA Approved Recommendations	
							2019 Scoring	2020 Pool & Later
<b>ELA</b>								
3	21060240	OP	Choice	LA 3.1.6.f	1	B	Retain	Remove
4	21079160	OP	Choice	LA 4.1.5.d	1	C	Retain	Remove
4	21079210	OP	Choice	LA 4.1.5.b	1	D	Suppress	Remove
6	11191450**	OP	Choice	LA 6.2.2.3	1	–	Suppress	Remove
7	21074700	OP	Choice	LA 7.1.6.i	1	B	Suppress	Remove
7	21074740***	HL	Choice	LA 7.1.5.d	1	B	Suppress	Remove
7	21095730	OP	Choice	LA 7.1.5.d	1	A	Retain	Remove
7	21096100	OP	Choice	LA 7.1.6.j	1	A	Retain	Remove
8	11188200	OP	Choice	LA 8.1.6.g	1	B	Retain	Remove
8	21048960	OP	Choice	LA 8.1.5.b	1	B	Suppress	Remove
8	21074740***	VL	Choice	LA 7.1.5.d	1	B	(see above)	(see above)
<b>Mathematics</b>								
5	31157900	OP	Composite	MA 5.2.3.a	2	–	Suppress	Remove
7	31160150	OP	Composite	MA 7.1.2.b	2	–	Suppress	Remove
7	31161220	OP	Composite	MA 7.4.3.e	2	–	Suppress	Remove

\*OP = operational. VL = vertical linking. HL = horizontal linking.

\*\*ELA Grade 6 Item 11191450 is from the fixed-form assessment. The item was flagged because all 25 students who took the test got a score of 0. This item was not included in the CAT pool.

\*\*\*ELA Item 21074740 is a HL item in Grade 7 and a VL item in Grade 8. This item is also on the English paper-pencil assessment (Grade 5, Item 30) and the Spanish assessment.

### 6.3. Differential Item Functioning (DIF)

DIF is a statistical procedure that flags items for potential bias. The fundamental measurement assumption of DIF is that the probability of a correct response to a test item is a function of the item’s difficulty and the student’s ability. This function is expected to remain invariant to other person characteristics unrelated to ability such as gender and ethnicity. Therefore, if two students with the same ability respond to the same item, they are assumed to have an equal probability of answering the item correctly. To test this assumption, responses to items by students sharing an aspect of a person characteristic (e.g., gender) are compared to responses to the same items by other students who share a different aspect of the same characteristic (e.g., males vs. females). The group representing students in a specific demographic group is referred to as the *focal* group. The group comprised of students from outside this group is referred to as the *reference* group. Table 6.9 presents the focal and reference groups for the NSCAS DIF analyses.

**Table 6.9. Focal and Reference Groups for Gender- and Ethnicity-Based DIF**

Group Type	Focal Group	Reference Group
Gender	Female	Male
Ethnicity	Black or African American	White
	Hispanic	White
	Asian	White
	Two or More Races	White

When DIF is detected and the fundamental measurement assumption does not hold (i.e., students with the same ability in different groups of interest have different probabilities of correctly answering an item), the item is said to be functioning differently for the two groups. The presence of DIF in an item suggests that the item is functioning unexpectedly regarding the groups included in the comparison. The cause of the unexpected functioning is not revealed in a DIF analysis. It may be that item content is inadvertently providing an advantage or disadvantage to members of one of the two groups. Content experts who have special knowledge of the groups involved can often identify a cause of this type. DIF may also result from differential instruction closely associated with group membership.

### 6.3.1. DIF Methods

The Mantel-Haenszel (MH) procedure was used to detect DIF for dichotomous items (Holland & Thayer, 1988), and the standardized mean difference (SMD) analysis, developed as an extension of the MH procedure, was used to detect DIF for polytomous items (Dorans & Schmitt, 1991; Zwick, Donoghue, & Grima, 1993). The MH method has been widely used in educational measurement due to its easy implementation in testing programs. The procedure compares the ratio of the probabilities of two groups of students (i.e., focal and reference groups) answering an item correctly across all score levels. The obtained estimate is known as the odds ratio, which is computed as follows:

$$\alpha_{MH} = \frac{\left(\frac{\sum_m R_{rm} W_{fm}}{N_m}\right)}{\left(\frac{\sum_m R_{fm} W_{rm}}{N_m}\right)} \quad (6.1)$$

where:

- $R_{rm}$  is the number of students in the reference group at ability level  $m$  answering the item correctly.
- $W_{fm}$  is the number of students in the focal group at ability level  $m$  answering the item incorrectly.
- $R_{fm}$  is the number of students in the focal group at ability level  $m$  answering the item correctly.
- $W_{rm}$  is the number of students in the reference group at ability level  $m$  answering the item incorrectly.
- $N_m$  is the total number of students at ability level  $m$ .

This value can then be used as follows (Holland & Thayer, 1988):

$$MH\ D - DIF = -2.35 \ln(\alpha_{MH}) \quad (6.2)$$

The MH chi-square statistic used to classify items into DIF categories is as follows:

$$MH\ CHISQ = \frac{\left(|\sum_m R_{rm} - \sum_m E(R_{rm})| - \frac{1}{2}\right)^2}{\sum_m Var(R_{rm})} \quad (6.3)$$

where:

- $E(R_{rm}) = \frac{N_{rm} R_{Nm}}{N_m}$ ,  $Var(R_{rm}) = \frac{N_{rm} N_{fm} R_{Nm} W_{Nm}}{N_m^2 (N_m - 1)}$
- $N_{rm}$  and  $N_{fm}$  are the numbers of students in the reference and focal groups, respectively.
- $R_{Nm}$  and  $W_{Nm}$  are the number of students who answered the item correctly and incorrectly, respectively.

SMD for polytomous items compares item performance of two subpopulations adjusting for differences in the distributions of the two subpopulations. The standardized mean difference statistic can be divided by the total standard deviation to obtain a measure of the effect size. A negative value of the standardized mean difference shows that the item is more difficult for the focal group, whereas a positive value indicates that it is more difficult for the reference group. The standardized mean difference used for polytomous items is defined as:

$$SMD = \sum p_{FK}m_{FK} - \sum p_{FK}m_{RK} \quad (6.4)$$

where:

- $p_{FK}$  is the proportion of the focal group students at the  $k^{\text{th}}$  level of the matching variable.
- $m_{FK}$  is the mean score for the focal group at the  $k^{\text{th}}$  level.
- $m_{RK}$  is the mean item score for the reference group at the  $k^{\text{th}}$  level.

The SMD is divided by the total item group standard deviation to get a measure of the effect size.

Table 6.10 and Table 6.11 present the Educational Testing Service (ETS) DIF categories for classifying the DIF results. The ETS method of categorizing DIF allows items exhibiting negligible DIF (Category A) to be differentiated from those exhibiting moderate DIF (Category B) and strong DIF (Category C). Categories B and C have a further breakdown as “+” (DIF is in favor of the focal group) or “-” (DIF is in favor of the reference group).

**Table 6.10. DIF Categories for Dichotomous Items**

DIF Category	Level of DIF	Definition
A	Negligible	<ul style="list-style-type: none"> <li>• Absolute value of the Mantel-Haenszel delta difference (MH D-DIF) is not significantly different from 0 or is less than one.</li> </ul>
B	Moderate	<ul style="list-style-type: none"> <li>• Absolute value of the MH D-DIF is significantly different from 0 but not from one, and is at least 1; or</li> <li>• Absolute value of the MH D-DIF is significantly different from 1, but less than 1.5.</li> <li>• Positive values are classified as “B+” and negative values as “B-”.</li> </ul>
C	Strong	<ul style="list-style-type: none"> <li>• Absolute value of the MH D-DIF is significantly different from 1, and is at least 1.5; and</li> <li>• Absolute value of the MH D-DIF is larger than 1.96 times the standard error of MH D-DIF.</li> <li>• Positive values are classified as “C+” and negative values are “C-”.</li> </ul>

**Table 6.11. DIF Categories for Polytomous Items**

DIF Category	Level of DIF	Definition
A	Negligible	Mantel $p$ -value $>0.05$ or chi-square $ SMD/SD  \leq 0.17$
B	Moderate	Mantel chi-square $p$ -value $<0.05$ and $ SMD/SD  >0.17$ , but $\leq 0.25$
C	Strong	Mantel chi-square $p$ -value $<0.05$ and $ SMD/SD  > 0.25$



### 6.3.2. DIF Results

Table 6.12 and Table 6.13 present the number of items assigned to each DIF category for operational and field test items, respectively. Male was the reference group for gender, and white was the reference group for ethnicity. DIF was not conducted if the sample size for either group was less than 250. Appendix I presents the item-level DIF statistics. The + sign next to the DIF category indicates that the item is in favor of the reference group, and the - sign indicates that the item is in favor of the focal group. As shown in the tables, most items were categorized as DIF Category A (negligible DIF).

**Table 6.12. DIF Results—Operational Items**

Grade	Focal Group	#Items by DIF Category					
		Total	A	B+	B-	C+	C-
<b>ELA</b>							
3	Female	100	97	1	1	1	–
	Black or African American	54	54	–	–	–	–
	Hispanic	82	82	–	–	–	–
	Asian	41	38	1	2	–	–
	Two or More Races	52	52	–	–	–	–
4	Female	198	183	9	1	1	4
	Black or African American	63	60	–	3	–	–
	Hispanic	112	109	–	1	–	2
	Asian	40	33	3	1	–	3
	Two or More Races	55	54	–	1	–	–
5	Female	153	142	2	5	–	4
	Black or African American	67	65	–	2	–	–
	Hispanic	92	90	–	–	–	2
	Asian	43	40	–	2	–	1
	Two or More Races	54	54	–	–	–	–
6	Female	155	142	5	7	–	1
	Black or African American	62	61	1	–	–	–
	Hispanic	103	100	0	1	–	2
	Asian	42	39	1	1	1	–
	Two or More Races	50	50	–	–	–	–
7	Female	146	137	5	1	1	2
	Black or African American	61	60	–	1	–	–
	Hispanic	89	85	–	3	–	1
	Asian	42	42	–	–	–	–
	Two or More Races	50	50	–	–	–	–
8	Female	97	92	2	1	–	2
	Black or African American	51	48	1	2	–	–
	Hispanic	70	67	–	2	–	1
	Asian	39	35	2	1	–	1
	Two or More Races	46	46	–	–	–	–

Grade	Focal Group	#Items by DIF Category					
		Total	A	B+	B-	C+	C-
<b>Mathematics</b>							
3	Female	162	145	3	9	–	5
	Black or African American	52	50	1	–	1	–
	Hispanic	115	110	1	4	–	–
	Asian	39	37	1	1	–	–
	Two or More Races	43	43	–	–	–	–
4	Female	167	157	–	10	–	–
	Black or African American	70	62	6	2	–	–
	Hispanic	121	121	–	–	–	–
	Asian	42	37	5	–	–	–
	Two or More Races	51	51	–	–	–	–
5	Female	200	174	7	10	–	9
	Black or African American	65	63	1	1	–	–
	Hispanic	140	132	1	7	–	–
	Asian	42	38	1	3	–	–
	Two or More Races	47	46	1	–	–	–
6	Female	212	196	2	8	–	6
	Black or African American	55	54	–	1	–	–
	Hispanic	139	135	1	3	–	–
	Asian	35	29	2	3	–	1
	Two or More Races	43	43	–	–	–	–
7	Female	162	155	–	5	1	1
	Black or African American	67	65	–	2	–	–
	Hispanic	89	88	–	1	–	–
	Asian	39	34	–	3	–	2
	Two or More Races	48	48	–	–	–	–
8	Female	146	139	6	1	–	–
	Black or African American	50	50	–	–	–	–
	Hispanic	105	100	2	2	1	–
	Asian	36	33	1	1	–	1
	Two or More Races	43	43	–	–	–	–
<b>Science</b>							
5	Female	50	48	2	–	–	–
	Black or African American	50	50	–	–	–	–
	Hispanic	50	50	–	–	–	–
	Asian	50	47	–	1	2	–
	Two or More Races	50	50	–	–	–	–
8	Female	60	54	2	4	–	–
	Black or African American	60	60	–	–	–	–
	Hispanic	60	60	–	–	–	–
	Asian	60	55	2	1	1	1
	Two or More Races	60	60	–	–	–	–

**Table 6.13. DIF Results—Field Test Items**

Grade	Focal Group	#Items by DIF Category					
		Total	A	B+	B-	C+	C-
<b>ELA</b>							
3	Female	180	162	7	9	1	1
	Black or African American	–	–	–	–	–	–
	Hispanic	2	2	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
4	Female	191	166	14	6	3	2
	Black or African American	–	–	–	–	–	–
	Hispanic	2	1	–	1	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
5	Female	185	165	11	6	2	1
	Black or African American	–	–	–	–	–	–
	Hispanic	–	–	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
6	Female	192	163	8	12	7	2
	Black or African American	–	–	–	–	–	–
	Hispanic	–	–	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
7	Female	195	164	13	7	8	3
	Black or African American	–	–	–	–	–	–
	Hispanic	–	–	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
8	Female	184	165	6	10	2	1
	Black or African American	–	–	–	–	–	–
	Hispanic	–	–	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
<b>Mathematics</b>							
3	Female	231	168	17	26	9	11
	Black or African American	1	1	–	–	–	–
	Hispanic	5	5	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
4	Female	114	97	5	11	–	1
	Black or African American	1	1	–	–	–	–
	Hispanic	5	5	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–

Grade	Focal Group	#Items by DIF Category					
		Total	A	B+	B-	C+	C-
5	Female	231	179	14	19	7	12
	Black or African American	–	–	–	–	–	–
	Hispanic	4	2	–	1	–	1
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
6	Female	59	51	4	3	1	–
	Black or African American	–	–	–	–	–	–
	Hispanic	4	4	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
7	Female	231	174	30	12	9	6
	Black or African American	–	–	–	–	–	–
	Hispanic	4	4	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
8	Female	231	188	13	12	7	11
	Black or African American	–	–	–	–	–	–
	Hispanic	–	–	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–

#### 6.4. IRT Calibration

The Rasch model (Rasch, 1960, 1980; Wright, 1977) for dichotomous items and the partial credit model (PCM; Masters, 1982) for polytomous items were used to calibrate items and create the NSCAS Summative scale. For all content areas, item parameter estimations were implemented using WINSTEPS 3.91.0.0 (Linacre, 2015) that used joint maximum likelihood estimation (MLE) as described by Wright and Masters (1982). The Rasch model has had a long-standing presence in applied testing programs and was the methodology used to calibrate the previous Nebraska State Accountability (NeSA) items. Under the Rasch model, the probability of a student with ability  $\theta$  responding correctly to item  $i$  is as follows, where  $\theta_j$  and  $b_i$  are the person and item parameters, respectively:

$$P(u_{ij} = 1 | \theta_j, b_i) = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}} \quad (6.5)$$

Under the PCM model, the probability of a student with ability  $\theta$  having a score at the  $k$ th level of item  $i$  is:

$$P(u_{ij} = k | \theta_i) = \frac{e^{[\sum_{u=1}^k Da_i(\theta_j - b_i + d_{iu})]}}{\sum_{v=1}^{m_i} e^{[\sum_{u=1}^k Da_i(\theta_j - b_i + d_{iu})]}} \quad (6.6)$$

where  $k$  is the score on the item,  $m_i$  is the total number of score categories for the item,  $d_{iu}$  is the threshold parameter for the threshold between scores  $u$  and  $u - 1$ , and  $\theta_j$  and  $b_i$  are the person and item parameters, respectively.

### 6.4.1. Checking Model Assumptions

It is important to check the three fundamental model assumptions of unidimensionality of the data, local independence, and item fit using the operational items. Overall, the principal component analysis (PCA) of residuals indicates one dominant dimension for all content areas and grades. The median residual correlations are close to 0 and the small number of items with correlations greater than 0.20, suggesting that local item independence generally holds for all content areas and grades. A small number of items outside of 0.7 and 1.3 in terms of infit mean square statistics indicates a good fit.

#### 6.4.1.1. Unidimensionality

Unidimensionality is the most commonly violated assumptions in the latent trait structure implied by the item response data. In most instances, it is sufficient to assume that all items in a test are sensitive to differences in examinees along a single latent trait. However, it is crucial to check if only one dominant dimension exists among the items. PCA is the most commonly used statistical procedure to check how many dimensions exist in the data. Table 6.14 presents the PCA of residuals results. For ELA, one dominant dimension explained most of the variance from 22.7 to 27.8%. A few grades have less than five dimensions with an eigenvalue bigger than 1. A similar pattern was observed for Mathematics and Science.

**Table 6.14. Unidimensionality: Results from PCA of Residuals**

Grade	Components	Eigenvalue	Explained Variance
<b>ELA</b>			
3	Measure	39.6	23.5
	1	2.0	1.2
	2	1.6	1.0
	3	1.6	0.9
	4	1.3	0.8
	5	–	–
4	Measure	108.6	24.3
	1	5.0	1.1
	2	4.0	0.9
	3	2.2	0.5
	4	2.1	0.5
	5	2.0	0.4
5	Measure	69.5	22.7
	1	4.0	1.3
	2	3.2	1.1
	3	2.9	1.0
	4	2.2	0.7
	5	2.1	0.7
6	Measure	75.3	23.0
	1	5.2	1.6
	2	4.1	1.2
	3	–	–
	4	–	–
	5	–	–

Grade	Components	Eigenvalue	Explained Variance
7	Measure	105.1	25.3
	1	4.0	1.0
	2	–	–
	3	–	–
	4	–	–
	5	–	–
8	Measure	42.8	27.8
	1	2.2	1.4
	2	2.0	1.3
	3	1.8	1.2
	4	1.4	0.9
	5	1.3	0.8
<b>Mathematics</b>			
3	Measure	79.6	26.1
	1	2.0	0.7
	2	1.7	0.5
	3	1.5	0.5
	4	1.4	0.5
	5	–	–
4	Measure	57.2	26.2
	1	2.2	1.0
	2	1.8	0.8
	3	1.7	0.8
	4	1.5	0.7
	5	1.5	0.7
5	Measure	59.6	22.4
	1	2.0	0.8
	2	1.8	0.7
	3	–	–
	4	–	–
	5	–	–
6	Measure	86.0	25.7
	1	2.1	0.6
	2	1.8	0.5
	3	1.8	0.5
	4	1.6	0.5
	5	1.6	0.5
7	Measure	60.7	23.2
	1	1.7	0.6
	2	1.4	0.6
	3	1.3	0.5
	4	1.3	0.5
	5	–	–
8	Measure	56.2	29.1
	1	1.7	0.9
	2	1.7	0.9
	3	1.4	0.7
	4	1.4	0.7
	5	1.3	0.7

Grade	Components	Eigenvalue	Explained Variance
<b>Science</b>			
5	Measure	16.9	25.3
	1	1.7	2.5
	2	1.4	2.1
	3	1.2	1.8
	4	–	–
	5	–	–
8	Measure	19.3	24.4
	1	1.9	2.4
	2	1.4	1.8
	3	1.3	1.7
	4	1.3	1.6
	5	1.1	1.4

#### 6.4.1.2. Local Independence

Local independence is a fundamental assumption of Rasch measurement. No relationship should exist between students' responses to different items after accounting for the abilities measured by a test. Many indicators of local independence are framed by the form of local independence proposed by McDonald (1979) that the conditional covariances of all item response pairs, conditioned on the abilities, must be equal to zero. The following residual item correlations provided in WINSTEPS for each item pair were used to assess local dependence among the NSCAS Summative items:

- Raw
- Standardized
- Logit

The raw score residual correlation corresponds to Yen's Q3 index, a popular local independence statistic. The expected value for the Q3 statistic is approximately  $-1/(k-1)$  when no local dependence exists, where  $k$  is test length (Yen, 1993). Thus, the expected Q3 values should be approximately  $-0.02$  for the NSCAS tests (since most NSCAS tests had more than 50 operational items). Index values greater than 0.20 indicate a degree of local dependence that should be examined by test developers (Chen & Thissen, 1997).

Since the three residual correlations are very similar, the default "standardized residual correlation" in WINSTEPS was used for these analyses. Table 6.15 presents the summary statistics for all the residual correlations for each test, including the median, interquartile range (IQR), minimum, maximum, and several percentiles (10, 25, 50, 75, and 90). The table also presents the total number of item pairs and the number of pairs with the residual correlations greater than 0.2. The median residual correlations were slightly negative, and the values were close to 0.0. Most of the correlations were very small, suggesting local item independence generally holds for NSCAS ELA, Mathematics, and Science.

**Table 6.15. Local Independence: Summary of Item Residual Correlations**

Grade	#Item Pairs		Median	IQR	Min.	Percentiles					Max.
	Total	> 0.2				P10	P25	P50	P75	P90	
<b>ELA</b>											
3	8,256	0	0.00	0.01	-1.00	-0.02	-0.01	0.00	0.00	0.00	0.17
4	56,953	53	0.00	0.00	-1.00	-0.01	0.00	0.00	0.00	0.00	1.00
5	27,966	26	0.00	0.00	-1.00	-0.01	0.00	0.00	0.00	0.00	1.00
6	31,626	39	0.00	0.00	-1.00	-0.01	0.00	0.00	0.00	0.00	1.00
7	47,895	47	0.00	0.00	-1.00	-0.01	0.00	0.00	0.00	0.00	1.00
8	6,105	6	0.00	0.01	-0.44	-0.02	-0.01	0.00	0.00	0.01	1.00
<b>Mathematics</b>											
3	25,425	0	0.00	0.00	-0.22	-0.01	0.00	0.00	0.00	0.00	0.17
4	12,880	4	0.00	0.01	-0.17	-0.02	-0.01	0.00	0.00	0.01	0.26
5	21,115	9	0.00	0.01	-0.18	-0.02	-0.01	0.00	0.00	0.00	0.48
6	30,876	10	0.00	0.00	-0.18	-0.02	0.00	0.00	0.00	0.00	0.44
7	20,100	0	0.00	0.00	-0.17	-0.01	0.00	0.00	0.00	0.00	0.16
8	9,316	1	0.00	0.01	-0.14	-0.02	-0.01	0.00	0.00	0.00	0.32
<b>Science</b>											
5	1,225	0	-0.02	0.02	-0.09	-0.04	-0.03	-0.02	-0.01	0.00	0.13
8	1,770	0	-0.02	0.03	-0.10	-0.04	-0.03	-0.02	0.00	0.01	0.18

**6.4.1.3. Item Fit**

Item fit refers to how well the data fit the calibration model. Infit and outfit are two item fit statistics in WINSTEPS used to evaluate the degree to which the Rasch model predicts the observed item responses for the NSCAS tests. Each fit statistic can be expressed as a mean square (MNSQ) statistic with an expected value of 1.0 and a different variance for each mean square or as a standardized (ZSTD) statistic with an expected mean of 0.0 and an expected variance of 1.0. Table 6.16 presents the summary MNSQ statistics, and Table 6.17 presents the summary ZSTD statistics. Overall, these results show that the data fit the model well.

MNSQ values are more difficult to interpret due to an asymmetrical distribution and unique variance, while ZSTD values are more oriented toward standardized statistical significance. Though both are informative, the ZSTD values are less likely to be sensitive to the large sample sizes and have better distributional properties (Smith, Schumacker, & Bush, 1998). The outfit statistic tends to be affected more by unexpected responses far from the person, item, or rating scale category measure (i.e., it is more sensitive to outlying, off-target, and low information responses that are very informative regarding fit). The infit statistic tends to be affected more by unexpected responses close to the person, item, or rating scale category measure (i.e., with more information, but contributing little to the understanding of fit).

The expected MNSQ value is 1.0 and can range from 0 to positive infinity. Values greater than 1.0 can be interpreted as indicating the presence of noise or lack of fit between the responses and the measurement model. Values less than 1.0 can be interpreted as item consistency or overfitting (i.e., too predictable and/or too much redundancy). Rules of thumb regarding “practically significant” MNSQ values vary. More conservative users might prefer items with MNSQ values that range from 0.8 to 1.2, while others believe that reasonable test results can be achieved with values from 0.5 to 1.5. In Table 6.16, values outside of 0.7 to 1.3 are given



practical importance. Just like previous years, more outfit values fell outside of [0.7, 1.3], which is not surprising given that the infit statistic mutes the effects of anomalous responses by extreme students. Compared to 2018, the number of items outside of [0.7, 1.3] increased from approximately 1% to up to 11% in ELA and 3% in Mathematics. This is because items in 2018 that were included in the calibration were included in scoring, whereas items in 2019 that were not included in the calibration were also included in scoring. The inclusion of not-calibrated items in scoring in 2019 was because 2019 vertical scaling was conducted to verify the 2018 vertical scales. Therefore, to properly compare pre- and post-equated vertical scale results, it was necessary to use the same number of items between them.

The expected ZSTD mean value is 0.0 with an expected variance, or SD, of 1.0. It can effectively range from -9.99 to +9.99 in WINSTEPS. Values greater than 0.0 can be interpreted as indicating the presence of noise or lack of fit between the items and the model (underfitting). Values less than 0.0 can be interpreted as item redundancy or overfitting items (i.e., too predictable and/or too much redundancy). Rules of thumb regarding “practically significant” ZSTD values vary. More conservative users might prefer items with ZSTD values that range from -2.0 to +2.0, while others believe that reasonable test results can be achieved with values from -3.0 to +3.0 (DRC, 2017). In Table 6.17, values outside of -3.0 to +3.0 are given practical importance. The fact that all infit and outfit means were negative suggests that, on average, the data overfit the Rasch model (i.e., the data were a bit more consistent than expected by probabilistic model). Similar to the MNSQ statistics, more items fell outside of [-3.0, 3.0], which again is because the not-calibrated items were included in the 2019 scoring.

**Table 6.16. Item Fit: Summary of Infit and Outfit MNSQ Statistics for Items**

Grade	#Items	Infit						Outfit					
		Mean	SD	Min.	Max.	# [0.7,1.3]	% [0.7,1.3]	Mean	SD	Min.	Max.	# [0.7,1.3]	% [0.7,1.3]
<b>ELA</b>													
3	129	1.03	0.24	0.71	2.51	7	5.4	1.04	0.33	0.62	2.83	10	7.8
4	338	1.02	0.28	0.17	3.35	39	11.5	1.02	0.30	0.17	3.44	44	13.0
5	237	1.00	0.19	0.39	2.43	16	6.8	1.00	0.23	0.39	2.75	22	9.3
6	252	1.01	0.34	0.17	5.49	24	9.5	1.02	0.45	0.17	5.49	32	12.7
7	310	0.99	0.29	0.21	4.36	31	10.0	0.98	0.33	0.16	5.01	37	11.9
8	111	0.99	0.28	0.28	3.64	4	3.6	0.98	0.29	0.27	3.51	8	7.2
<b>Mathematics</b>													
3	226	0.96	0.15	0.59	3.00	5	2.2	0.96	0.17	0.41	2.87	7	3.1
4	161	0.98	0.10	0.75	1.51	3	1.9	0.98	0.15	0.67	1.82	11	6.8
5	206	0.98	0.26	0.22	4.24	7	3.4	0.98	0.27	0.22	4.24	11	5.3
6	249	0.97	0.12	0.11	1.74	7	2.8	0.97	0.17	0.07	2.10	15	6.0
7	201	0.96	0.07	0.67	1.28	1	0.5	0.97	0.10	0.67	1.55	4	2.0
8	137	0.97	0.07	0.84	1.36	1	0.7	0.99	0.32	0.65	4.54	3	2.2
<b>Science</b>													
5	50	0.99	0.08	0.82	1.19	0	0.0	0.99	0.15	0.69	1.28	2	4.0
8	60	1.00	0.09	0.84	1.24	0	0.0	0.99	0.16	0.63	1.36	4	6.7

**Table 6.17. Item Fit: Summary of Infit and Outfit ZSTD Statistics for Items**

Grade	#Items	Infit						Outfit					
		Mean	SD	Min.	Max.	# [-3.0,3.0]	% [-3.0,3.0]	Mean	SD	Min.	Max.	# [-3.0,3.0]	% [-3.0,3.0]
<b>ELA</b>													
3	129	-1.11	6.01	-9.90	9.90	75	58.1	-1.22	5.85	-9.90	9.90	79	61.2
4	338	-0.77	4.23	-9.90	9.90	135	39.9	-0.73	4.26	-9.90	9.90	130	38.5
5	237	-0.68	4.95	-9.90	9.90	89	37.6	-0.70	4.85	-9.90	9.90	85	35.9
6	252	-0.58	4.87	-9.90	9.90	113	44.8	-0.66	4.85	-9.90	9.90	111	44.0
7	310	-1.01	4.45	-9.90	9.90	109	35.2	-0.99	4.41	-9.90	9.90	113	36.5
8	111	-2.63	6.45	-9.90	9.90	76	68.5	-3.06	6.21	-9.90	9.90	75	67.6
<b>Mathematics</b>													
3	226	-4.60	4.58	-9.90	9.90	177	78.3	-4.51	4.49	-9.90	9.90	172	76.1
4	161	-3.76	5.86	-9.90	9.90	126	78.3	-3.45	5.98	-9.90	9.90	124	77.0
5	206	-5.01	5.84	-9.90	9.90	170	82.5	-4.80	5.80	-9.90	9.90	167	81.1
6	249	-4.29	5.52	-9.90	9.90	206	82.7	-4.25	5.40	-9.90	9.90	205	82.3
7	201	-2.70	4.67	-9.90	9.90	118	58.7	-2.50	4.65	-9.90	9.90	118	58.7
8	137	-4.13	4.91	-9.90	9.90	111	81.0	-3.80	5.17	-9.90	9.90	112	81.8
<b>Science</b>													
5	50	-0.58	7.78	-9.9	9.9	40	80.0	-0.45	7.85	-9.9	9.9	41	82.0
8	60	-1.04	7.51	-9.9	9.9	50	83.3	-1.12	8.14	-9.9	9.9	52	86.7

**6.4.2. Summary IRT Item Statistics**

Table 6.18 and Table 6.19 present the summary IRT item statistics across all operational and field test items, respectively. Appendix J presents the item-level IRT item statistics. Operational item parameter means increase by grade for ELA and Mathematics, as can be expected for vertical scales.

**Table 6.18. Summary IRT Item Statistics—Operational Items**

Grade	#Items	#Parameters	Mean	SD	Min.	Max.	Range (Max. – Min.)
<b>ELA</b>							
3	129	149	-0.391	0.897	-2.430	2.716	5.146
4	338	366	-0.480	1.125	-3.326	3.151	6.476
5	237	265	-0.034	1.206	-3.005	4.268	7.273
6	252	282	-0.218	1.115	-2.779	3.345	6.124
7	310	326	-0.091	1.001	-2.442	2.583	5.025
8	111	126	0.494	1.103	-2.050	3.180	5.230
<b>Mathematics</b>							
3	226	237	-0.968	1.138	-3.820	2.589	6.409
4	161	173	-0.009	1.228	-2.301	3.908	6.209
5	206	222	0.020	1.152	-2.822	3.695	6.517
6	249	272	0.439	1.251	-1.965	4.737	6.702
7	201	218	1.118	1.128	-1.390	4.545	5.935
8	137	145	1.095	1.116	-1.413	5.641	7.054
<b>Science</b>							
5	50	50	-0.691	0.912	-2.445	0.882	3.327
8	60	60	-0.629	0.891	-2.662	1.431	4.093

**Table 6.19. Summary IRT Item Statistics—Field Test Items**

Grade	#Items	#Parameters	Mean	SD	Min.	Max.	Range (Max. – Min.)
<b>ELA</b>							
3	180	220	0.069	1.064	-3.088	5.630	8.718
4	191	245	-0.085	1.180	-2.665	4.097	6.762
5	185	243	0.262	1.148	-2.490	3.686	6.176
6	192	251	0.641	1.063	-2.306	3.747	6.052
7	195	258	0.689	1.059	-2.162	4.134	6.296
8	184	243	0.794	1.118	-2.204	3.804	6.008
<b>Mathematics</b>							
3	231	258	-0.412	1.331	-3.404	6.297	9.701
4	231	259	0.436	1.128	-2.612	3.737	6.349
5	231	259	0.399	1.083	-2.439	3.220	5.659
6	231	259	1.061	1.350	-3.653	5.479	9.131
7	231	259	1.412	1.371	-2.005	5.285	7.290
8	231	259	1.704	1.360	-1.780	5.114	6.893

### 6.5. Vertical Scaling (ELA and Mathematics)

Vertical scales are constructed using multiple test levels (such as the grade level for the NSCAS tests), each of which is developed to be appropriate for students at a certain grade. A vertical scale score facilitates the estimation of an individual’s growth over time since it can describe student performance on the continuum for any levels of a test (Petersen, Kolen, & Hoover, 1989). In other words, vertical scales can permit the assessment of growth at the student level and provide the assessment of progress toward goals in subsequent grades on the same metric. When their use is appropriate and their construction is sound, vertical scales can provide a systematic way to examine the developmental characteristics and appropriateness of systems of state performance standards across grades (Patz, 2007).

Following the 2018 linking design, the NSCAS ELA and Mathematics vertical scales were created based on the following decisions. Please refer to Section 2.1 of this technical report for details on the linking design.

- Data collection design: Common item design
- Selection of the vertical scaling items: Above grade and below grade
- Scaling method: IRT Rasch and partial-credit models
- Calibration method: Concurrent calibration across grades
- Theta estimators and software: MLE in WINSTEPS
- Score transformation: The transformation constants determined in 2018

#### 6.5.1. Linking Item Selection

Unlike 2018 when linking items were selected as a subset of the 2018 paper-pencil forms, the 2019 anchors were selected separately. The purpose of having the horizontal linking items in the 2018 paper-pencil ELA and Mathematics forms was to calibrate paper-pencil forms, if needed. However, NDE’s effort of encouraging schools to take online tests worked, and less than 80 students took the PP test per content area and grade in 2018. Considering that calibration would not be possible with this small number of students, NWEA recommended not including a horizontal anchor item set in the PP form and instead using the 2019 pool

characteristics as statistical references in selecting the anchors. The TOS was used as the reference for anchor items so that the percentages were within 10% difference for each reporting category, which remained the same as in 2018. Anchor items were included as horizontal linking items and as vertical linking items in the lower/upper grades for the adaptive assessment (e.g., of the 28 horizontal anchors for ELA Grade 5, 14 of them were ELA Grade 4 vertical linking items and the other 14 were ELA Grade 6 vertical linking items).

### *6.5.2. Vertical Scaling Process*

NWEA performed four steps of calibration to verify the 2018 vertical scales for ELA and Mathematics:

1. Calibrate the HL and VL items across grades in a single calibration run where six grades were concurrently calibrated. All items across Grades 3–8 were placed on the same scale through vertical linking.
2. Equate VL and HL items from Step 1, employing the Robust Z method and mean-b (i.e. mean/mean) transformation.
3. Calibrate HL and operational items by grade while fixing the HL item parameter estimate from Step 2.
4. Run fully anchored WINSTEPS to score students using all items administered in 2019. For items calibrated in 2019, item parameters were fixed to those in Step 3. For items that were not calibrated due to low n-counts (i.e., less than 100 students), their item parameters were fixed to those in the bank.

NWEA followed the previous procedure of post-equating check for Science, employing the Robust Z statistic (Huynh, 2000; Huynh & Rawls, 2009; Huynh & Meyer, 2010) after unanchored calibration. The ELA and Mathematics results for operational items in this section were obtained from Step 2 and those for field test items were obtained from Step 2.

### *6.5.3. Vertical Scale Evaluation*

Vertical scaling in 2019 was scheduled to verify the 2018 vertical scales. Vertical linking items were placed on the assessment with the intent to assess the stability of the original 2018 vertical scale to show growth across grades (i.e., the vertical scale is common across grades), and horizontal linking items were placed on the assessment with the intent to equate back to the base scale for each grade from 2018. However, nonequivalent groups were administered vertical linking item sets from the initial test engine configuration. Therefore, NWEA did the following:

1. Construct the 2019 vertical scale without sampling following the same procedure for the 2018 vertical scale. All students were used, regardless of dissimilar student scores in the vertical linking item sets. The results were compared to the pre-equated results, applying the same comparison criteria used for the consistency check. The distribution of differences in student scores was also examined, and the test characteristic curves (TCCs) using all the 2019 calibrated items were compared.
2. Construct the 2019 vertical scales with sampling by selecting samples for each vertical linking item set. The 2019 vertical scales were then compared to the pre-equated results as described in Step 1.
3. Horizontally equate using the HL items.

Specifically, the following four cases were used to evaluate the original 2018 vertical scale. Case 1 used the 2018 item parameter estimates, whereas Cases 2–4 used new parameter estimates.

- Case 1: Pre-equating
- Case 2: Post-equating without sampling
- Case 3: Post-equating with sampling
- Case 4: Horizontal equating

The following analyses were conducted to evaluate the vertical scale for Cases 1–4:

- TCCs
- Conditional standard error of measurement (CSEM) curves
- Grade-to-grade growth
- Grade-to-grade variability
- Separation of grade distribution

Table 6.20 – Table 6.23 and Figure 6.1 – Figure 6.5 summarize the results of the vertical scale evaluation.<sup>7</sup> Based on the results, NWEA recommended the following for 2019 scoring, which were approved by NDE:

- Case 1: Pre-equating for ELA Grades 3–8 and Mathematics Grades 3–6
- Case 4: Horizontal linking for Mathematics Grades 7 and 8

**Table 6.20. Scale Score Difference Between 2018 and 2019 Final Recommendations**

Grade	2019 Scale Score			2018 Scale Score			Mean Difference (2019–2018)	Effect Size
	N	Mean	SD	N	Mean	SD		
<b>ELA</b>								
3	23,410	2485.99	72.21	23,769	2481.31	76.47	4.68	0.06
4	23,935	2513.99	73.41	23,783	2511.56	71.98	2.43	0.03
5	23,939	2525.88	69.57	22,198	2531.34	66.88	-5.46	-0.08
6	22,352	2538.20	68.16	23,231	2538.47	66.68	-0.27	0.00
7	23,397	2545.11	72.79	22,870	2550.49	73.79	-5.38	-0.07
8	23,054	2558.20	68.27	23,165	2560.89	66.26	-2.69	-0.04
<b>Mathematics</b>								
3	23,369	1194.90	71.81	23,740	1192.17	71.13	2.73	0.04
4	23,892	1224.98	66.89	23,734	1227.00	66.92	-2.02	-0.03
5	23,889	1244.31	70.59	22,154	1241.60	66.10	2.71	0.04
6	22,302	1252.91	68.16	23,189	1253.86	72.33	-0.95	-0.01
7	23,384	1252.22	64.33	22,806	1254.73	67.33	-2.51	-0.04
8	23,023	1267.75	71.35	23,096	1270.73	71.32	-2.98	-0.04

<sup>7</sup> For full results for Cases 1–4, please refer to the 2019 vertical scale evaluation report (NWEA, 2019e).

**Table 6.21. Achievement Level Distributions—2019, 2018, and %Difference**

Grade	2019*					2018*					%Difference*			
	N	%Dev	%OT	%CCR	%OT+ %CCR	N	%Dev	%OT	%CCR	%OT+ %CCR	%Dev	%OT	%CCR	%OT+ %CCR
<b>ELA</b>														
3	23,410	43.6	39	17.4	56.4	23,769	46.6	37.4	16.0	53.4	-3.0	1.6	1.4	3.0
4	23,935	41.7	39.5	18.8	58.3	23,783	43.4	40.5	16.1	56.6	-1.7	-1.0	2.7	1.7
5	23,939	51.8	32.9	15.3	48.2	22,198	48.5	35.4	16.1	51.5	3.3	-2.5	-0.8	-3.3
6	22,352	50.7	31.6	17.7	49.3	23,231	52.4	30.4	17.2	47.6	-1.7	1.2	0.5	1.7
7	23,397	51.1	38.1	10.8	48.9	22,870	52.4	32.7	14.9	47.6	-1.3	5.4	-4.1	1.3
8	23,054	49.5	36.1	14.5	50.6	23,165	48.9	37.2	14.0	51.1	0.6	-1.1	0.5	-0.6
<b>Mathematics</b>														
3	23,369	44.9	45.4	9.8	55.2	23,740	50.1	39.6	10.3	49.9	-5.2	5.8	-0.5	5.2
4	23,892	48.3	43.6	8.1	51.7	23,734	50.1	39.4	10.4	49.9	-1.8	4.2	-2.3	1.8
5	23,889	45.7	43.5	10.7	54.2	22,154	49.4	41.2	9.5	50.6	-3.7	2.3	1.2	3.7
6	22,302	44.8	45.4	9.8	55.2	23,189	45.1	44.6	10.3	54.9	-0.3	0.8	-0.5	0.3
7	23,384	51.0	40.4	8.6	49.0	22,806	50.5	39.3	10.2	49.5	0.5	1.1	-1.6	-0.5
8	23,023	52.5	37.1	10.4	47.5	23,096	49.4	41.1	9.5	50.6	3.1	-4.0	0.9	-3.1

\*Dev = Developing. OT = On Track. CCR = College and Career Readiness.

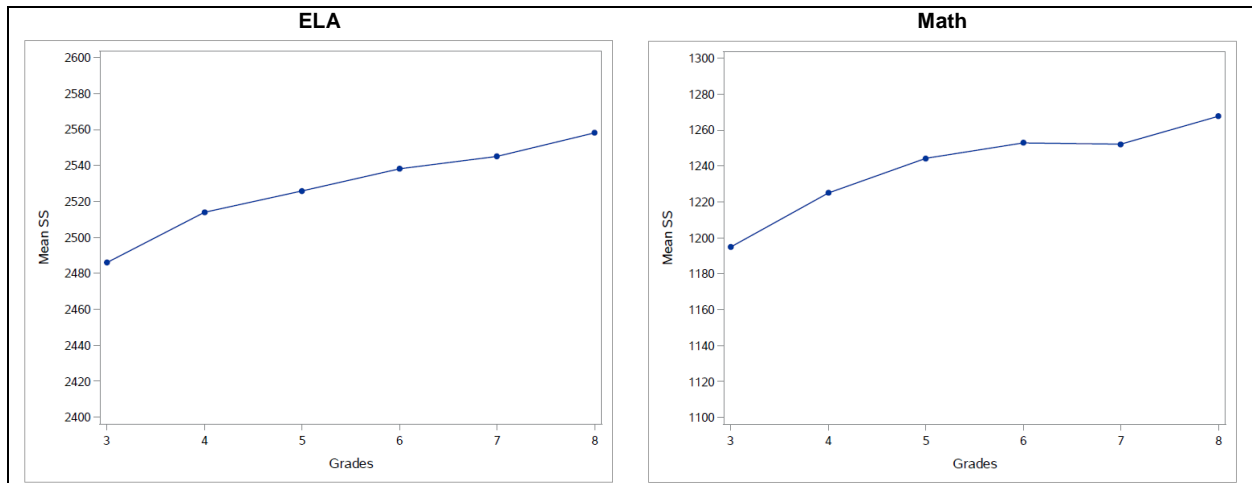
**Table 6.22. Scale Scores by Percentile Rank**

Grade	Case	Percentile Rank									
		P5	P15	P25	P35	P45	P55	P65	P75	P85	P95
<b>ELA</b>											
3	1	2365	2404	2431	2457	2480	2500	2518	2538	2562	2601
4	1	2391	2435	2462	2486	2507	2527	2546	2567	2591	2630
5	1	2411	2450	2477	2500	2519	2536	2554	2575	2599	2636
6	1	2421	2467	2493	2514	2532	2550	2569	2587	2609	2644
7	1	2415	2464	2497	2523	2543	2562	2580	2599	2619	2653
8	1	2435	2483	2512	2533	2553	2571	2588	2606	2631	2664
<b>Mathematics</b>											
3	1	1074	1117	1147	1170	1190	1207	1224	1244	1269	1310
4	1	1115	1154	1179	1198	1216	1232	1251	1270	1294	1335
5	1	1134	1172	1196	1215	1234	1250	1268	1286	1314	1363
6	1	1143	1182	1206	1226	1244	1260	1279	1299	1324	1368
7	4	1159	1188	1206	1223	1238	1252	1270	1291	1319	1372
8	4	1165	1195	1216	1233	1250	1269	1289	1313	1346	1396

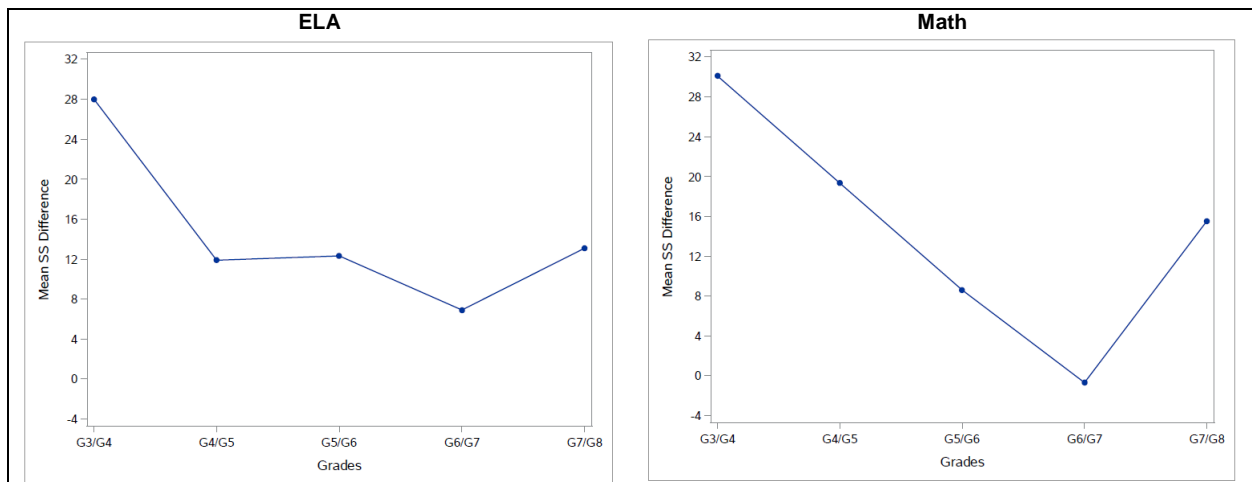
**Table 6.23. Effect Size and Horizontal Distance**

Grades	Effect Size	Percentile Rank						
		P5	P10	P25	P50	P75	P90	P95
<b>ELA</b>								
3/4	0.38	26	27	31	26	29	30	29
4/5	0.17	20	18	15	11	8	8	6
5/6	0.18	10	16	16	13	12	8	8
6/7	0.10	-6	-6	4	12	12	9	9
7/8	0.19	20	24	15	9	7	13	11
<b>Mathematics</b>								
3/4	0.43	41	39	32	27	26	27	25
4/5	0.28	19	20	17	17	16	25	28
5/6	0.12	9	9	10	10	13	5	5
6/7	-0.01	16	9	0	-7	-8	-2	4
7/8	0.23	6	6	10	14	22	30	24
3/4	0.43	41	39	32	27	26	27	25

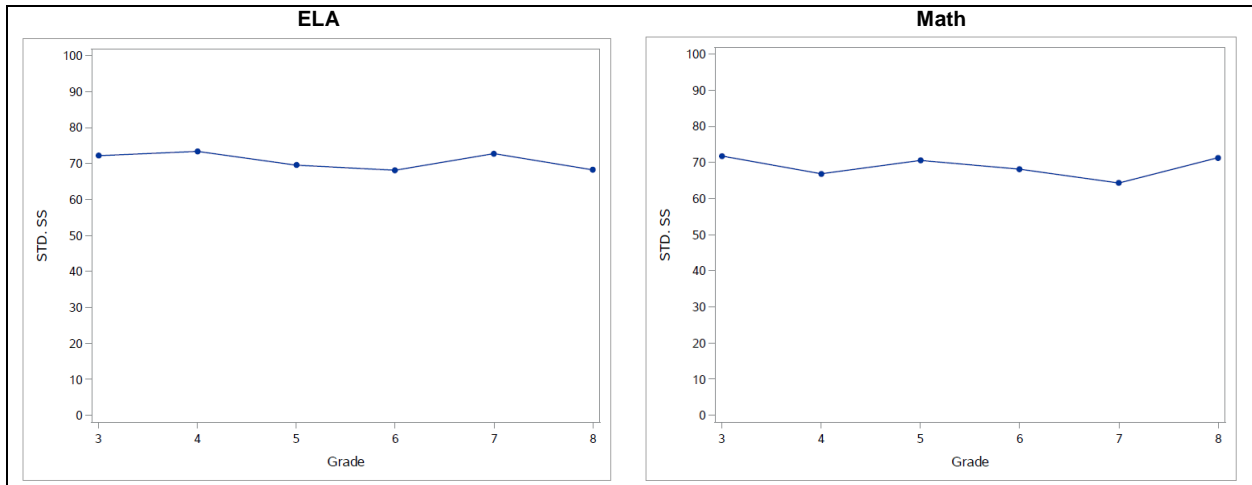
**Figure 6.1. Mean Scale Score by Grade**



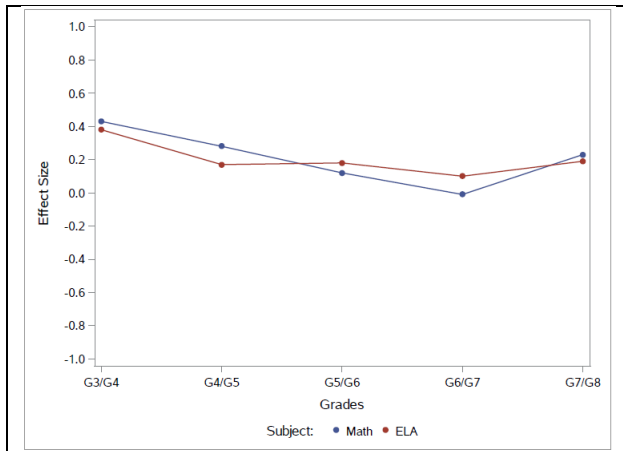
**Figure 6.2. Mean Scale Score Differences Between Adjacent Grades**



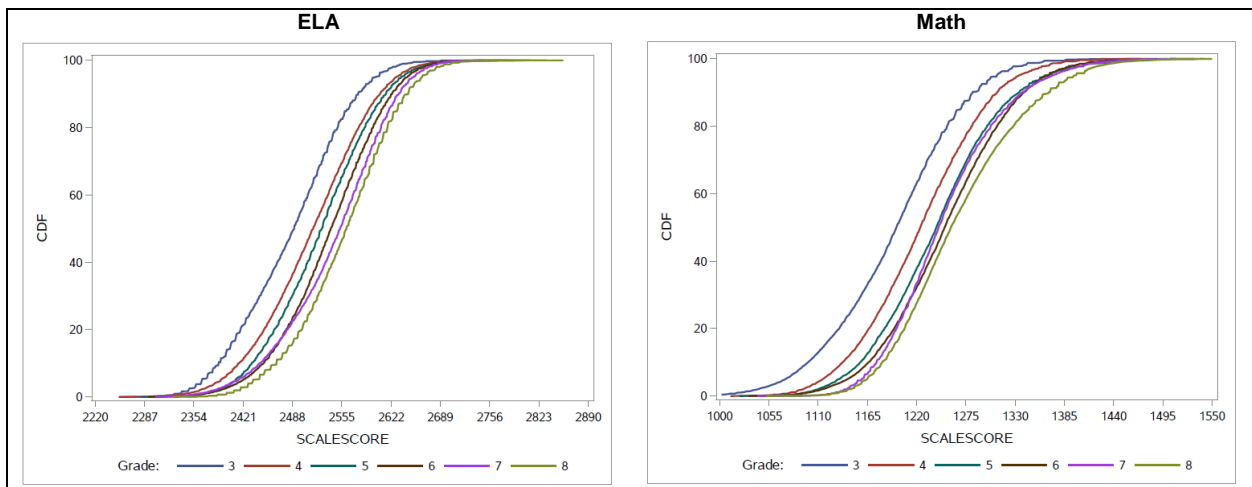
**Figure 6.3. SD of Scale Score by Grade**



**Figure 6.4. Effect Sizes between Adjacent Grades**



**Figure 6.5. Cumulative Distribution Function (CDF)**





#### 6.5.4. 2020 Scaling Considerations

To ensure the quality of vertical scales, the same scaling process will be applied in 2020 to verify them. Both vertical and horizontal linking items will be included and new sets of linking items will be selected and administered for 2020, but the overall design and process will be similar to 2018 and 2019. The results from 2020 vertical scaling will be compared to that from 2020 pre-equating so that the vertical scales can be verified. Further, operational items with significant item parameter estimate changes using the Robust Z method will not be included for the 2020 item pool. A total of 95 items were flagged using the Robust Z statistics of  $\pm 1.645$  critical value, as reported in Table 6.24. Some of these items may remain in the 2020 pool to meet the TOS, including items to be used for the breach forms. Overall, compared to 2018 when a total of 234 items were moved from the 2019 pool, much less items were flagged, which indicates that the pool, as a whole, is getting more stable.

**Table 6.24. Number of Items with a Large Parameter Change**

Grade	#Items		
	ELA	Mathematics	Total
3	8	1	9
4	11	–	11
5	3	3	6
6	23	–	23
7	46	–	46
8	–	–	–
<b>Total</b>	<b>91</b>	<b>4</b>	<b>95</b>

#### 6.6. Post-Equating Check (Science)

NWEA followed the previous procedure of post-equating check for Science Grades 5 and 8 employing the Robust Z statistic (Huynh, 2000; Huynh & Rawls, 2009; Huynh & Meyer, 2010) after unanchored calibration.

##### 6.6.1. Post-Equating Method

Because the 2019 Science forms were created from previously operational items, all the 2019 operational items were used as the linking set. This means that the raw-to-scale score conversion tables were established prior to the operational administration. This is referred to as pre-equating. However, it may not be appropriate to assume that the operational items maintained their relative difficulty across administrations. The same item can perform differently across administrations due to changes in the item's position or changes in the students' experiences. Further, in Spring 2019 before the test administration, Science Grade 5 had key changes for seven items that were changed from D to the other options, as there were 23 items with key D (i.e., almost half of the 50-item test). Therefore, when the 2019 operational test data became available, the item difficulty equivalence was checked using the Robust Z post-equating check procedure to identify items that show significant difficulty changes from the bank values, paying special attention to the seven items with key changes. If no unstable items are identified, the 2019 equating process would result in the pre-equating solution, whereas a post-equating solution would be used if items are found to be outside the normal estimation error.

The subset of 2019 operational items (with any items identified for large difficulty changes excluded) was used as the set to estimate the link constant to map the 2019 test to the bank scale. This equating process is known as post-equating, and the raw-to-scale score conversion is generated based on the operational test data. As part of the post-equating check procedures, the item difficulty equivalence was checked by comparing the old banked item calibration (i.e., pre-calibration) with a new unanchored calibration of the 2019 data (i.e., post-calibration) using WINSTEPS 3.91.0.0 (Linacre, 2015). The evaluations were conducted for each grade using the Robust Z statistic (Huynh & Meyer, 2010). This method focuses on the correlations between the pre- and post-calibrated item difficulties and the ratio of standard deviations (RSD) between the two calibrations, just like in previous years. The correlation between the two item difficulty estimates should be 0.95 or higher, and the RSD between the two sets of item difficulty estimates should range between 0.90 and 1.10 (Huynh & Meyer, 2010). To detect inconsistent item difficulty estimates, a critical value for the Robust Z statistic of  $\pm 1.645$  was used. Items that exceeded the Robust Z critical value were deleted, one item at a time, until both the item difficulty correlation and SD ratio fell within the prescribed limits.

### 6.6.2. Post-Equating Results

Table 6.25 presents the 2019 Science correlation statistics and SD ratio following the process described above. Table 6.26 presents the percentage of students at each achievement level for both 2018 and 2019. The percentage of students at Below the Standards increases slightly by approximately 0.6% and 3.8% for Grades 5 and 8, respectively.

**Table 6.25. Science Pre- and Post-Equating Comparison**

Grade	Iteration	SD Pre	SD Post	RSD	Correlation
5	1	0.96	0.93	1.03	0.939
5	2, excluded 41145680	0.94	0.94	1.00	0.946
5	3, excluded 41145730	0.87	0.91	0.95	0.950
5	4, excluded 41143970	0.87	0.89	0.97	0.951
8	1	0.89	0.90	0.99	0.964

**Table 6.26. Science Achievement Level Distribution for 2018 and 2019**

Grade	2018*				2019*			
	N	%Below	%Meets	%Exceeds	N	%Below	%Meets	%Exceeds
5	22,136	30.1	54.4	15.5	23,897	30.7	55.3	13.9
8	23,043	32.9	47.6	19.5	23,019	36.7	50.1	13.2

\*The 2018 percentages are from the 2018 technical report, which includes all students. The 2019 percentages are from the 2019 psychometric analyses data, which includes students taking online tests who attempted 10 or more operational items.

Table K.1, found in Appendix K, presents the item parameter estimates for Grades 5 and 8 when all items were used. The item difficulty correlation is 0.939 for Grade 5 Science when all items were used, which did not meet the Robust Z criteria. Consequently, Grade 5 items with the highest absolute Robust Z statistic were excluded one item at a time (Item 41145680 first, followed by Items 41145730 and 41143970). With these three items excluded, the correlation is 0.951 and the RSD is 0.97, which met both criteria. None of the seven Grade 5 items with key changes got flagged. The item difficulty correlation is 0.963 and the RSD is 0.99 for Grade 8 Science when all items were used, which met both criteria. Considering these results, post-equating was conducted for Grade 5 using the updated item parameter estimates for the three excluded items, and pre-equating was conducted for Grade 8.

Table K.2 compares the pre- and post-equated scoring tables for Grade 5 with student frequency. As shown in the “SS Diff (Pre – Post)” and “AL Diff” columns, scale scores differ up to three points, but achievement Levels 1 and 2 differ by one point (i.e. Level 2 starts from the raw score of 28 and 27 in pre- and post-equating, respectively, and Level 1 starts from the raw score of 42 and 41 in pre- and post-equating, respectively). Table K.3 presents the pre-equated scoring tables for Grade 8 with student frequency.

### 6.7. Scaling

The previously set scaling constants for Science were used again in 2019. For ELA and Mathematics, scaling constants were set in 2018 without anchoring cut scores so that scale scores could be presented at the standard setting and cut score review meetings, as well as the Nebraska State Board of Education meeting on August 2, 2018. After constructing the vertical scales for ELA and Mathematics, descriptive statistics of student scale scores were examined to determine the following scaling constants of slope and intercept:

- A slope of  $66.6/\sigma_{G5}$  (i.e., slope= $72.47244$ ) and intercept of 2500 for ELA
- A slope of  $66.6/\sigma_{G5}$  (i.e., slope= $54.92622$ ) and intercept of 1200 for Mathematics

where  $\sigma_{G5}$  is the standard deviation of the Grade 5 theta score.

The theta estimate,  $\theta$ , and associated  $\theta$ -CSEM of students were then expressed on the NSCAS reporting scale by applying the linear transformation, slope and intercept (A and B, respectively), as follows:

$$\begin{aligned} SS &= (\theta \times A) + B \\ SSCSEM &= \theta\text{-CSEM} \times A. \end{aligned} \tag{6.10}$$

$\theta$ -CSEM are defined as the reciprocal of the square root of the test information function and can be estimated across all points of the ability continuum (Hambleton & Swaminathan, 1985):

$$\theta\text{-CSEM} = \text{CSEM}(\theta_i) = \frac{1}{\sqrt{I(\theta_i)}} \tag{6.11}$$

where  $I(\theta)$  is the test information function, as a sum of item information function, obtained as:

$$I(\theta_i) = \sum_j \frac{p'_{ij}(\theta_i)^2}{p_{ij}(\theta_i)q_{ij}(\theta_i)} \tag{6.12}$$

where  $p'_{ij}(\theta_i)$  is the derivative of  $p_{ij}(\theta_i)$  and  $q_{ij}(\theta_i) = 1 - p_{ij}(\theta_i)$ . Once the linear transformation was applied, the scaled scores and associated CSEMs were rounded to an integer value. There was no adjustment made around cut scores or the scale score CSEM (SSCSEM). Final adjustments were made to scale scores that fell outside of the HOSS or the LOSS.

In setting the HOSS for ELA and Mathematics, the following guidelines were considered. In setting the LOSS, similar guidelines were considered.

1. The HOSS must increase as the grade increases for tests on a vertical scale.
2. The HOSS should be high enough that it does not cause an unnecessary "pile-up" of scale scores at the HOSS, targeting less than 1%.
3. The HOSS should be low enough that  $SSCSEM(HOSS) < 10 \times \text{Min}(SSCSEM)$ .
4. The HOSS may be high enough that  $SSCSEM(\text{Penultimate HOSS}) < 5 \times \text{Min}(SSCSEM)$ .
5. The HOSS gap should not be too small, as a future test form may be slightly more difficult. It is also important that the gap is not too large, as that will tend to impact the mean of the distribution for cases with many perfect scores.
6. The gaps should change smoothly over score points, and the HOSS gap should transition smoothly across grades. It is more difficult, and less important, to keep the gaps smooth over score points and grades than it is to keep the SSCSEM values smooth over score points and SSCSEM (HOSS) transitions smooth across grade levels.

Based on these guidelines, the LOSS and HOSS presented in Table 6.27 were used. To be consistent with ELA and Mathematics with score ranges, the LOSS of Science was changed from 1 to 0. This did not change actual scores in that a score of 0 were assigned to students who attempted 0 items and a score of 1 were assigned to students who attempted 1–9 operational items. However, this change did make the communication consistent: The LOSS of each grade was used for students with 0 items attempted, the score of one point higher than LOSS were used for students with 1–9 operational items attempted, and the score of two points higher than LOSS were used for students with 10 or more operational items attempted.

**Table 6.27. Score Range (LOSS and HOSS) and Assigned Score**

Grade	LOSS	HOSS	Assigned score for students with 0 OP items attempted	Assigned score for students with 1–9 OP items attempted	Lowest calculated score for students with 10 or more OP items attempted
<b>ELA</b>					
3	2220	2840	2220	2221	2222
4	2250	2850	2250	2251	2252
5	2280	2860	2280	2281	2282
6	2290	2870	2290	2291	2292
7	2300	2880	2300	2301	2302
8	2310	2890	2310	2311	2312
<b>Mathematics</b>					
3	1000	1470	1000	1001	1002
4	1010	1500	1010	1011	1012
5	1020	1510	1020	1021	1022
6	1030	1530	1030	1031	1032
7	1040	1540	1040	1041	1042
8	1050	1550	1050	1051	1052
<b>Science</b>					
5	0	200	0	1	2
8	0	200	0	1	2

Table 6.28 summarizes the cut score implementation, or the conversions of student ability (theta) to scale scores that were used for scoring. Specifically, the table presents the calculations of the slopes and intercepts for all grades of the scale score conversions, including the cut scores set during standard setting.

**Table 6.28. Conversion of Theta to Scale Scores**

Grade	Scale Score Ranges			Cut Scores		Conversion		Cuts (Theta)*	
	Developing	On Track	CCR	On Track	CCR	Slope <i>b</i>	Intercept <i>a</i>	On Track	CCR
<b>ELA</b>									
3	2220–2476	2477–2556	2557–2840	2477	2557	72.47244	2500	-0.3193	0.7867
4	2250–2499	2500–2581	2582–2850	2500	2582	72.47244	2500	-0.0024	1.1291
5	2280–2530	2531–2598	2599–2860	2531	2599	72.47244	2500	0.4309	1.3599
6	2290–2542	2543–2602	2603–2870	2543	2603	72.47244	2500	0.5970	1.4212
7	2300–2555	2556–2629	2630–2880	2556	2630	72.47244	2500	0.7741	1.7938
8	2310–2560	2561–2631	2632–2890	2561	2632	72.47244	2500	0.8389	1.8146
<b>Mathematics</b>									
3	1000–1189	1190–1285	1286–1470	1190	1286	54.92622	1200	-0.1821	1.5657
4	1010–1221	1222–1316	1317–1500	1222	1317	54.92622	1200	0.4005	2.1301
5	1020–1235	1236–1330	1331–1510	1236	1331	54.92622	1200	0.6554	2.3850
6	1030–1243	1244–1341	1342–1530	1244	1342	54.92622	1200	0.8011	2.5853
7	1040–1246	1247–1345	1346–1540	1247	1346	54.92622	1200	0.8557	2.6581
8	1050–1263	1264–1364	1365–1550	1264	1365	54.92622	1200	1.1652	3.0040
<b>Science</b>									
5	0–84	85–134	135–200	85	135	32.15095	100.49331	-0.4971	1.0580
8	0–84	85–134	135–200	85	135	33.50958	99.73252	-0.4543	1.0378

\*For ELA, theta cuts are based on equipercentile linking, as reported in “2018 NSCAS Vertical Scale Evaluation Report 2018-07-02.docx,” except for the Grade 7 CCR cut that was adjusted from 2632 to 2630 to be vertically aligned with Grade 8. For Mathematics, theta cuts were calculated using scale score cuts, slope, and intercept for each grade.

## Section 7: Standard Setting

No standard setting was held in 2018–2019. Nebraska’s statewide assessment system for ELA and Mathematics underwent significant changes between the 2016 and 2017 administrations, so cut scores for ELA and Mathematics were set following the Spring 2018 administration at standard setting and cut score review meetings from July 26–28, 2018, using the Item-Descriptor (ID) Matching method to delineate the Developing, On Track, and CCR Benchmark achievement levels. The purpose of the standard setting was to set new cut scores for Mathematics, whereas the purpose of the cut score review was to validate the existing cut scores for ELA. This section summarizes the process and results from those meetings. For more in-depth information, please refer to the full standard setting and cut score review reports (EdMetric, 2018a, 2018b). No changes were made to the Science standards or assessments, so a standard setting was not necessary.

### 7.1. Overview

In 2016–2017, the NSCAS ELA assessments underwent a shift in focus from basic proficiency to alignment with Nebraska’s College and Career Ready Standards for ELA to create a logical coherence in the transition from the grade-level assessments to the ACT assessment for high school students. Concurrent with the change in focus for the 2017 administration, NDE conducted a series of standard setting events for the NSCAS ELA Grades 3–8 assessments and the Nebraska administration of the ACT in Summer 2017. These events began with a Nebraska-specific ACT standard setting, followed by a Grade 8 NSCAS ELA standard setting, and, finally, a NSCAS ELA Grades 3–7 standard setting. This sequencing allowed the Nebraska ACT performance standards to inform development of the NSCAS ELA Grade 8 standards and the NSCAS ELA Grade 8 standards, in turn, to inform the development of the NSCAS ELA Grades 3–7 standards. The intended result was coherence across the entire system, from Grade 3 to high school.

NDE examined the percent of students achieving proficiency based on the 2017 cut scores for the NSCAS and ACT ELA assessments and confirmed that the cut scores did reflect coherence across the grade levels. NDE framed the release of the 2017 scores to stakeholders with the expectation that the percent of students meeting the CCR Benchmark would increase as educators and schools had opportunities to align curriculum, instructional materials, and instructional strategies to the College and Career Ready Standards and to adjust to the paradigm shift away from “basic proficiency” to college and career readiness. Because new ELA standards had already been set in 2017 and the updates to the test reflected a change in test structure, rather than a change in the constructs being measured, NDE conducted a review of the cut scores in 2018 to ensure that they were still appropriate.

The development and update schedule for the NSCAS Mathematics assessments is one administration cycle after that of the ELA assessments. Therefore, concurrently with the ELA cut score review, NDE conducted a full standard setting for the NSCAS Mathematics assessments. NDE’s intention was to maintain system-level coherence by using the ACT CCR Benchmark as a reference point for the Mathematics standard setting. Beginning with the Mathematics CCR Benchmark cut scores established during the Nebraska-specific ACT standard setting, preliminary cut scores were extrapolated for each grade level. These cut scores were then used to create a range within which panelists could determine their recommended cut scores for each grade and achievement level.

To ensure that the NSCAS standard setting and cut score review meetings were completed with fidelity to the intended processes and with the necessary technical expertise, NWEA subcontracted with EdMetric, an industry leader in standard setting. EdMetric facilitated and trained panelists and table leaders in the process of examining test items and content to recommend the cut scores, whereas NDE provided policy guidance and historical perspective, NWEA provided resources and content expertise, and Nebraska educators participated actively as panelists and table leaders. Specifically, 67 panelists participated in the Mathematics standard setting and 62 panelists participated in the ELA cut score review, representing 44 Nebraska school districts.

## **7.2. ID Matching Method**

The *Standards* (AERA et al., 2014) emphasize the selection of a standard setting methodology that is appropriate for the assessment being administered. Based on the technical characteristics of the NSCAS ELA and Mathematics assessments and their intended uses, NWEA and EdMetric, with the input of NDE's TAC, determined that the ID Matching method would be most appropriate for the standard setting and cut score review. The ID Matching method brings together diverse panels of experts (typically a wide representation of classroom educators) who complete a deep study of the content of the items and content standards to which they are aligned to determine recommended scale score cut points that fall between each achievement level. ID Matching is particularly appropriate for assessments that are scaled using IRT and assessments that include multiple item types because panelists consider the content of items that are presented in ascending order of difficulty based on IRT item statistics derived from actual student performance. Panelists match item demands to those described in the ALDs.

## **7.3. Meeting Process**

The meetings included an overview of the NSCAS and meeting goals, training, ID Matching training, multiple rounds of judgments, ALD revision, and vertical articulation. Mathematics and ELA panelists participated in a joint opening session before moving to content-specific workshop activities. A small group of panelists then participated in vertical articulation once the cut scores were set to finalize the recommended cut scores. Specifically, Mathematics panelists completed the following activities during the multiple rounds of judgments:

- Round 1: Panelists experienced the adaptive student assessment, studied the ALDs and OIB, completed the item matching activity, and recommended cut scores.
- Round 2: Panelists reviewed the dispersion of their Round 1 recommendations, reviewed benchmark cut score ranges, and revisited their cut scores.
- Round 3: Panelists reviewed impact data, discussed their Round 2 recommendations, and revisited their cut scores.
- Round 4: Panelists reviewed impact data, discussed their Round 3 recommendations, and recommended final cut scores.
- Vertical Articulation: In a cross-grade activity, a small group of panelists examined the system of cut scores and impact data to ensure coherence across the grades.

ELA panelists completed the following activities during the multiple rounds of judgments:

- Round 1: Panelists experienced the adaptive student assessment, studied the ALDs and OIB, studied the placement of the 2017 cut scores, and recommended cut scores.

- Round 2: Panelists reviewed impact data, discussed their Round 1 recommendations, and recommended final cut scores.
- Vertical Articulation: In a cross-grade activity, a small group of panelists examined the system of cut scores and impact data to ensure coherence across the grades.

#### 7.4. ALD Revision

The ID Matching method requires clear ALDs that describe the KSAs of a student at a particular achievement level. Using those ALDs to identify a cut score ensures alignment of the assessment system and allows educators to focus on the ALDs during instructional adaptations to effect change in student learning and performance. Draft ELA and Mathematics Range ALDs were brought to the standard setting and cut score meetings to be reviewed and refined by educators who were trained on the tenets of the Range ALD process by an expert in the development of ALDs. The training and presenter were the same as was given to the original set of teachers who reviewed the Mathematics ALDs during their original development process. While the training given to participants was the same regarding the framework of ALD constructional principals, the work participants engaged in to develop the Reporting ALDs differed. The final Range ALDs, after being finalized and approved by NDE, are provided in the standard setting and cut score review reports (EdMetric, 2018a, 2018b), as well as posted online on NDE’s website.

Specifically for ELA, participants used items in the OIBs to support the development of Range ALDs for each indicator by contrasting items from the same indicator that were in different achievement levels. Participants in each grade were divided into four groups: (a) Reading Vocabulary, (b) Reading Comprehension, (c) Writing Process, and (d) Writing Modes. When each group finished an initial draft, another table reviewed and suggested edits for the draft. By the end of the workshop, working drafts of ALDs for all ELA indicators were completed. For Mathematics, participants identified items in the OIB that they felt had not matched the ALDs during the standard setting process. Participants were trained that the order in the OIB showed how difficult items were for students. Using the content-recommended cut scores, participants could study the items that were inconsistent with the ALDs and suggest edits to the ALDs. The grade-level groups began this task at their own pace. NWEA reviewed the participants’ recommendations as the ALDs were finalized along with the items in the OIB.

#### 7.5. Final Results

The recommended cut scores were presented to the Nebraska State Board of Education on August 2, 2018. Table 7.1 presents the final approved cut scores that were used for subsequent scoring. The table also presents the accompanying impact data, or the percent of students in each achievement level based on the cut scores, that are based on the standard setting data.

**Table 7.1. Final Approved Cut Scores and Impact Data—ELA and Mathematics**

Content Area	Grade	Cut Scores		Impact Data			
		On Track	CCR	Developing	On Track	CCR	On Track + CCR
ELA	3	2477	2557	46.7	37.3	15.9	53.2
	4	2500	2582	43.4	40.5	16.1	56.6
	5	2531	2599	48.6	35.3	16.1	51.4
	6	2543	2603	52.4	30.4	17.2	47.6
	7	2556	2630	52.4	32.7	14.9	47.6
	8	2561	2632	49.0	37.1	13.9	51.0



Content Area	Grade	Cut Scores		Impact Data			
		On Track	CCR	Developing	On Track	CCR	On Track + CCR
Mathematics	3	1190	1286	50.2	39.5	10.3	49.8
	4	1222	1317	50.2	39.4	10.4	49.8
	5	1236	1331	49.5	41.1	9.4	50.5
	6	1244	1342	45.2	44.6	10.3	54.9
	7	1247	1346	50.6	39.2	10.2	49.4
	8	1264	1365	49.4	41.1	9.5	50.6

## Section 8: Test Results

All students who took the online, paper-pencil, and Spanish forms of the 2019 NSCAS Summative assessments were included in the test results. For results based on demographics and accommodations, all participants (i.e., student who attempted at least one item) were included. For all other results in this section, students who attempted at least 10 operational items on the online and paper-pencil forms were used (i.e., Spanish test-takers were not included). Results presented in this section are not from the state student file that NDE received and may therefore differ slightly from the official state summary report due to ongoing resolution of test materials and slight differences in the application of exclusion rules.

### 8.1. Demographics and Accommodations

Table 8.1 – Table 8.6 present the number of tested students by demographics for each grade and content area, including gender, ethnicity, free and reduced lunch (FRL) status, limited English proficiency (LEP) status, special education (SPED) status, use of universal features (i.e., answer eliminator, highlighter, notepad, and zoom), and use of accommodations (text-to-speech (TTS), paper-pencil form, Spanish online or paper-pencil form, Braille, and large print). Starting in 2018, both current and former English language learner (ELL) students are considered to have LEP status, resulting in more LEP students compared to previous years.

As shown in these tables, more than 22,000 students took the assessment in each grade and content area. Of those students across grades, half are males, half are females, two thirds are white, and about one fifth are Hispanic. Among the students across grades, about 45% to 47% are eligible for FRL, 14–16% have LEP status, and 13–16% belong to at least one SPED category. For all three of these programs/categories, the participation rate is slightly lower for upper-grade students. In terms of the test accommodations, the calculator is used by most students (80% or higher for Grades 6–8 in Mathematics). In general, the answer choice eliminator was the most-used tool and TTS was the least-used tool across all grades and content areas. These percentages are very similar to last year.

**Table 8.1. Number of Students Tested by Demographics—Grade 3**

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
<b>Total N-Count</b>		<b>23,475</b>	<b>100.00</b>	<b>23,446</b>	<b>100.00</b>
Gender	Female	11,358	48.38	11,348	48.40
	Male	12,117	51.62	12,098	51.60
Ethnicity	AI/AN	287	1.22	289	1.23
	Asian	675	2.88	675	2.88
	Black or African American	1,557	6.63	1,556	6.64
	Hispanic	4,567	19.45	4,562	19.46
	NH/PI	31	0.13	31	0.13
	White	15,315	65.24	15,289	65.21
	Two or More Races	1,043	4.44	1,044	4.45
FRL	Yes	11,037	47.02	11,029	47.04
	No	12,438	52.98	12,417	52.96
LEP	Yes	3,663	15.6	3,663	15.62
	No	19,812	84.4	19,783	84.38

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
SPED	Yes	3,758	16.01	3,743	15.96
	No	19,717	83.99	19,703	84.04
Universal Features & Accommodations	Answer Choice Eliminator	13,466	57.36	13,868	59.15
	Highlighter	11,341	48.31	7,564	32.26
	Line Reader	14,178	60.40	6,277	26.77
	Notepad	9,299	39.61	8,802	37.54
	Text-to-Speech (TTS)	3,944	16.80	3,714	15.84
	Zoom	9,100	38.76	5,079	21.66
	Paper-Pencil (PP)	18	0.08	16	0.07
	Spanish Online	41	0.17	54	0.23
	Spanish Paper-Pencil (PP)	2	0.01	3	0.01
	Braille**	2	–	2	–
	Large Print**	6	–	6	–

\*AI/AN = American Indian or Alaskan Native. NH/PI = Native Hawaiian or Other Pacific Islander. FRL = free and reduced lunch. LEP = limited English proficient. SPED = special education.

\*\*Braille and Large Print counts are based on students who actually tested and were not included in the total n-count.

**Table 8.2. Number of Students Tested by Demographics—Grade 4**

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
<b>Total N-Count</b>		<b>23,982</b>	<b>100.00</b>	<b>23,967</b>	<b>100.00</b>
Gender	Female	11,703	48.80	11,699	48.81
	Male	12,279	51.20	12,268	51.19
Ethnicity	AI/AN	307	1.28	307	1.28
	Asian	659	2.75	659	2.75
	Black or African American	1,595	6.65	1,592	6.64
	Hispanic	4,781	19.94	4,778	19.94
	NH/PI	30	0.13	30	0.13
	White	15,552	64.85	15,544	64.86
	Two or More Races	1,058	4.41	1,057	4.41
FRL	Yes	11,273	47.01	11,265	47.00
	No	12,709	52.99	12,702	53.00
LEP	Yes	3,715	15.49	3,712	15.49
	No	20,267	84.51	20,255	84.51
SPED	Yes	3,945	16.45	3,926	16.38
	No	20,037	83.55	20,041	83.62

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
Universal Features & Accommodations	Answer Choice Eliminator	15,022	62.64	15,781	65.84
	Highlighter	10,564	44.05	7,741	32.30
	Line Reader	14,118	58.87	6,168	25.74
	Notepad	10,094	42.09	10,749	44.85
	Text-to-Speech (TTS)	3,950	16.47	3,563	14.87
	Zoom	8,872	36.99	5,080	21.20
	Paper-Pencil (PP)	25	0.10	24	0.10
	Spanish Online	15	0.06	44	0.18
	Spanish Paper-Pencil (PP)	2	0.01	2	0.01
	Braille**	1	–	1	–
Large Print**	5	–	5	–	

\*AI/AN = American Indian or Alaskan Native. NH/PI = Native Hawaiian or Other Pacific Islander. FRL = free and reduced lunch. LEP = limited English proficient. SPED = special education.

\*\*Braille and Large Print counts are based on students who actually tested and were not included in the total n-count.

**Table 8.3. Number of Students Tested by Demographics—Grade 5**

Demographic Sub-Group*		ELA		Mathematics		Science	
		N	%	N	%	N	%
<b>Total N-Count</b>		<b>24,009</b>	<b>100.00</b>	<b>23,998</b>	<b>100.00</b>	<b>24,002</b>	<b>100.00</b>
Gender	Female	11,652	48.53	11,648	48.54	11,645	48.52
	Male	12,357	51.47	12,350	51.46	12,357	51.48
Ethnicity	AI/AN	292	1.22	292	1.22	292	1.22
	Asian	680	2.83	680	2.83	680	2.83
	Black or African American	1,690	7.04	1,689	7.04	1,688	7.03
	Hispanic	4,657	19.40	4,648	19.37	4,655	19.39
	NH/PI	37	0.15	37	0.15	37	0.15
	White	15,640	65.14	15,639	65.17	15,638	65.15
	Two or More Races	1,013	4.22	1,013	4.22	1,012	4.22
FRL	Yes	11,198	46.64	11,185	46.61	11,268	46.95
	No	12,811	53.36	12,813	53.39	12,734	53.05
LEP	Yes	3,720	15.49	3,718	15.49	3,722	15.51
	No	20,289	84.51	20,280	84.51	20,280	84.49
SPED	Yes	3,878	16.15	3,873	16.14	3,873	16.14
	No	20,131	83.85	20,125	83.86	20,129	83.86

Demographic Sub-Group*		ELA		Mathematics		Science	
		N	%	N	%	N	%
Universal Features & Accommodations	Answer Choice Eliminator	13,971	58.19	14,672	61.14	13,966	58.19
	Highlighter	8,939	37.23	5,565	23.19	4,896	20.40
	Line Reader	12,341	51.40	4,348	18.12	4,294	17.89
	Notepad	8,957	37.31	8,804	36.69	6,471	26.96
	Text-to-Speech (TTS)	3,663	15.26	3,038	12.66	3,243	13.51
	Zoom	7,473	31.13	3,402	14.18	4,920	20.50
	Calculator (basic)	–	–	208	0.87	–	–
	Paper-Pencil (PP)	28	0.12	27	0.11	28	0.12
	Spanish Online	36	0.15	70	0.29	71	0.30
	Spanish Paper-Pencil (PP)	3	0.01	4	0.02	3	0.01
	Braille**	4	–	4	–	4	–
	Large Print**	5	–	5	–	5	–

\*AI/AN = American Indian or Alaskan Native. NH/PI = Native Hawaiian or Other Pacific Islander. FRL = free and reduced lunch. LEP = limited English proficient. SPED = special education.

\*\*Braille and Large Print counts are based on students who actually tested and were not included in the total n-count.

**Table 8.4. Number of Students Tested by Demographics—Grade 6**

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
<b>Total N-Count</b>		<b>22,431</b>	<b>100.00</b>	<b>22,414</b>	<b>100.00</b>
Gender	Female	10,830	48.28	10,823	48.29
	Male	11,601	51.72	11,591	51.71
Ethnicity	AI/AN	309	1.38	310	1.38
	Asian	598	2.67	597	2.66
	Black or African American	1,514	6.75	1,514	6.75
	Hispanic	4,320	19.26	4,317	19.26
	NH/PI	39	0.17	39	0.17
	White	14,738	65.70	14,724	65.69
	Two or More Races	913	4.07	913	4.07
FRL	Yes	10,234	45.62	10,111	45.11
	No	12,197	54.38	12,303	54.89
LEP	Yes	3,133	13.97	3,130	13.96
	No	19,298	86.03	19,284	86.04
SPED	Yes	3,282	14.63	3,266	14.57
	No	19,149	85.37	19,148	85.43

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
Universal Features & Accommodations	Answer Choice Eliminator	11,285	50.31	14,164	63.19
	Highlighter	6,810	30.36	4,513	20.13
	Line Reader	9,779	43.60	4,436	19.79
	Notepad	6,778	30.22	9,135	40.76
	Text-to-Speech (TTS)	2,658	11.85	1,936	8.64
	Zoom	5,921	26.40	2,485	11.09
	Calculator (basic)	–	–	16,584	73.99
	Calculator (scientific)	–	–	927	4.14
	Paper-Pencil (PP)	25	0.11	25	0.11
	Spanish Online	51	0.23	80	0.36
	Spanish Paper-Pencil (PP)	1	0.00	1	0.00
	Braille**	2	–	2	–
	Large Print**	2	–	2	–

\*AI/AN = American Indian or Alaskan Native. NH/PI = Native Hawaiian or Other Pacific Islander. FRL = free and reduced lunch. LEP = limited English proficient. SPED = special education.

\*\*Braille and Large Print counts are based on students who actually tested and were not included in the total n-count.

**Table 8.5. Number of Students Tested by Demographics—Grade 7**

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
<b>Total N-Count</b>		<b>23,562</b>	<b>100.00</b>	<b>23,547</b>	<b>100.00</b>
Gender	Female	11,501	48.81	11,496	48.82
	Male	12,061	51.19	12,051	51.18
Ethnicity	AI/AN	315	1.34	316	1.34
	Asian	629	2.67	629	2.67
	Black or African American	1,624	6.89	1,620	6.88
	Hispanic	4,570	19.40	4,569	19.40
	NH/PI	32	0.14	32	0.14
	White	15,496	65.77	15,486	65.77
	Two or More Races	896	3.80	895	3.80
FRL	Yes	10,860	46.09	10,736	45.59
	No	12,702	53.91	12,811	54.41
LEP	Yes	3,577	15.18	3,575	15.18
	No	19,985	84.82	19,972	84.82
SPED	Yes	3,416	14.50	3,405	14.46
	No	20,146	85.50	20,142	85.54

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
Universal Features & Accommodations	Answer Choice Eliminator	9,508	40.35	11,628	49.38
	Highlighter	5,884	24.97	3,671	15.59
	Line Reader	8,878	37.68	3,895	16.54
	Notepad	5,392	22.88	7,847	33.32
	Text-to-Speech (TTS)	2,477	10.51	1,699	7.22
	Zoom	4,667	19.81	2,771	11.77
	Calculator (basic)	–	–	1,044	4.43
	Calculator (scientific)	–	–	21,233	90.17
	Paper-Pencil (PP)	30	0.13	30	0.13
	Spanish Online	53	0.22	93	0.39
	Spanish Paper-Pencil (PP)	3	0.01	3	0.01
	Braille**	3	–	3	–
	Large Print**	4	–	4	–

\*AI/AN = American Indian or Alaskan Native. NH/PI = Native Hawaiian or Other Pacific Islander. FRL = free and reduced lunch. LEP = limited English proficient. SPED = special education.

\*\*Braille and Large Print counts are based on students who actually tested and were not included in the total n-count.

**Table 8.6. Number of Students Tested by Demographics—Grade 8**

Demographic Sub-Group*		ELA		Mathematics		Science	
		N	%	N	%	N	%
<b>Total N-Count</b>		<b>23,226</b>	<b>100.00</b>	<b>23,217</b>	<b>100.00</b>	<b>23,214</b>	<b>100.00</b>
Gender	Female	11,270	48.52	11,268	48.53	11,268	48.54
	Male	11,956	51.48	11,949	51.47	11,946	51.46
Ethnicity	AI/AN	278	1.20	279	1.20	277	1.19
	Asian	613	2.64	613	2.64	613	2.64
	Black or African American	1,600	6.89	1,598	6.88	1,594	6.87
	Hispanic	4,408	18.98	4,414	19.01	4,410	19.00
	NH/PI	39	0.17	39	0.17	39	0.17
	White	15,419	66.39	15,405	66.35	15,415	66.40
	Two or More Races	869	3.74	869	3.74	866	3.73
FRL	Yes	10,369	44.64	10,219	44.02	10,215	44.00
	No	12,857	55.36	12,998	55.98	12,999	56.00
LEP	Yes	3,348	14.41	3,349	14.42	3,349	14.43
	No	19,878	85.59	19,868	85.58	19,865	85.57
SPED	Yes	3,251	14.00	3,243	13.97	3,235	13.94
	No	19,975	86.00	19,974	86.03	19,979	86.06

Demographic Sub-Group*		ELA		Mathematics		Science	
		N	%	N	%	N	%
Universal Features & Accommodations	Answer Choice Eliminator	6,675	28.74	11,994	51.66	6,633	28.57
	Highlighter	4,992	21.49	2,489	10.72	1,859	8.01
	Line Reader	6,638	28.58	3,616	15.57	2,018	8.69
	Notepad	4,180	18.00	6,477	27.90	2,735	11.78
	Text-to-Speech (TTS)	1,930	8.31	1,149	4.95	1,475	6.35
	Zoom	3,185	13.71	1,753	7.55	1,823	7.85
	Calculator (scientific)	–	–	20,056	86.38	–	–
	Paper-Pencil (PP)	56	0.24	56	0.24	55	0.24
	Spanish Online	97	0.42	131	0.56	132	0.57
	Spanish Paper-Pencil (PP)	2	0.01	2	0.01	1	0.00
	Braille**	0	–	0	–	0	–
	Large Print**	5	–	4	–	4	–

\*AI/AN = American Indian or Alaskan Native. NH/PI = Native Hawaiian or Other Pacific Islander. FRL = free and reduced lunch. LEP = limited English proficient. SPED = special education.

\*\*Braille and Large Print counts are based on students who actually tested and were not included in the total n-count.

## 8.2. Administration Mode (Online vs. Paper-Pencil)

The 2019 NSCAS assessments were administered online to the extent practical, and a very small number of students took the paper-pencil test. As shown in Table 8.7, less than 1% of students took the assessment in the paper-based version across all grades and content areas.

**Table 8.7. Number of Students Tested by Administration Mode**

Grade	Total #Students	Online N	Paper-Pencil	
			N	%
<b>ELA</b>				
3	23,422	23,406	16	0.1
4	23,960	23,935	25	0.1
5	23,967	23,940	27	0.1
6	22,376	22,351	25	0.1
7	23,491	23,461	30	0.1
8	23,108	23,054	54	0.2
<b>Mathematics</b>				
3	23,385	23,369	16	0.1
4	23,914	23,890	24	0.1
5	23,920	23,893	27	0.1
6	22,327	22,304	23	0.1
7	23,442	23,412	30	0.1
8	23,068	23,018	50	0.2
<b>Science</b>				
5	23,923	23,897	26	0.1
8	23,070	23,019	51	0.2



### 8.3. Testing Time

Table 8.8, Table 8.9, and Table 8.10 present the number of minutes students took to complete the Spring 2019 NSCAS ELA, Mathematics, and Science assessments, respectively. Specifically, the tables present the number and percent of students who completed the tests in various time ranges. As shown in the tables, most students completed the ELA test in 40–120 minutes, the Mathematics test in 20–100 minutes, and the Science test in 10–60 minutes. Most students finished the tests within 120 minutes, and the percentage of students who took more than 180 minutes is less than 2%.

**Table 8.8. Testing Time in Minutes—ELA**

Time in Minutes	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	N	%	N	%	N	%	N	%	N	%	N	%
<10	16	0.1	12	0.1	13	0.1	16	0.1	37	0.2	32	0.1
10 – <20	102	0.4	92	0.4	66	0.3	63	0.3	89	0.4	101	0.4
20 – <30	444	1.9	292	1.2	248	1.0	173	0.8	306	1.3	296	1.3
30 – <40	1,253	5.4	1,042	4.4	662	2.8	570	2.5	926	3.9	767	3.3
40 – <50	2,385	10.2	2,100	8.8	1,468	6.1	1,428	6.4	2,051	8.7	1,817	7.9
50 – <60	3,278	14.0	3,178	13.3	2,456	10.3	2,576	11.5	3,336	14.2	3,023	13.1
60 – <70	3,618	15.5	3,806	15.9	3,280	13.7	3,255	14.6	3,966	16.9	3,797	16.5
70 – <80	3,120	13.3	3,455	14.4	3,544	14.8	3,373	15.1	3,670	15.6	3,570	15.5
80 – <90	2,596	11.1	2,941	12.3	3,128	13.1	2,973	13.3	2,849	12.1	2,991	13.0
90 – <100	1,963	8.4	2,177	9.1	2,565	10.7	2,376	10.6	2,148	9.1	2,185	9.5
100 – <110	1,386	5.9	1,573	6.6	1,971	8.2	1,721	7.7	1,475	6.3	1,578	6.8
110 – <120	1,038	4.4	1,110	4.6	1,400	5.8	1,202	5.4	863	3.7	1,083	4.7
120 – <130	685	2.9	706	2.9	956	4.0	837	3.7	613	2.6	615	2.7
130 – <140	440	1.9	451	1.9	689	2.9	534	2.4	383	1.6	408	1.8
140 – <150	332	1.4	349	1.5	434	1.8	399	1.8	291	1.2	269	1.2
150 – <160	212	0.9	211	0.9	334	1.4	264	1.2	153	0.7	170	0.7
160 – <170	138	0.6	133	0.6	232	1.0	163	0.7	91	0.4	112	0.5
170 – <180	111	0.5	84	0.4	152	0.6	126	0.6	82	0.3	81	0.4
>=180	297	1.3	228	1.0	344	1.4	305	1.4	147	0.6	176	0.8
<b>Total</b>	<b>23,414</b>	<b>100.0</b>	<b>23,940</b>	<b>100.0</b>	<b>23,942</b>	<b>100.0</b>	<b>22,354</b>	<b>100.0</b>	<b>23,476</b>	<b>100.0</b>	<b>23,071</b>	<b>100.0</b>

**Table 8.9. Testing Time in Minutes—Mathematics**

Time in Minutes	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	N	%	N	%	N	%	N	%	N	%	N	%
<10	4	0.0	11	0.0	4	0.0	11	0.0	29	0.1	44	0.2
10 – <20	80	0.3	81	0.3	65	0.3	53	0.2	102	0.4	147	0.6
20 – <30	901	3.9	486	2.0	329	1.4	229	1.0	321	1.4	400	1.7
30 – <40	3,512	15.0	1,931	8.1	1,750	7.3	915	4.1	844	3.6	1,122	4.9
40 – <50	5,012	21.4	3,745	15.7	3,872	16.2	2,224	10.0	2,009	8.6	2,589	11.2
50 – <60	4,519	19.3	4,471	18.7	4,793	20.1	3,453	15.5	3,264	13.9	4,150	18.0
60 – <70	3,490	14.9	3,892	16.3	4,139	17.3	3,791	17.0	3,846	16.4	4,272	18.6
70 – <80	2,188	9.4	3,010	12.6	3,188	13.3	3,409	15.3	3,543	15.1	3,513	15.3
80 – <90	1,332	5.7	2,114	8.8	2,092	8.8	2,695	12.1	2,942	12.6	2,478	10.8
90 – <100	849	3.6	1,430	6.0	1,265	5.3	1,945	8.7	2,130	9.1	1,655	7.2
100 – <110	478	2.0	952	4.0	817	3.4	1,216	5.5	1,458	6.2	1,068	4.6
110 – <120	326	1.4	594	2.5	535	2.2	816	3.7	953	4.1	603	2.6
120 – <130	209	0.9	388	1.6	363	1.5	525	2.4	625	2.7	369	1.6
130 – <140	145	0.6	274	1.1	206	0.9	332	1.5	414	1.8	216	0.9
140 – <150	85	0.4	170	0.7	168	0.7	237	1.1	294	1.3	137	0.6
150 – <160	68	0.3	101	0.4	95	0.4	142	0.6	197	0.8	86	0.4
160 – <170	50	0.2	82	0.3	68	0.3	92	0.4	124	0.5	59	0.3
170 – <180	25	0.1	54	0.2	41	0.2	65	0.3	94	0.4	43	0.2
>=180	100	0.4	111	0.5	107	0.4	158	0.7	232	1.0	77	0.3
<b>Total</b>	<b>23,373</b>	<b>100.0</b>	<b>23,897</b>	<b>100.0</b>	<b>23,897</b>	<b>100.0</b>	<b>22,308</b>	<b>100.0</b>	<b>23,421</b>	<b>100.0</b>	<b>23,028</b>	<b>100.0</b>

**Table 8.10. Testing Time in Minutes—Science**

Time in Minutes	Grade 5		Grade 8	
	N	%	N	%
<10	23	0.1	53	0.2
10 – <20	1,175	4.9	1,583	6.9
20 – <30	7,006	29.3	7,961	34.6
30 – <40	7,195	30.1	6,664	28.9
40 – <50	4,249	17.8	3,359	14.6
50 – <60	2,071	8.7	1,675	7.3
60 – <70	1,041	4.4	793	3.4
70 – <80	526	2.2	438	1.9
80 – <90	294	1.2	213	0.9
90 – <100	136	0.6	111	0.5
100 – <110	87	0.4	78	0.3
110 – <120	41	0.2	30	0.1
120 – <130	15	0.1	24	0.1
130 – <140	12	0.1	13	0.1
140 – <150	5	0.0	7	0.0
150 – <160	6	0.0	5	0.0
160 – <170	4	0.0	4	0.0
170 – <180	3	0.0	3	0.0
>=180	11	0.0	12	0.1
<b>Total</b>	<b>23,900</b>	<b>100.0</b>	<b>23,026</b>	<b>100.0</b>

## 8.4. Achievement Level Distributions

Table 8.11 presents the achievement level distributions for the Spring 2019 NSCAS Summative assessments. Appendix L provides the achievement level distributions by demographic group. For ELA, 42–51% of students are at Developing and 48–58% of students are at On Track or CCR Benchmark. For Mathematics, 45–52% of students are at Developing and 48–55% of students are at On Track or CCR Benchmark. For Science, 31–37% of students are at Below the Standards and 63–69% are at Meets or Exceeds the Standards.

**Table 8.11. Achievement Level Distributions**

Grade	Total N-Count	Level 3*		Level 2*		Level 1*		Level 2 + Level 1	
		N-Count	%	N-Count	%	N-Count	%	N-Count	%
<b>ELA</b>									
3	23,422	10,213	43.6	9,135	39.0	4,074	17.4	13,209	56.4
4	23,960	9,991	41.7	9,460	39.5	4,509	18.8	13,969	58.3
5	23,967	12,427	51.9	7,878	32.9	3,662	15.3	11,540	48.1
6	22,376	11,353	50.7	7,073	31.6	3,950	17.7	11,023	49.3
7	23,491	12,009	51.1	8,946	38.1	2,536	10.8	11,482	48.9
8	23,108	11,450	49.5	8,315	36.0	3,343	14.5	11,658	50.5
<b>Mathematics</b>									
3	23,385	10,497	44.9	10,609	45.4	2,279	9.7	12,888	55.1
4	23,914	11,554	48.3	10,415	43.6	1,945	8.1	12,360	51.7
5	23,920	10,949	45.8	10,405	43.5	2,566	10.7	12,971	54.2
6	22,327	10,010	44.8	10,137	45.4	2,180	9.8	12,317	55.2
7	23,442	11,963	51.0	9,456	40.3	2,023	8.6	11,479	49.0
8	23,068	12,113	52.5	8,562	37.1	2,393	10.4	10,955	47.5
<b>Science</b>									
5	23,923	7,364	30.8	12,391	51.8	4,168	17.4	16,559	69.2
8	23,070	8,492	36.8	11,537	50.0	3,041	13.2	14,578	63.2

\*Achievement levels for ELA and Mathematics = Level 3: Developing, Level 2: On Track, and Level 1: CCR Benchmark. Achievement levels for Science = Level 3 = Below the Standards, Level 2 = Meets the Standards, and Level 1 = Exceeds the Standards.

## 8.5. Descriptive Statistics of Scale Scores

Table 8.12 presents the descriptive statistics for the scale scores, including the mean, standard deviation (SD), and scores at the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles. Appendix L also presents the descriptive statistics by demographic group. The mean scale score increases with the grade for ELA and Mathematics, as expected.

**Table 8.12. Scale Score Descriptive Statistics**

Grade	N-Count	Mean	SD	Percentiles						
				P5	P10	P25	P50	P75	P90	P95
<b>ELA</b>										
3	23,422	2485.96	72.21	2365	2388	2431	2491	2538	2578	2601
4	23,960	2513.91	73.43	2391	2415	2462	2517	2567	2608	2630
5	23,967	2525.87	69.71	2411	2433	2477	2528	2575	2616	2636
6	22,376	2538.13	68.18	2421	2449	2493	2541	2587	2624	2644
7	23,491	2544.99	72.86	2414	2443	2497	2553	2599	2633	2653
8	23,108	2558.03	68.33	2435	2467	2512	2562	2606	2645	2664

Grade	N-Count	Mean	SD	Percentiles						
				P5	P10	P25	P50	P75	P90	P95
<b>Mathematics</b>										
3	23,385	1194.87	71.80	1074	1099	1147	1198	1244	1283	1310
4	23,914	1224.90	66.91	1115	1138	1178	1224	1270	1310	1335
5	23,920	1244.23	70.60	1134	1157	1196	1242	1286	1335	1363
6	22,327	1252.83	68.20	1143	1167	1206	1252	1299	1340	1368
7	23,442	1252.14	64.36	1159	1175	1207	1244	1291	1338	1372
8	23,068	1267.63	71.38	1165	1182	1215	1259	1313	1368	1397
<b>Science</b>										
5	23,923	102.66	32.09	50	63	82	104	126	146	159
8	23,070	97.42	32.17	46	54	73	97	118	141	150

### 8.6. Reporting Category Correlations

For each grade and content area, Pearson’s correlation coefficients were calculated between reporting category scores to provide information on score dimensionality, which is part of validity evidence based on the tests’ internal structure. Disattenuated correlations provide an estimate of the relationships between reporting categories if there is no measurement error. Table 8.13 – 7.18 provide the reporting category correlations, and Table 8.19 – 7.24 present the disattenuated correlations.

The correlations between reporting categories within the content areas are positive and moderate in value, ranging from 0.54 (between Geometry and Data for Grade 4) to 0.77 (between Number and Algebra for Grade 5). The correlations between reporting categories across the content areas are positive and low to moderate in value, ranging from 0.45 (between Writing Skills and Data for Grade 8) and 0.68 (between Reading Vocabulary and Data for Grade 6). These ranges are similar to those from last year. In general, the within-content-area reporting category correlations are higher than the across-content-area reporting category correlations.

The disattenuated correlation are higher than the correlations, which is expected given that none of the reporting categories has perfect reliabilities (see Table 9.1 – Table 9.3). The disattenuated correlations between reporting categories within the content areas are positive and high in value: 0.84 (between Algebra and Data for Grade 6) or higher. The disattenuated correlations between reporting categories across the content areas are positive and moderate in value, ranging from 0.71 (between Writing Skills and Data for Grade 6,) or higher. These ranges are similar to those from last year. The high disattenuated correlations within the content suggest that reporting categories might be measuring essentially the same construct, which is one evidence based on internal structure. In other words, the internal structure of the assessments is consistent with the structure of the content standards.

**Table 8.13. Reporting Category Correlations—Grade 3**

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00						
Reading Comprehension	0.73	1.00					

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Writing Skills	0.61	0.68	1.00				
Number	0.63	0.66	0.56	1.00			
Algebra	0.55	0.59	0.50	0.68	1.00		
Geometry	0.59	0.63	0.53	0.75	0.63	1.00	
Data	0.63	0.68	0.57	0.75	0.66	0.71	1.00

**Table 8.14. Reporting Category Correlations—Grade 4**

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00						
Reading Comprehension	0.72	1.00					
Writing Skills	0.63	0.71	1.00				
Number	0.54	0.61	0.56	1.00			
Algebra	0.59	0.65	0.60	0.73	1.00		
Geometry	0.53	0.58	0.54	0.65	0.64	1.00	
Data	0.49	0.54	0.49	0.60	0.61	0.54	1.00

**Table 8.15. Reporting Category Correlations—Grade 5**

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data	Inquiry, Nature of Science & Tech	Physical Science	Life Science	Earth/Space Sciences
Reading Vocabulary	1.00										
Reading Comprehension	0.68	1.00									
Writing Skills	0.63	0.73	1.00								
Number	0.57	0.64	0.60	1.00							
Algebra	0.55	0.62	0.59	0.77	1.00						
Geometry	0.55	0.59	0.57	0.68	0.65	1.00					
Data	0.54	0.61	0.58	0.64	0.62	0.59	1.00				
Inquiry, Nature of Science & Tech	0.60	0.67	0.63	0.61	0.60	0.58	0.59	1.00			
Physical Science	0.57	0.59	0.56	0.57	0.56	0.56	0.51	0.60	1.00		
Life Science	0.59	0.63	0.59	0.57	0.55	0.56	0.54	0.62	0.63	1.00	
Earth/Space Sciences	0.56	0.58	0.55	0.53	0.52	0.54	0.51	0.59	0.60	0.62	1.00

**Table 8.16. Reporting Category Correlations—Grade 6**

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00						
Reading Comprehension	0.69	1.00					
Writing Skills	0.59	0.70	1.00				
Number	0.54	0.62	0.56	1.00			
Algebra	0.57	0.65	0.59	0.75	1.00		
Geometry	0.49	0.58	0.53	0.68	0.69	1.00	
Data	0.45	0.54	0.48	0.61	0.61	0.57	1.00

**Table 8.17. Reporting Category Correlations—Grade 7**

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00						
Reading Comprehension	0.72	1.00					
Writing Skills	0.62	0.70	1.00				
Number	0.52	0.59	0.54	1.00			
Algebra	0.56	0.64	0.59	0.72	1.00		
Geometry	0.50	0.56	0.52	0.65	0.68	1.00	
Data	0.54	0.60	0.55	0.66	0.70	0.63	1.00

**Table 8.18. Reporting Category Correlations—Grade 8**

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data	Inquiry, Nature of Science & Tech	Physical Science	Life Science	Earth/Space Sciences
Reading Vocabulary	1.00										
Reading Comprehension	0.63	1.00									
Writing Skills	0.59	0.74	1.00								
Number	0.48	0.58	0.58	1.00							
Algebra	0.53	0.65	0.63	0.75	1.00						
Geometry	0.49	0.60	0.59	0.71	0.74	1.00					
Data	0.46	0.58	0.55	0.62	0.65	0.63	1.00				
Inquiry, Nature of Science & Tech	0.52	0.63	0.60	0.59	0.64	0.61	0.55	1.00			
Physical Science	0.49	0.59	0.57	0.55	0.60	0.57	0.51	0.63	1.00		
Life Science	0.52	0.64	0.60	0.54	0.59	0.56	0.51	0.63	0.66	1.00	
Earth/Space Sciences	0.50	0.60	0.57	0.55	0.59	0.57	0.52	0.62	0.66	0.68	1.00

**Table 8.19. Reporting Category Disattenuated Correlations—Grade 3**

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00						
Reading Comprehension	1.00	1.00					
Writing Skills	0.95	0.93	1.00				
Number	0.88	0.81	0.78	1.00			
Algebra	0.90	0.84	0.81	0.99	1.00		
Geometry	0.87	0.81	0.78	0.99	0.97	1.00	
Data	0.92	0.87	0.83	0.98	1.00	0.98	1.00

**Table 8.20. Reporting Category Disattenuated Correlations—Grade 4**

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00						
Reading Comprehension	0.97	1.00					
Writing Skills	0.98	0.98	1.00				
Number	0.74	0.74	0.78	1.00			
Algebra	0.85	0.83	0.88	0.95	1.00		
Geometry	0.78	0.76	0.81	0.86	0.90	1.00	
Data	0.79	0.77	0.81	0.88	0.94	0.85	1.00

**Table 8.21. Reporting Category Disattenuated Correlations—Grade 5**

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data	Inquiry, Nature of Science & Tech	Physical Science	Life Science	Earth/Space Sciences
Reading Vocabulary	1.00										
Reading Comprehension	0.96	1.00									
Writing Skills	0.96	0.96	1.00								
Number	0.81	0.78	0.79	1.00							
Algebra	0.81	0.78	0.81	0.98	1.00						
Geometry	0.89	0.82	0.87	0.95	0.95	1.00					
Data	0.93	0.90	0.93	0.95	0.96	1.00	1.00				
Inquiry, Nature of Science & Tech	1.02	0.97	1.00	0.89	0.91	0.98	1.00	1.00			
Physical Science	0.93	0.83	0.86	0.81	0.82	0.91	0.88	1.00	1.00		
Life Science	0.94	0.86	0.88	0.79	0.79	0.89	0.91	1.00	1.00	1.00	
Earth/Space Sciences	0.95	0.84	0.87	0.78	0.79	0.91	0.91	1.00	1.00	1.00	1.00

**Table 8.22. Reporting Category Disattenuated Correlations—Grade 6**

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00						
Reading Comprehension	0.98	1.00					
Writing Skills	0.92	0.92	1.00				
Number	0.80	0.77	0.76	1.00			
Algebra	0.83	0.79	0.79	0.95	1.00		
Geometry	0.76	0.76	0.76	0.92	0.92	1.00	
Data	0.72	0.73	0.71	0.85	0.84	0.84	1.00

**Table 8.23. Reporting Category Disattenuated Correlations—Grade 7**

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00						
Reading Comprehension	0.99	1.00					
Writing Skills	0.98	0.95	1.00				
Number	0.77	0.76	0.80	1.00			
Algebra	0.80	0.79	0.84	0.96	1.00		
Geometry	0.76	0.74	0.79	0.93	0.93	1.00	
Data	0.82	0.79	0.83	0.94	0.95	0.92	1.00

**Table 8.24. Reporting Category Disattenuated Correlations—Grade 8**

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data	Inquiry, Nature of Science & Tech	Physical Science	Life Science	Earth/Space Sciences
Reading Vocabulary	1.00										
Reading Comprehension	0.97	1.00									
Writing Skills	0.97	0.96	1.00								
Number	0.78	0.74	0.79	1.00							
Algebra	0.82	0.79	0.82	0.96	1.00						
Geometry	0.78	0.75	0.79	0.93	0.93	1.00					
Data	1.00	1.00	1.00	1.15	1.00	1.00	1.00				
Inquiry, Nature of Science & Tech	0.92	0.88	0.90	0.87	0.90	0.88	1.00	1.00			
Physical Science	0.84	0.79	0.82	0.78	0.81	0.79	1.00	0.98	1.00		
Life Science	0.88	0.85	0.86	0.76	0.79	0.77	0.99	0.97	0.98	1.00	
Earth/Space Sciences	0.84	0.79	0.80	0.76	0.78	0.77	1.00	0.94	0.96	0.99	1.00



## 8.7. Correlations with MAP Growth

Table 8.25 presents the correlation coefficients between MAP Growth and NSCAS scores for students who took both tests in Spring 2019. As shown in the table, the correlation coefficients range from 0.83 to 0.85 for ELA/Reading, 0.80 to 0.81 for ELA/Language Usage, and 0.87 to 0.89 for Mathematics. In general, these high correlations indicate that the relationship between MAP Growth and NSCAS test scores is strong, which can be considered validity evidence based on other variables.

**Table 8.25. Correlation and Descriptive Statistics of NSCAS and MAP Growth Scores**

Grade	N	<i>r</i>	NSCAS*				MAP Growth*			
			Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
<b>ELA/Reading</b>										
3	17,155	0.85	2483	71.63	2253	2748	200	15.55	143	245
4	17,429	0.84	2511	73.08	2252	2778	207	15.19	143	251
5	17,378	0.83	2524	69.81	2282	2833	212	15.31	142	256
6	16,275	0.84	2536	68.4	2292	2810	216	14.92	148	258
7	16,110	0.84	2540	73.13	2302	2750	218	15.61	144	263
8	15,507	0.84	2554	69.09	2326	2853	221	16.20	150	275
<b>ELA/Language Usage</b>										
3	5,019	0.80	2487	67.75	2291	2745	202	12.8	153	239
4	5,349	0.80	2514	69.75	2254	2738	208	12.25	149	245
5	5,154	0.80	2527	65.23	2282	2768	213	12.17	155	246
6	6,142	0.80	2545	64.82	2322	2810	218	11.96	154	263
7	5,872	0.81	2555	67.27	2302	2750	221	12.89	149	261
8	5,864	0.80	2569	65.41	2326	2853	224	12.93	156	268
<b>Mathematics</b>										
3	17,059	0.89	1193	71.25	1002	1470	204	13.89	127	257
4	17,282	0.87	1224	66.59	1012	1500	213	15.29	138	270
5	17,476	0.88	1241	69.65	1022	1510	222	17.03	137	283
6	16,402	0.88	1250	67.83	1032	1530	226	16.34	136	302
7	15,680	0.87	1249	63.43	1042	1540	230	18.01	143	294
8	15,286	0.87	1266	70.31	1052	1550	234	19.32	139	310

\*SD = standard deviation. Min. = minimum. Max. = maximum.

## 8.8. Score Differences between 2018 and 2019

To evaluate students' annual progress toward college and career readiness, the 2018 and 2019 data were merged by students who advanced by one grade using student ID and grade (i.e., students who repeated a grade or skipped a grade were not included). Therefore, there was no Grade 3 or Science output. About 95% or more students were merged for each grade and content area.

Table 8.26 presents the descriptive statistics of score differences from one grade to the next, including the minimum and maximum differences, the mean, standard deviation (SD), and score differences by percentile. For ELA, the mean score difference is highest for Grade 4 (33.26 score increase), and in general the mean decreases as the grade increases (15.14 for Grade 5, then about 8 for Grades 6–8). The median score difference is 9 or higher. The pattern for

Mathematics is similar. The mean score difference is highest for Grade 4 (33.207 score increase), and in general the mean decreases as the grade increases (18.06 for Grade 5, then about 12 for Grades 6 and 8). An exception is Grade 7 that has a negative score difference of -1.32. The median score difference is 0 for Grade 7 but 12 or higher for the rest of the grades.

Table 8.27 presents the number and percentage of students with a minimum score difference greater than or equal to 1, 10, 20, etc. Of particular interest are the  $\geq 1$  results. For ELA, 57% or more students show a minimum of 1 score point difference (thus, score increase from 2018), with Grade 4 having the highest percentage (76%). For Mathematics (without counting Grade 7), 63% or more students show a minimum of 1 score difference, with Grade 4 also having the highest percentage (79%). This demonstrates a score increase from 2018 to 2019 for more than half of the students across grades. For Mathematics Grade 7, 49% of students show a minimum of 1 score difference.

Table 8.28 presents the number and percentage of students who dropped either one or two achievement levels or increased one or two achievement levels from 2018 to 2019. For ELA, approximately 70% of students remained in the same achievement level, while 11–19% dropped to a lower achievement and 11–19% moved to a higher achievement level. For Mathematics, 73–77% of students remained in the same achievement level, while 11–17% dropped to a lower achievement level and 9–16% moved to a higher achievement level.

**Table 8.26. Descriptive Statistics of Score Point Differences from 2018 to 2019**

Grade	Total N	Min.	Max.	Mean	SD	Percentiles						
						P5	P10	P25	P50	P75	P90	P95
<b>ELA</b>												
4	22,828	-190	375	33.26	47.59	-44	-26	2	34	64	92	111
5	22,885	-206	395	15.14	44.70	-59	-41	-13	15	44	70	87
6	21,242	-213	359	7.43	42.76	-62	-46	-19	8	35	60	77
7	22,370	-344	356	7.35	44.65	-67	-48	-20	9	37	62	77
8	22,084	-290	311	8.35	43.68	-63	-46	-19	9	36	62	79
<b>Mathematics</b>												
4	22,797	-179	266	33.07	41.57	-36	-20	6	34	60	85	100
5	22,849	-220	304	18.06	41.74	-50	-34	-8	18	45	69	85
6	21,199	-229	276	11.91	41.30	-56	-39	-14	12	39	63	78
7	22,339	-209	257	-1.32	40.99	-69	-53	-28	0	25	49	65
8	22,044	-306	318	13.89	38.38	-48	-33	-10	14	38	61	76

**Table 8.27. Minimum Score Point Differences from 2018 to 2019**

Grade	Total N	Minimum Score Point Differences											
		$\geq 1$		$\geq 10$		$\geq 20$		$\geq 30$		$\geq 40$		$\geq 50$	
		N	%	N	%	N	%	N	%	N	%	N	%
<b>ELA</b>													
4	22,828	17,417	76.3	15,990	70.1	14,195	62.2	12,244	53.6	10,196	44.7	8,212	36.0
5	22,885	14,584	63.7	12,701	55.5	10,541	46.1	8,461	37.0	6,560	28.7	4,845	21.2
6	21,242	12,124	57.1	10,238	48.2	8,205	38.6	6,219	29.3	4,582	21.6	3,251	15.3
7	22,370	12,855	57.5	11,013	49.2	8,852	39.6	6,847	30.6	5,092	22.8	3,620	16.2
8	22,084	12,781	57.9	10,910	49.4	8,705	39.4	6,694	30.3	5,006	22.7	3,521	15.9

Grade	Total N	Minimum Score Point Differences											
		≥1		≥10		≥20		≥30		≥40		≥50	
		N	%	N	%	N	%	N	%	N	%	N	%
<b>Mathematics</b>													
4	22,797	18,045	79.2	16,550	72.6	14,594	64.0	12,470	54.7	10,152	44.5	7,871	34.5
5	22,849	15,372	67.3	13,413	58.7	11,089	48.5	8,793	38.5	6,650	29.1	4,855	21.3
6	21,199	13,120	61.9	11,157	52.6	9,038	42.6	7,034	33.2	5,132	24.2	3,663	17.3
7	22,339	10,935	49.0	8,871	39.7	6,702	30.0	4,809	21.5	3,346	15.0	2,202	9.9
8	22,044	14,265	64.7	12,169	55.2	9,615	43.6	7,283	33.0	5,245	23.8	3,529	16.0

**Table 8.28. Changes in Achievement Level from 2018 to 2019**

Grade	Total N	Achievement Level Change									
		-2		-1		Same		+1		+2	
		N	%	N	%	N	%	N	%	N	%
<b>ELA</b>											
4	22,828	36	0.2	2,541	11.1	15,902	69.7	4,220	18.5	129	0.6
5	22,885	96	0.4	4,351	19.0	15,930	69.6	2,474	10.8	34	0.2
6	21,242	103	0.5	3,083	14.5	14,960	70.4	2,999	14.1	97	0.5
7	22,370	118	0.5	3,781	16.9	15,633	69.9	2,790	12.5	48	0.2
8	22,084	48	0.2	2,756	12.5	15,855	71.8	3,351	15.2	74	0.3
<b>Mathematics</b>											
4	22,797	18	0.1	2,971	13.0	16,850	73.9	2,946	12.9	12	0.1
5	22,849	11	0.1	2,434	10.7	16,774	73.4	3,612	15.8	18	0.1
6	21,199	19	0.1	2,316	10.9	15,424	72.8	3,419	16.1	21	0.1
7	22,339	30	0.1	3,649	16.3	16,594	74.3	2,061	9.2	5	0.0
8	22,044	10	0.1	2,692	12.2	16,918	76.8	2,424	11.0	0	0.0

## Section 9: Reliability

The *Standards* refer to reliability as the “consistency of scores across replications of a testing procedure” (AERA et al., 2014, p. 33). The level of reliability/precision of scores has implications for validity. In other words, scores must be consistent and precise enough to be useful for intended purposes. If scores are to be meaningful, tests should produce stable scores if the same group of students were to take the same test repeatedly without any fatigue or memory of the test. In addition, the range of certainty around the score should be small enough to support educational decisions. The reliability/precision of the 2019 NSCAS assessments was examined through analysis of measurement error in simulated and operational conditions, as follows:

- Score precision and reliability of the constraint-based engine (see Section 5.2.4)
- Marginal reliability
- Conditional standard error of measurement (CSEM)
- Cronbach’s alpha and standard error of measurement (SEM) for fixed forms

Combined, these data provide several ways of looking at the reliability of the NSCAS Summative assessments. Simulation results and marginal reliability statistics, as well as Cronbach’s alpha and SEM for the Science fixed forms, operate at the content level and provide estimates of reliability for student scores on a test. CSEM and classification accuracy provide important information related to the NSCAS achievement level classifications. These are of particular interest in the context of state accountability requirements.

### 9.1. Marginal Reliability

Marginal reliability is typically used in adaptive assessments to investigate score stability and is estimated as the ratio of mean of true score variance (i.e. observed score variance minus mean error variance) to observed score variance, as explained in Section 5.2.1. Table 9.1, Table 9.2, and Table 9.3 present marginal reliabilities of scale scores by grade and reporting category for ELA, Mathematics, and Science, respectively. Marginal reliability estimates for the total scores are well above 0.80 (0.87 or higher), which is typically considered the minimally acceptable level of reliability. Because reliability for reporting categories are based on fewer items, they have lower reliability than total scores. Appendix M provides marginal reliability estimates for the total scores by demographic sub-group.

As shown in Table 9.4, reliability varies by overall score levels (i.e., deciles). Observed variance is from the total score, and error variance is calculated for each decile. All students take the same number of items, but the information delivered by the items differs. The most information, and hence lower error and higher reliability, is found where the pool has the most items. The NSCAS item pools have more items in the middle than the both end and are easy relative to the population, resulting in lower reliability with higher scores (Deciles 9 and 10).

**Table 9.1. Marginal Reliability of Scale Scores—ELA**

Grade	N	Total Score	Reading Vocabulary	Reading Comprehension	Writing Skills
3	23,432	0.90	0.64	0.84	0.64
4	23,965	0.90	0.66	0.84	0.63
5	23,970	0.90	0.61	0.83	0.70
6	22,379	0.90	0.59	0.84	0.69
7	23,506	0.90	0.63	0.84	0.64
8	23,127	0.90	0.51	0.83	0.72

**Table 9.2. Marginal Reliability of Scale Scores—Mathematics**

Grade	N	Total Score	Number	Algebra	Geometry	Data
3	23,389	0.92	0.80	0.59	0.72	0.73
4	23,921	0.92	0.81	0.73	0.70	0.58
5	23,924	0.92	0.82	0.76	0.62	0.55
6	22,333	0.92	0.78	0.80	0.70	0.66
7	23,451	0.91	0.72	0.78	0.68	0.69
8	23,084	0.92	0.75	0.81	0.78	0.39

**Table 9.3. Marginal Reliability of Scale Scores—Science**

Grade	N	Total Score	Inquiry, Nature of Science, & Tech	Physical Science	Life Science	Earth/Space Sciences
5	23,928	0.87	0.57	0.61	0.64	0.57
8	23,081	0.90	0.62	0.67	0.68	0.70

**Table 9.4. Marginal Reliability: Variance**

Content Area	Grade	N	Variance	Overall	Deciles									
					1	2	3	4	5	6	7	8	9	10
ELA	3	23,432	5217.16	0.90	0.88	0.91	0.91	0.92	0.92	0.91	0.91	0.90	0.89	0.84
	4	23,965	5399.76	0.90	0.88	0.91	0.92	0.92	0.92	0.92	0.91	0.91	0.89	0.84
	5	23,970	4859.32	0.90	0.86	0.89	0.90	0.91	0.92	0.92	0.92	0.91	0.90	0.85
	6	22,379	4648.92	0.90	0.87	0.90	0.91	0.91	0.91	0.91	0.91	0.91	0.90	0.86
	7	23,506	5324.46	0.90	0.87	0.91	0.91	0.92	0.92	0.91	0.91	0.91	0.89	0.85
	8	23,127	4695.98	0.90	0.86	0.90	0.90	0.91	0.91	0.91	0.91	0.91	0.90	0.86
Mathematics	3	23,389	5158.62	0.92	0.93	0.94	0.94	0.94	0.94	0.94	0.93	0.92	0.91	0.85
	4	23,921	4480.75	0.92	0.91	0.93	0.93	0.93	0.93	0.93	0.92	0.92	0.91	0.86
	5	23,924	4984.97	0.92	0.93	0.94	0.94	0.95	0.94	0.94	0.93	0.93	0.91	0.79
	6	22,333	4652.60	0.92	0.92	0.93	0.94	0.94	0.94	0.93	0.93	0.92	0.92	0.87
	7	23,451	4147.51	0.91	0.90	0.91	0.92	0.92	0.92	0.92	0.92	0.92	0.91	0.88
	8	23,084	5105.46	0.92	0.92	0.93	0.94	0.94	0.94	0.94	0.94	0.93	0.92	0.86
Science	5	23,928	1030.14	0.87	0.89	0.90	0.90	0.90	0.90	0.89	0.88	0.87	0.84	0.73
	8	23,081	1036.54	0.90	0.90	0.92	0.92	0.92	0.92	0.90	0.90	0.89	0.88	0.80

## 9.2. Conditional Standard Error of Measurement (CSEM)

The CSEM represents the degree of measurement error in scale score units and are conditioned on the ability of the student, meaning that the test has different levels of error at different points along the ability scale. When applied to an adaptive assessment, the CSEM will vary for the same scale score. It is therefore necessary to report averages.

CSEMs are especially useful for characterizing measurement precision regarding score levels used for decision making, such as the cut score that determines student proficiency on an assessment. Table 9.5 presents the CSEMs for the achievement level cut scores that demark proficiency on the NSCAS tests (i.e., On Track and CCR Benchmark for ELA and Mathematics, Meets and Standards and Exceeds the Standards for Science), including the number of students  $\pm 10$  scale score points from the cut scores, the mean CSEMs of students near the cut, and the standard deviation (SD) of the CSEMs.

Table 9.6 then presents the overall and by-decile CSEM. The overall CSEM is slightly higher for ELA (from 21.5 to 22.9) than for Mathematics (from 18.7 to 19.5). The low CSEM for Science is expected as its conversion slope is smaller than ELA or Mathematics. CSEM is also relatively similar between Deciles 2 and 9, while the CSEM tends to be higher at the first and last decile. This suggests that item pools have more items in the middle than at both ends and that more difficult items are needed for both ELA and Mathematics, which is consistent with reliability results. Appendix N presents scatterplots for scale score CSEM by reporting category for each content area and grade.

**Table 9.5. CSEMs at the Proficient Cut Scores**

Content Area	Grade	Level 3/Level 2 Cut Score*			Level 2/Level 1 Cut Score*		
		N	Mean CSEM	SD	N	Mean CSEM	SD
ELA	3	2,272	21.0	0.1	1,885	23.9	0.8
	4	2,452	20.5	0.5	2,046	23.9	0.8
	5	2,803	20.0	0.2	1,784	22.1	0.4
	6	2,648	20.0	0.2	2,005	21.5	0.5
	7	2,732	21.2	0.4	1,724	24.9	0.4
	8	2,675	20.6	0.6	1,832	22.0	0.1
Mathematics	3	2,765	17.7	0.5	1,169	22.9	0.7
	4	3,119	18.0	0.1	1,081	21.4	0.6
	5	2,851	16.9	0.4	1,073	22.6	1.0
	6	2,802	17.4	0.5	1,116	20.5	0.5
	7	3,215	17.9	0.3	900	20.1	0.4
	8	2,483	18.0	0.0	849	21.5	0.6
Science	5	5,485	10.0	0.0	3,503	12.5	0.5
	8	5,272	9.0	0.0	3,570	11.6	0.7

\*ELA and Mathematics: Level 3 = Developing, Level 2 = On Track, and Level 1 = CCR Benchmark. Science: Level 3 = Below the Standards, Level 2 = Meets the Standards, and Level 1 = Exceeds the Standards.

**Table 9.6. Mean CSEMs by Deciles**

Content Area	Grade	Mean CSEM	Mean CSEM by Decile									
			1	2	3	4	5	6	7	8	9	10
ELA	3	22.8	24.6	22.0	21.2	21.0	21.0	21.4	22.0	22.3	24.2	28.4
	4	22.9	25.8	22.2	21.1	20.5	20.5	21.0	21.8	22.6	24.6	29.0
	5	22.0	25.7	22.6	21.7	20.5	20.0	20.0	20.0	20.7	22.1	26.5
	6	21.5	24.8	21.4	20.2	20.0	20.0	20.0	20.0	20.6	21.9	25.6
	7	22.9	25.9	22.1	21.4	21.0	21.0	21.4	22.0	22.4	24.1	27.7
	8	22.0	25.2	22.0	21.4	21.0	20.9	20.5	21.0	21.0	21.8	25.3
Mathematics	3	19.4	19.5	17.6	17.0	17.0	17.7	18.1	18.9	19.8	21.3	27.4
	4	19.1	19.9	18.0	17.9	17.6	17.9	18.0	18.5	19.1	20.1	24.4
	5	19.1	18.4	17.0	16.8	16.4	16.8	17.4	18.0	18.9	20.9	30.7
	6	18.7	19.6	17.9	17.2	17.1	17.3	17.9	18.1	18.8	19.5	23.9
	7	18.9	20.8	18.8	18.0	18.0	18.0	17.9	17.8	18.0	19.0	22.3
	8	19.5	20.7	18.6	18.0	18.0	18.0	18.0	18.1	19.0	20.4	26.4
Science	5	11.3	10.7	10.0	10.0	10.0	10.0	10.5	11.0	11.5	12.6	16.3
	8	10.3	10.2	9.4	9.0	9.0	9.0	10.0	10.0	10.5	11.3	14.1

### 9.3. Classification Accuracy

Classification accuracy is a measure of how accurately test scores place students into reporting category levels. It refers to the agreement between the actual classifications using observed cut scores and true classifications based on known true cut scores. It is common to estimate classification accuracy by using a psychometric model to find true scores corresponding to observed scores. The likelihood of inaccurate placement depends on the amount of error associated with scores, especially those nearest cut points.

Classification accuracy was calculated as follows (SBAC, 2016):

1. For each student, a normal distribution was constructed with means equal to the scale score estimate and standard deviation equal to the SEM as a plausible true score distribution.
2. For each student, the proportion of that normal distribution that fell within each achievement level was calculated.
3. Within the groups of students assigned to a particular achievement level (Level 3, 2, or 1 for the overall score and for the reporting category scores), the sums of the proportions over students were computed. This provided estimates of the number of students whose true score falls within a level for each assigned achievement level. These sums were then expressed as a proportion of the total sample (i.e., expected proportion).
4. With the table of expected proportions, correct classification rates were then defined. This is the proportion of students whose true classification agrees the assigned level among the subset of students with that assigned level.
5. The overall classification rate is the sum of the proportions of students whose true score level agrees the assigned level, divided by the total proportion of students assigned to a level.

Table 9.7, Table 9.8, and Table 9.9 present the classification accuracy results by grade, achievement level, and reporting category. Overall classification accuracy ranges from 0.841 (ELA Grade 6) to 0.888 (Mathematics Grade 8). In general, classification accuracy is moderate to high. Considering that the magnitude of classification accuracy is influenced by key features of test design including the number of items, number of cut scores, and the reliability and associated SEM, the classification accuracy for 2019 suggests that accurate level classifications are being made for Nebraska students on the NSCAS assessments. Overall classification accuracy by achievement level ranges from 0.731 (ELA Grade 6 On Track) to 0.927 (Mathematics Grade 3 Developing).

**Table 9.7. Classification Accuracy by Achievement Level and Reporting Category—ELA**

Grade	Achievement Level	N	%	Expected Proportion*			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
<b>Overall</b>								
3	Developing	10,214	0.44	0.40	0.04	0.00	0.920	0.851
	On Track	9,135	0.39	0.04	0.31	0.04	0.795	
	CCR Benchmark	4,074	0.17	0.00	0.03	0.14	0.805	
4	Developing	9,994	0.42	0.38	0.04	0.00	0.909	0.846
	On Track	9,460	0.40	0.04	0.31	0.04	0.795	
	CCR Benchmark	4,509	0.19	0.00	0.04	0.15	0.814	
5	Developing	12,427	0.52	0.48	0.04	0.00	0.915	0.850
	On Track	7,878	0.33	0.04	0.25	0.03	0.763	
	CCR Benchmark	3,662	0.15	0.00	0.03	0.12	0.810	
6	Developing	11,353	0.51	0.47	0.04	0.00	0.919	0.841
	On Track	7,073	0.32	0.05	0.23	0.04	0.731	
	CCR Benchmark	3,950	0.18	0.00	0.03	0.14	0.814	
7	Developing	12,016	0.51	0.47	0.04	0.00	0.922	0.843
	On Track	8,946	0.38	0.05	0.29	0.04	0.759	
	CCR Benchmark	2,536	0.11	0.00	0.03	0.08	0.769	
8	Developing	11,461	0.50	0.45	0.04	0.00	0.911	0.845
	On Track	8,315	0.36	0.05	0.28	0.03	0.775	
	CCR Benchmark	3,343	0.15	0.00	0.03	0.11	0.786	
<b>Reading Vocabulary</b>								
3	Developing	9,648	0.41	0.35	0.06	0.00	0.854	0.720
	On Track	6,680	0.29	0.08	0.15	0.06	0.526	
	CCR Benchmark	7,093	0.30	0.00	0.07	0.22	0.719	
4	Developing	10,521	0.44	0.37	0.07	0.00	0.836	0.724
	On Track	7,539	0.32	0.08	0.17	0.07	0.549	
	CCR Benchmark	5,900	0.25	0.00	0.06	0.18	0.748	
5	Developing	12,457	0.52	0.43	0.08	0.00	0.819	0.712
	On Track	5,981	0.25	0.07	0.12	0.06	0.476	
	CCR Benchmark	5,527	0.23	0.00	0.05	0.17	0.723	



Grade	Achievement Level	N	%	Expected Proportion*			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
6	Developing	10,774	0.48	0.41	0.07	0.00	0.840	0.704
	On Track	6,072	0.27	0.08	0.11	0.08	0.413	
	CCR Benchmark	5,526	0.25	0.00	0.04	0.19	0.757	
7	Developing	12,242	0.52	0.45	0.06	0.00	0.871	0.719
	On Track	6,463	0.28	0.08	0.12	0.07	0.447	
	CCR Benchmark	4,775	0.20	0.00	0.05	0.14	0.700	
8	Developing	11,815	0.51	0.42	0.08	0.00	0.818	0.686
	On Track	7,233	0.31	0.10	0.14	0.08	0.454	
	CCR Benchmark	4,051	0.18	0.00	0.04	0.13	0.720	
<b>Reading Comprehension</b>								
3	Developing	11,008	0.47	0.43	0.05	0.00	0.904	0.819
	On Track	8,088	0.35	0.05	0.25	0.04	0.728	
	CCR Benchmark	4,326	0.19	0.00	0.04	0.14	0.773	
4	Developing	9,660	0.40	0.35	0.05	0.00	0.876	0.803
	On Track	9,373	0.39	0.06	0.29	0.05	0.737	
	CCR Benchmark	4,927	0.21	0.00	0.04	0.16	0.786	
5	Developing	12,457	0.52	0.46	0.06	0.00	0.892	0.813
	On Track	7,380	0.31	0.05	0.21	0.04	0.695	
	CCR Benchmark	4,130	0.17	0.00	0.04	0.14	0.785	
6	Developing	11,087	0.50	0.45	0.05	0.00	0.901	0.801
	On Track	6,996	0.31	0.06	0.20	0.05	0.649	
	CCR Benchmark	4,293	0.19	0.00	0.04	0.15	0.792	
7	Developing	11,432	0.49	0.44	0.05	0.00	0.901	0.803
	On Track	9,113	0.39	0.06	0.27	0.06	0.693	
	CCR Benchmark	2,946	0.13	0.00	0.03	0.10	0.760	
8	Developing	11,282	0.49	0.44	0.05	0.00	0.898	0.803
	On Track	8,382	0.36	0.06	0.25	0.05	0.686	
	CCR Benchmark	3,444	0.15	0.00	0.03	0.12	0.779	
<b>Writing Skills</b>								
3	Developing	9,511	0.41	0.34	0.07	0.00	0.828	0.709
	On Track	9,682	0.41	0.09	0.23	0.09	0.567	
	CCR Benchmark	4,228	0.18	0.00	0.04	0.14	0.768	
4	Developing	10,872	0.45	0.39	0.06	0.00	0.855	0.718
	On Track	8,114	0.34	0.09	0.18	0.07	0.525	
	CCR Benchmark	4,972	0.21	0.00	0.05	0.15	0.731	
5	Developing	12,367	0.52	0.45	0.07	0.00	0.864	0.741
	On Track	7,887	0.33	0.08	0.18	0.07	0.538	
	CCR Benchmark	3,713	0.16	0.00	0.03	0.12	0.761	
6	Developing	12,184	0.55	0.46	0.08	0.00	0.850	0.742
	On Track	6,494	0.29	0.07	0.15	0.07	0.517	
	CCR Benchmark	3,694	0.17	0.00	0.03	0.13	0.782	

Grade	Achievement Level	N	%	Expected Proportion*			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
7	Developing	13,696	0.58	0.49	0.09	0.00	0.832	0.732
	On Track	6,455	0.28	0.07	0.15	0.06	0.531	
	CCR Benchmark	3,336	0.14	0.00	0.04	0.10	0.711	
8	Developing	11,040	0.48	0.41	0.06	0.00	0.864	0.746
	On Track	8,066	0.35	0.08	0.20	0.07	0.564	
	CCR Benchmark	4,000	0.17	0.00	0.04	0.14	0.786	

\*L3: Developing, L2: On Track, and L1: CCR Benchmark.

**Table 9.8. Classification Accuracy by Achievement Level and Reporting Category—Mathematics**

Grade	Achievement Level	N	%	Expected Proportion*			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
<b>Overall</b>								
3	Developing	10,499	0.45	0.42	0.03	0.00	0.927	0.877
	On Track	10,609	0.45	0.05	0.38	0.03	0.844	
	CCR Benchmark	2,279	0.10	0.00	0.02	0.08	0.804	
4	Developing	11,556	0.48	0.44	0.04	0.00	0.917	0.877
	On Track	10,415	0.44	0.04	0.37	0.02	0.844	
	CCR Benchmark	1,945	0.08	0.00	0.02	0.07	0.827	
5	Developing	10,949	0.46	0.42	0.03	0.00	0.926	0.882
	On Track	10,405	0.44	0.05	0.37	0.02	0.848	
	CCR Benchmark	2,566	0.11	0.00	0.02	0.09	0.832	
6	Developing	10,011	0.45	0.41	0.04	0.00	0.922	0.880
	On Track	10,137	0.45	0.05	0.39	0.02	0.852	
	CCR Benchmark	2,180	0.10	0.00	0.02	0.08	0.816	
7	Developing	11,966	0.51	0.46	0.05	0.00	0.910	0.880
	On Track	9,456	0.40	0.05	0.34	0.02	0.849	
	CCR Benchmark	2,023	0.09	0.00	0.01	0.07	0.860	
8	Developing	12,119	0.53	0.49	0.04	0.00	0.926	0.888
	On Track	8,562	0.37	0.04	0.32	0.02	0.849	
	CCR Benchmark	2,393	0.10	0.00	0.02	0.09	0.837	
<b>Number</b>								
3	Developing	10,621	0.45	0.40	0.06	0.00	0.879	0.803
	On Track	9,503	0.41	0.07	0.30	0.04	0.734	
	CCR Benchmark	3,261	0.14	0.00	0.03	0.11	0.763	
4	Developing	11,618	0.49	0.43	0.05	0.00	0.889	0.812
	On Track	9,300	0.39	0.06	0.28	0.04	0.730	
	CCR Benchmark	2,996	0.13	0.00	0.03	0.10	0.768	
5	Developing	10,909	0.46	0.40	0.05	0.00	0.882	0.818
	On Track	9,597	0.40	0.06	0.30	0.04	0.753	
	CCR Benchmark	3,414	0.14	0.00	0.03	0.11	0.797	

Grade	Achievement Level	N	%	Expected Proportion*			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
6	Developing	10,222	0.46	0.40	0.06	0.00	0.867	0.793
	On Track	8,670	0.39	0.06	0.28	0.05	0.727	
	CCR Benchmark	3,435	0.15	0.00	0.04	0.11	0.740	
7	Developing	11,709	0.50	0.43	0.07	0.00	0.852	0.778
	On Track	8,706	0.37	0.08	0.25	0.04	0.685	
	CCR Benchmark	3,021	0.13	0.00	0.03	0.10	0.760	
8	Developing	11,736	0.51	0.44	0.07	0.00	0.870	0.798
	On Track	8,284	0.36	0.06	0.26	0.04	0.710	
	CCR Benchmark	3,045	0.13	0.00	0.03	0.10	0.758	
<b>Algebra</b>								
3	Developing	11,794	0.50	0.42	0.08	0.00	0.835	0.741
	On Track	8,840	0.38	0.08	0.24	0.06	0.630	
	CCR Benchmark	2,746	0.12	0.00	0.03	0.08	0.701	
4	Developing	11,515	0.48	0.41	0.07	0.00	0.855	0.779
	On Track	9,308	0.39	0.07	0.27	0.05	0.697	
	CCR Benchmark	3,091	0.13	0.00	0.03	0.10	0.744	
5	Developing	10,691	0.45	0.38	0.06	0.00	0.859	0.781
	On Track	9,504	0.40	0.07	0.28	0.05	0.695	
	CCR Benchmark	3,723	0.16	0.00	0.03	0.12	0.776	
6	Developing	9,378	0.42	0.36	0.06	0.00	0.864	0.797
	On Track	10,202	0.46	0.07	0.34	0.05	0.737	
	CCR Benchmark	2,747	0.12	0.00	0.03	0.10	0.789	
7	Developing	11,305	0.48	0.42	0.07	0.00	0.861	0.804
	On Track	9,928	0.42	0.08	0.31	0.03	0.741	
	CCR Benchmark	2,209	0.09	0.00	0.02	0.08	0.798	
8	Developing	12,201	0.53	0.46	0.07	0.00	0.870	0.819
	On Track	8,035	0.35	0.06	0.26	0.03	0.759	
	CCR Benchmark	2,830	0.12	0.00	0.03	0.10	0.772	
<b>Geometry</b>								
3	Developing	10,598	0.45	0.40	0.06	0.00	0.876	0.765
	On Track	8,069	0.35	0.08	0.23	0.03	0.678	
	CCR Benchmark	4,718	0.20	0.00	0.06	0.13	0.663	
4	Developing	11,504	0.48	0.41	0.07	0.00	0.859	0.758
	On Track	9,599	0.40	0.08	0.26	0.06	0.643	
	CCR Benchmark	2,807	0.12	0.00	0.03	0.09	0.744	
5	Developing	12,503	0.52	0.44	0.08	0.00	0.845	0.748
	On Track	8,975	0.38	0.07	0.24	0.06	0.632	
	CCR Benchmark	2,436	0.10	0.00	0.03	0.07	0.676	
6	Developing	9,880	0.44	0.37	0.07	0.00	0.842	0.757
	On Track	9,396	0.42	0.09	0.28	0.05	0.670	
	CCR Benchmark	3,044	0.14	0.00	0.03	0.10	0.750	

Grade	Achievement Level	N	%	Expected Proportion*			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
7	Developing	11,739	0.50	0.42	0.08	0.00	0.840	0.769
	On Track	8,989	0.38	0.08	0.26	0.04	0.680	
	CCR Benchmark	2,708	0.12	0.00	0.03	0.09	0.750	
8	Developing	11,991	0.52	0.45	0.07	0.00	0.867	0.799
	On Track	7,961	0.35	0.07	0.24	0.04	0.704	
	CCR Benchmark	3,116	0.14	0.00	0.03	0.11	0.778	
<b>Data</b>								
3	Developing	10,998	0.47	0.40	0.07	0.00	0.857	0.774
	On Track	8,745	0.37	0.06	0.26	0.05	0.687	
	CCR Benchmark	3,637	0.16	0.00	0.04	0.11	0.731	
4	Developing	11,997	0.50	0.41	0.09	0.00	0.807	0.732
	On Track	8,883	0.37	0.08	0.23	0.06	0.629	
	CCR Benchmark	3,030	0.13	0.00	0.03	0.09	0.732	
5	Developing	11,258	0.47	0.40	0.07	0.00	0.845	0.719
	On Track	9,150	0.38	0.11	0.23	0.05	0.590	
	CCR Benchmark	3,506	0.15	0.00	0.05	0.10	0.646	
6	Developing	11,990	0.54	0.46	0.08	0.00	0.853	0.763
	On Track	7,301	0.33	0.07	0.21	0.04	0.645	
	CCR Benchmark	3,028	0.14	0.00	0.04	0.09	0.691	
7	Developing	11,796	0.50	0.42	0.09	0.00	0.825	0.761
	On Track	8,649	0.37	0.07	0.25	0.05	0.675	
	CCR Benchmark	2,994	0.13	0.00	0.03	0.10	0.758	
8	Developing	10,182	0.44	0.39	0.06	0.00	0.871	0.704
	On Track	9,427	0.41	0.12	0.22	0.07	0.528	
	CCR Benchmark	3,450	0.15	0.00	0.03	0.10	0.687	

\*L3: Developing, L2: On Track, and L1: CCR Benchmark.

**Table 9.9. Classification Accuracy by Achievement Level and Reporting Category—Science**

Grade	Achievement Level	N	%	Expected Proportion*			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
<b>Overall</b>								
5	Below	7,365	0.31	0.27	0.04	0.00	0.860	0.845
	Meets	12,391	0.52	0.04	0.45	0.03	0.861	
	Exceeds	4,168	0.17	0.00	0.04	0.13	0.770	
8	Below	8,497	0.37	0.33	0.04	0.00	0.897	0.861
	Meets	11,537	0.50	0.04	0.43	0.03	0.854	
	Exceeds	3,041	0.13	0.00	0.03	0.10	0.788	
<b>Inquiry, Nature of Science, &amp; Tech</b>								
5	Below	8,132	0.34	0.28	0.06	0.00	0.815	0.722
	Meets	10,967	0.46	0.08	0.29	0.08	0.638	
	Exceeds	4,823	0.20	0.00	0.04	0.15	0.757	

Grade	Achievement Level	N	%	Expected Proportion*			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
8	Below	8,143	0.35	0.28	0.07	0.00	0.796	0.732
	Meets	8,579	0.37	0.08	0.25	0.04	0.680	
	Exceeds	6,347	0.28	0.00	0.07	0.20	0.720	
<b>Physical Science</b>								
5	Below	7,684	0.32	0.24	0.08	0.00	0.745	0.723
	Meets	12,497	0.52	0.07	0.36	0.09	0.686	
	Exceeds	3,741	0.16	0.00	0.03	0.13	0.808	
8	Below	7,934	0.34	0.28	0.06	0.00	0.826	0.743
	Meets	10,562	0.46	0.09	0.32	0.05	0.699	
	Exceeds	4,573	0.20	0.00	0.06	0.14	0.702	
<b>Life Science</b>								
5	Below	8,726	0.37	0.29	0.08	0.00	0.789	0.735
	Meets	9,390	0.39	0.07	0.27	0.05	0.692	
	Exceeds	5,807	0.24	0.00	0.07	0.18	0.720	
8	Below	8,732	0.38	0.32	0.06	0.00	0.834	0.753
	Meets	10,211	0.44	0.07	0.30	0.07	0.684	
	Exceeds	4,125	0.18	0.00	0.04	0.13	0.749	
<b>Earth/Space Sciences</b>								
5	Below	8,192	0.34	0.27	0.08	0.00	0.781	0.717
	Meets	10,453	0.44	0.08	0.30	0.06	0.691	
	Exceeds	5,277	0.22	0.00	0.07	0.15	0.670	
8	Below	8,846	0.38	0.33	0.05	0.00	0.870	0.762
	Meets	11,412	0.50	0.10	0.34	0.05	0.683	
	Exceeds	2,808	0.12	0.00	0.03	0.09	0.738	

\*L3: Below the Standards, L2: Meets the Standards, and L1: Exceeds the Standards.

#### 9.4. Reliability for Fixed Forms (Science)

Cronbach's alpha reliability coefficient is a frequently used measure of internal consistency over the responses to a set of items measuring an underlying, unidimensional trait. Reliability coefficient alpha expresses the consistency of test scores as the ratio of true score variance to total score (observed) variance (true score variance + error variance). A larger index would indicate that test scores were influenced less by random sources of error. The reliability coefficient is a "unitless" index, which can be compared from test to test and ranges from 0.0 to 1.0, where 0.80 is typically considered the minimally acceptable level of reliability for assessments like NSCAS. While sensitive to random error associated with content sampling variability, the index is not sensitive to other types of errors, such as temporal stability or variability in performance that might occur across different testing occasions. Cronbach's alpha is computed as follows (Crocker & Algina, 1986):

$$\hat{\alpha} = \frac{k}{k-1} \left( 1 - \frac{\sum \sigma_j^2}{\sigma_x^2} \right) \quad (9.1)$$

where  $k$  = number of items,  $\sigma_x^2$  = the total score variance, and  $\sigma_j^2$  = the variance of item  $j$ . The SEM is an index of the random variability in test scores in raw score units and is defined as follows:

$$SEM = SD\sqrt{1 - \hat{\alpha}} \quad (9.2)$$

where SD represents the standard deviation of the raw score distribution and  $\hat{\alpha}$  represents Cronbach's alpha, as expressed in Equation 9.1. The overall SEM is expressed in raw score units and is a test-level statistic. Table 9.10 presents Cronbach's alpha reliability coefficients by demographics for the Science fixed forms, along with the SEMs. The alpha reliability coefficients are similar to marginal reliability (reported in Table 9.3) and to the 2018 results.

**Table 9.10. Cronbach's Alpha (Internal Consistency) by Demographics for Science Fixed Forms**

Grade	Demographic Group*		#Items	Reliability	SEM
5	<b>Grade 5 Overall</b>		<b>50</b>	<b>0.88</b>	<b>11.12</b>
	Gender	Female	50	0.87	11.12
		Male	50	0.89	11.01
	Ethnicity	AI/AN	50	0.86	10.52
		Asian	50	0.90	10.79
		Black or African American	50	0.86	10.78
		Hispanic	50	0.85	10.66
		NH/PI	50	0.87	10.54
		White	50	0.87	11.16
		Two or More Races	50	0.88	10.81
	FRL	Yes	50	0.87	10.65
		No	50	0.86	11.47
	LEP	Yes	50	0.85	10.69
		No	50	0.88	11.02
SPED	Yes	50	0.87	10.96	
	No	50	0.88	11.43	
8	<b>Grade 8 Overall</b>		<b>60</b>	<b>0.90</b>	<b>10.17</b>
	Gender	Female	60	0.89	10.04
		Male	60	0.91	10.14
	Ethnicity	AI/AN	60	0.88	10.30
		Asian	60	0.91	10.40
		Black or African American	60	0.87	9.94
		Hispanic	60	0.87	10.21
		NH/PI	60	0.88	10.57
		White	60	0.89	10.18
		Two or More Races	60	0.89	10.36
	FRL	Yes	60	0.88	10.20
		No	60	0.89	10.12
	LEP	Yes	60	0.87	10.18
		No	60	0.89	10.47
SPED	Yes	60	0.86	10.23	
	No	60	0.89	10.57	

\*AI/AN = American Indian or Alaskan Native. NH/PI = Native Hawaiian or Other Pacific Islander. FRL = free and reduced lunch. LEP = limited English proficient. SPED = special education.

## Section 10: Validity

Validity is defined by the *Standards* as the “the degree to which evidence and theory support the interpretations of test scores for proposed uses. Validity is, therefore, the most fundamental consideration in developing and evaluating tests” (AERA et al., 2014, p. 11). Validating a test score interpretation is not a quantifiable property but an ongoing process, beginning at initial conceptualization of the construct and continuing throughout the entire assessment process. Every aspect of an assessment development and administration process provides evidence in support of (or a challenge to) the validity of the intended inferences about what students know based on their score, including design, content specifications, item development, test constraints, psychometric quality, standard setting, and administration.

As the technical report has progressed, it has covered the different phases of the testing cycle and provided different pieces of technical quality evidence along the way. It provides relevant evidence and a rationale in support of test score interpretations and intended uses based on the *Standards*, as the *Standards* are considered to be “the most authoritative statement of professional consensus regarding the development and evaluation of educational and psychological tests” (Linn, 2006, p. 27). The validity argument begins with a statement of the assessment’s intended purposes, followed by the evidentiary framework where available validity evidence is provided to support the argument that the test actually measures what it purports to measure (SBAC, 2016).

### 10.1. Intended Purposes and Uses of Test Scores

The purposes of the NSCAS Summative assessment are as follows:

1. To measure and report Nebraska students’ depth of achievement regarding Nebraska’s College and Career Ready Standards for ELA and Mathematics in Grades 3–8 and Nebraska’s Science standards for Grades 5 and 8.
2. To report if student achievement is sufficient academic proficiency in ELA and Mathematics to be on track for achieving college readiness.
3. To measure students’ annual progress toward college and career readiness in ELA and Mathematics.
4. To inform teachers how student thinking differs along different areas of the scale as represented by the ALDs as information to support instructional planning.
5. To assess students’ construct relevant achievement in ELA, Mathematics, and Science for all students and subgroups of students.

As the *Standards* note, “validation is the joint responsibility of the test developer and the test user...the test user is ultimately responsible for evaluating the evidence in the particular setting in which the test is to be used” (AERA et al., 2014, p. 13). This report provides information about test content and technical quality but does not interfere in the use of scores. Ultimate use of test scores is determined by Nebraska educators. However, some intended uses of the NSCAS test results include the following:

- To supplement teachers’ observations and classroom assessment data and to improve the decisions teachers make about sequencing instructional goals, designing instructional materials, and selecting instructional approaches for groups and individuals
- To identify individuals for summer school and other remediation programs

- To gauge and improve the quality of education at the class, school, system, and state levels throughout Nebraska
- To assess the performance of a teacher, school, or system in conjunction with other sources of information

The unintended uses of the NSCAS are as follows:

- To place students in special education classes
- To apply group differences in test scores to admission and class grouping
- To narrow a school's curriculum to exclude learning of objectives that are not assessed

## 10.2. Sources of Validity Evidence

The *Standards* describe validation as a process of constructing and evaluating arguments for the intended interpretation and use of test scores:

“A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses. . . . Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system (AERA et al., 2014, pp. 21–22).”

The *Standards* (AERA et al., 2014, pp. 13–19) outline the following five main sources of validity evidence:

- Evidence based on test content
- Evidence based on response processes
- Evidence based on internal structure
- Evidence based on relations to other variables
- Evidence for validity and consequences of testing

Evidence based on test design refers to traditional forms of content validity or content-related evidence. Evidence based on response processes refers to the cognitive process engaged in by students when answering test items, or the “evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees” (AERA et al., 2014, p. 15). Evidence based on internal structure refer to the psychometric analyses of “the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based” (AERA et al., 2014, p. 16). Evidence based on relations to other variables refers to traditional forms of criterion-related validity evidence such as predictive and concurrent validity, and evidence based on validity and consequences of testing refers to the evaluation of the intended and unintended consequences associated with a testing program.

This technical report summarizes development and performance of the test instrument itself, addressing test content, response processes, internal structure, and other variables. Other elements addressing testing consequences are not reported within this report and may be addressed in future as supplemental research projects or third-party studies.



### 10.3. Evidentiary Validity Framework

Table 10.1 presents an overview of the validity components covered in this technical report. Table 10.2 – Table 10.5 then examine the types of evidence available for each intended purpose of the NSCAS Summative assessments.

**Table 10.1. Sources of Validity Evidence for Each NSCAS Test Purpose**

Test Purpose	Sources of Validity Evidence			
	Test Content	Response Processes	Internal Structure	Relations to Other Variables
1. Measure and report Nebraska students' depth of achievement regarding Nebraska's standards.	✓	✓	✓	✓
2. Report if student achievement is sufficient academic proficiency in ELA and Mathematics to be on track for achieving college readiness.	✓	✓	✓	
3. Measure students' annual progress toward college and career readiness in ELA and Mathematics.	✓	✓	✓	
4. Inform teachers how student thinking differs along different areas of the scale as represented by the ALDs as information to support instructional planning.	✓	✓	✓	
5. Assess students' construct relevant achievement in ELA, Mathematics, and Science for all students and subgroups of students.	✓	✓	✓	

**Table 10.2. Sources of Validity Evidence based on Test Content**

Test Purpose	Summary of Evidence	Tech Report Sections
1. Measure and report Nebraska students' depth of achievement regarding Nebraska's standards.	<ul style="list-style-type: none"> <li>• Bias is minimized through Universal Design and accessibility resources.</li> <li>• TOS, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity.</li> <li>• The item pool and item selection procedures adequately support the test design.</li> </ul>	2,9
2. Report if student achievement is sufficient academic proficiency in ELA and Mathematics to be on track for achieving college readiness.	<ul style="list-style-type: none"> <li>• Nebraska's College and Career Ready Standards are based on skills leading to college and career readiness across grades.</li> <li>• TOS, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity.</li> </ul>	2
3. Measure students' annual progress toward college and career readiness in ELA and Mathematics.	<ul style="list-style-type: none"> <li>• Nebraska's College and Career Ready Standards are based on skills leading to college and career readiness across grades.</li> <li>• TOS, passage specifications and item specifications are aligned to grade-level content, process skills, and associated cognitive complexity.</li> </ul>	2

<b>Test Purpose</b>	<b>Summary of Evidence</b>	<b>Tech Report Sections</b>
4. Inform teachers how student thinking differs along different areas of the scale as represented by the ALDs as information to support instructional planning.	<ul style="list-style-type: none"> <li>• TOS, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity.</li> <li>• TOS and ALDs were developed in consultation with Nebraska educators.</li> <li>• Reporting categories align with the structure of the Nebraska standards to support the interpretation of the test results.</li> </ul>	2,4,7
5. Assess students' construct relevant achievement in ELA, Mathematics, and Science for all students and subgroups of students.	<ul style="list-style-type: none"> <li>• Bias is minimized through Universal Design and accessibility resources.</li> <li>• DIF analysis completed for all items across all required sub-groups.</li> <li>• Assessments are administered with appropriate accommodations.</li> </ul>	2,3,6,9

**Table 10.3. Sources of Validity Evidence based on Response Process**

<b>Test Purpose</b>	<b>Summary of Evidence</b>	<b>Tech Report Sections</b>
1. Measure and report Nebraska students' depth of achievement regarding Nebraska's standards.	<ul style="list-style-type: none"> <li>• Bias is minimized through Universal Design and accessibility resources.</li> <li>• TOS, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity.</li> <li>• Achievement levels were set consistent with best practice.</li> </ul>	2
2. Report if student achievement is sufficient academic proficiency in ELA and Mathematics to be on track for achieving college readiness.	<ul style="list-style-type: none"> <li>• TOS, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity.</li> <li>• Achievement levels were vertically articulated.</li> </ul>	2
3. Measure students' annual progress toward college and career readiness in ELA and Mathematics.	<ul style="list-style-type: none"> <li>• TOS, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity.</li> <li>• Achievement levels were vertically articulated.</li> </ul>	2
4. Inform teachers how student thinking differs along different areas of the scale as represented by the ALDs as information to support instructional planning.	<ul style="list-style-type: none"> <li>• TOS, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity.</li> <li>• Range and Policy ALDs were developed in consultation with Nebraska educators with the goal of providing information to Nebraska educators.</li> </ul>	2
5. Assess students' construct relevant achievement in ELA, Mathematics, and Science for all students and subgroups of students.	<ul style="list-style-type: none"> <li>• Bias is minimized through Universal Design and accessibility resources.</li> <li>• DIF analysis completed for all items across all required sub-groups.</li> <li>• Assessments are administered with appropriate accommodations.</li> </ul>	2,3,6,9

**Table 10.4. Sources of Validity Evidence based on Internal Structure**

<b>Test Purpose</b>	<b>Summary of Evidence</b>	<b>Tech Report Sections</b>
1. Measure and report Nebraska students' depth of achievement regarding Nebraska's standards.	<ul style="list-style-type: none"> <li>• The assessment supports precise measurement and consistent classification.</li> <li>• Achievement levels were set consistent with best practice.</li> </ul>	6,8,9
2. Report if student achievement is sufficient academic proficiency in ELA and Mathematics to be on track for achieving college readiness.	<ul style="list-style-type: none"> <li>• Scale is vertically articulated.</li> <li>• Achievement levels were vertically articulated.</li> </ul>	6, 7
3. Measure students' annual progress toward college and career readiness in ELA and Mathematics.	<ul style="list-style-type: none"> <li>• The assessment supports precise measurement and consistent classification to support analysis and reporting of longitudinal data.</li> <li>• Scale is vertically articulated.</li> <li>• Achievement levels were vertically articulated.</li> </ul>	6,7,9
4. Inform teachers how student thinking differs along different areas of the scale as represented by the ALDs as information to support instructional planning.	<ul style="list-style-type: none"> <li>• Range and Policy ALDs were developed in consultation with Nebraska educators with the goal of providing information to Nebraska educators.</li> <li>• Reporting categories align with the structure of the Nebraska standards to support the interpretation of the test results.</li> <li>• Items aligned with ALDs to support item writing processes.</li> </ul>	2,7
5. Assess students' construct relevant achievement in ELA, Mathematics, and Science for all students and subgroups of students.	<ul style="list-style-type: none"> <li>• The assessment supports precise measurement and consistent classification for all students.</li> <li>• DIF analysis completed for all items across all required subgroups.</li> </ul>	6,9

**Table 10.5. Sources of Validity Evidence based on Other Variables**

<b>Test Purpose</b>	<b>Summary of Evidence</b>	<b>Tech Report Sections</b>
1. Measure and report Nebraska students' depth of achievement regarding Nebraska's standards.	<ul style="list-style-type: none"> <li>• Correlations with MAP Growth are high.</li> </ul>	8
2. Report if student achievement is sufficient academic proficiency in ELA and Mathematics to be on track for achieving college readiness.		

Test Purpose	Summary of Evidence	Tech Report Sections
3. Measure students' annual progress toward college and career readiness in ELA and Mathematics.	<ul style="list-style-type: none"> <li>A summary of score differences between 2018 and 2019 indicates that, overall, students are increasing scores in subsequent grades.</li> </ul>	8
4. Inform teachers how student thinking differs along different areas of the scale as represented by the ALDs as information to support instructional planning.		
5. Assess students' construct relevant achievement in ELA, Mathematics, and Science for all students and subgroups of students.		

#### 10.4. Interpretive Argument Claims

The test scores for the 2019 NSCAS support their intended purpose, and the interpretation of the test scores after the careful development of the Reporting ALDs support that the test scores describe where the students were in their learning at the end of the year based on the Nebraska College and Career Ready standards. The claims to support this documented in the technical report are shown in Table 10.6.

**Table 10.6. Interpretive Argument Claims, Evidence to Support the Essential Validity Elements**

Arguments	Tech Report Section(s)	Evidence
Careful test and item development through iteration occurred to ensure that the test measured the College and Career Ready standards.	2. Test Design and Development	Description of the development and review process for item, passage, and test
Test score interpretations are comparable across students.	9. Reliability 6. Psychometric Analyses	Simulations, analysis of test information, conditional standard errors of measurement, classification accuracy, and reliability estimates; blueprint comparability across students; item analysis, IRT model, vertical scaling and equating procedures
Test administrations were secure and standardized.	3. Test Administration and Security	Test administration procedures, including administration training, test accommodations, test security, and availability of help desk during testing window
Scoring was standardized and accurate.	4. Scoring and Reporting	Scoring rules and procedures; quality control of operational scoring

Arguments	Tech Report Section(s)	Evidence
Achievement standards were rigorous and technically sound.	7. Standard Setting	Documentation of the Mathematics standard setting procedures and ELA cut score review process, including the methodology, identification of workshop participants, and implementation process, and ALD development and validation
Assessments were accessible to all students and fair across student subgroups.	3. Test Administration and Security 6. Psychometric Analyses	Accommodation policy and implementation, sensitivity review, availability of translations, and DIF analyses

### 10.5. NSCAS Validity Argument

The test development and technical quality of the NSCAS Summative assessments supports the intended test score interpretations that are provided through the Reporting ALDs and scale scores. The TOS, passage specifications, item specifications, and ALD development process show that the NSCAS Summative assessments are aligned to grade-level content. For ELA and Mathematics, there is evidence that the student response processes associated with cognitive complexity specified in the standards and TOS is behaving as intended. As an added dimension for adaptive testing, the NSCAS Summative ELA and Mathematics assessments demonstrated that the tests administered to students conform to the TOS during the constraint-based engine simulation studies and post-hoc analyses.

The item pool and item selection procedures used for the adaptive administration adequately support the test design and TOS. Content experts developed expanded item types that allow response processes to reveal skills and knowledge. All items were carefully reviewed through multiple cycles of the item development process for ambiguity, bias, sensitivity, irrelevant clues, and inaccuracy to ensure the fit between the construct and the nature of performance.

NSCAS test scores are suitable for use in accountability systems. Reporting category scores indicate directions for gaining further instructional information through the interim system or classroom observation. The assessment also supports precise measurement and consistent classification for all students. Achievement levels were vertically articulated, beginning with writing ALDs and continuing through a rigorous process of setting achievement criteria. The vertical scale was constructed to provide measurement across grades, facilitating estimates of progress toward career and college readiness for ELA and Mathematics.

To demonstrate the internal structure of the NSCAS Summative assessments, this report includes principal component analysis (PCA) that shows one dominant dimension, as well as indices of measurement precision such as test reliability, classification accuracy, CSEMs, test information, and DIF. The high correlations between NSCAS and MAP Growth show a strong relationship between the two test scores and provide concurrent evidence based on other variables. Future studies may include a predictive validity study using ACT or SAT, as well as a concurrent validity study using NAEP.

Studies for evidence based on consequences of testing have not been included within the scope of work undertaken to date by NWEA. The evidence may be added in future studies, such as evaluation of the effects of testing on instruction, evaluation of the effects of testing on issues such as high school dropout rates, analyses of students' opportunity to learn, and analyses of changes in textbooks and instructional approaches (SBAC, 2016). The evaluation of unintended consequences may include changes in instruction, diminished morale among teachers and students, increased pressure on students leading to increased dropout rates, or the pursuit of college majors and careers that are less challenging (SBAC, 2016).

Teacher surveys or focus groups can be used to collect information regarding the use of the tests and how the tests impacted the curriculum and instruction. A better understanding of the extent to which performance gains on assessments reflect improved instruction and student learning, rather than more superficial interventions such as narrow test preparation activities, would also provide evidence based on consequences of test use. Longitudinal test data along with additional information collected from Nebraska educators (e.g., information on understanding of learning standards, motivation and effort to adapt the curriculum and instruction to content standards, instructional practices, classroom assessment format and content, use and nature of test assessment preparation activities, professional development) would allow for meaningful analyses and interpretations of the score gain and uniformity of standards, learning expectations, and consequences for all students.

## References

- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265–289.
- CRESST. (2015, June). *Simulation-based evaluation of the Smarter Balanced summative assessments*. National Center for Research on Evaluation, Standards, & Student Testing. Retrieved from <https://portal.smarterbalanced.org/library/en/simulation-based-evaluation-of-the-smarter-balanced-summative-assessments.pdf>.
- Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Group/Thompson Learning.
- Data Recognition Corporation (DRC). (2017). *Spring 2018 Nebraska State Accountability (NeSA) ELA, mathematics, and science technical report*. Retrieved from <https://cdn.education.ne.gov/wp-content/uploads/2017/11/Final-NeSA-2017-Technical-Report.pdf>.
- Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (ETS Research Report No. ETS-RR-91-47). Princeton, NJ: Educational Testing Service.
- EdMetric. (2018a). *Nebraska Student-Centered Assessment System – mathematics standard setting technical report*. Report provided to NDE.
- EdMetric. (2018b). *Nebraska Student-Centered Assessment System – English language arts cut score review technical report*. Report provided to NDE.
- EdMetric. (2019). *Alignment study for Nebraska Student-Centered Assessment System, mathematics grades 3–8*. Report provided to NDE.
- Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice and a proposed framework. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). New York: Routledge.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Hingham, MA: Kluwer-Nijhoff Pub.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, and H. I. Brown (Eds.), *Test Validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Huff, K, Warner, Z., & Schweid, J. (2016). Large-scale standards based assessments of educational achievement. In A. A. Rupp & J.P. Leighton, (Eds). *The handbook of cognition assessment: Frameworks, methodologies, and applications*, pp. 399–426.



- Huynh, H. (2000). Guidelines for Rasch linking for PACT. Memorandum to Paul Sandifer on June 18, 2000. Columbia, SC: Available from Author.
- Huynh, H., & Rawls, A. (2009). A comparison between Robust z and 0.3-logit difference procedures in assessing stability of linking items for the Rasch model. In E. V. Smith, Jr., & G. E. Stone (Eds.) *Applications of Rasch measurement in criterion-referenced testing*. (pp. 429–442). Maple Grove, MN: JAM Press.
- Huynh, H., & Meyer, P. (2010). Use of Robust Z in detecting unstable items in item response theory models. *Practical Assessment, Research & Evaluation*, 15(2), 1–8.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
- Linacre, J. M. (2015). Winsteps® Rasch measurement computer program (V3.91.0.0). Beaverton, OR: winsteps.com.
- Linn, R. L. (2006). Following the Standards: Is it time for another revision? *Educational Measurement: Issues and Practice*, 25(3), 54–56.
- Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- McDonald, R. P. (1979). The structural analysis of multivariate data: A sketch of general theory. *Multivariate Behavioral Research*, 14, 21–38.
- Mulkey, J., Maynes, D., & Scott, M. (2019). *Data forensics report: Nebraska Student-Centered Assessment System, grades 3–8 English language arts and math, and grades 5 & 8 science, spring 2019 administration*. Midvale, UT: Caveon.
- Nebraska Department of Education (NDE). (2018, December). *Nebraska Student-Centered Assessment System (NSCAS) summative & alternate accessibility manual*. Retrieved from <https://cdn.education.ne.gov/wp-content/uploads/2019/01/NSCAS-Summative-and-Alternate-Accessibility-Manual-12.10.pdf>.
- NWEA. (2019a, January). *2018–2019 CAT engine simulation report for the Nebraska ELA and mathematics assessments*. Report provided to NDE. Portland, OR: NWEA.
- NWEA (2019b, May). *2019 operational CAT engine evaluation report for the NSCAS ELA and mathematics assessments*. Report provided to NDE. Portland, OR: NWEA.
- NWEA. (2019c, May). *2019 NSCAS science pilot technical report*. Report provided to NDE. Portland, OR: NWEA.
- NWEA. (2019d, May). *NSCAS science ALD development report*. Report provided to NDE. Portland, OR: NWEA.
- NWEA. (2019e). *2019 vertical scale evaluation report for the NSCAS summative ELA and mathematics assessments*. Portland, OR: NWEA.



- Patz, R.J. (2007). *Vertical scaling in standards-based educational assessment and accountability systems*. Paper prepared for the Technical Issues in Large Scale Assessment State Collaborative on Assessment and Student Standards of the Council of Chief State School Officers.
- Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments, *Educational Psychologist* 51(1), p. 59–81.
- Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 221–261). New York: American Council on Educational and Macmillan.
- Rasch, G. (1960, 1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogiske Institut.
- Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement*, 18(3), 229–244.
- Schneider, M. C., Huff, K. L., Egan, K. L., Gaines, M. L., & Ferrara, S. (2013). Relationships among item cognitive complexity, contextual response demands, and item difficulty: Implications for achievement level descriptors. *Educational Assessment*, 18(2), 99–121.
- Schneider, M. C., & Johnson, R. L. (2018). *Creating and implementing student learning objectives to support student learning and teacher evaluation*. Under contract. Taylor and Francis.
- Smarter Balanced Assessment Consortium (SBAC). (2016). *Smarter Balanced Assessment Consortium: 2014-15 technical report*. Retrieved from <https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf>.
- Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, 2(1), 66–78.
- U.S. Department of Education (2018). *A state's guide to the U.S. Department of Education's assessment peer review process*. Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education.
- Webb, N. (1997). *Criteria for alignment of expectations and assessments on mathematics and science education*. (Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6). Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.
- Webb, N. L. (1999). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. Washington, DC: Council of Chief State School Officers.
- Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, 20(1), 7- 25.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14(2), 97–116.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187–213.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, 30(3), 233–251.