



**Spring 2022 NSCAS Growth
ELA, Mathematics, and Science Technical Report**

Table of Contents

Executive Summary	10
Section 1: Introduction	13
1.1. NSCAS Overview.....	13
1.2. Background.....	14
1.3. Schedule of Major Events	14
1.4. Building a Validity Argument	15
1.5. Intended Purposes and Uses of Test Results	16
1.6. Theory of Action.....	17
Section 2: Test Design and Development	19
2.1. Test Designs.....	19
2.2. Academic Content Standards.....	21
2.3. Blueprints.....	21
2.4. Item Types.....	21
2.5. Depth of Knowledge (DOK).....	22
2.6. ALD Development.....	23
2.6.1. Policy ALDs.....	23
2.6.2. Range ALDs.....	23
2.6.2.1. ELA and Mathematics.....	24
2.6.2.2. Science.....	26
2.6.3. Reporting ALDs.....	27
2.7. ELA Passage Development	29
2.8. Item Development.....	29
2.8.1. Item Specifications	29
2.8.2. ELA and Mathematics	30
2.8.3. Science	30
2.8.4. Item Retirement.....	30
2.9. Content Alignment	30
2.9.1. Alignment and Adaptive Testing.....	31
2.9.2. 2019 Mathematics Alignment Study	31
2.10. Universal Design.....	32
2.11. Sensitivity and Fairness	32
2.12. Test Construction (ELA and Mathematics)	33
2.12.1. Fixed-Forms	33
2.12.2. MAP Growth Item Selection	34
2.13. Data Review	34
Section 3: Test Administration and Security	36
3.1. User Roles and Responsibilities.....	37
3.2. Administration Training	37
3.3. Item Type Samplers.....	38
3.4. Accommodations and Accessibility Features	39
3.5. User Acceptance Testing (UAT).....	41
3.6. Student Participation.....	42

3.6.1. Paper-Pencil Participation Criteria.....	42
3.6.2. Participation of English Language Learners (ELLs)	42
3.6.3. Participation of Recently Arrived Limited English Proficient Students	43
3.7. Test Security.....	43
3.7.1. Test Security	44
3.7.1.1. Physical Warehouse Security	44
3.7.1.2. Secure Destruction of Test Materials	44
3.7.1.3. Shipping Security.....	44
3.7.1.4. Electronic Security of Test Materials and Data.....	44
3.7.2. Caveon Test Security	44
3.8. Partner Support.....	44
Section 4: Scoring and Reporting.....	46
4.1. Scoring Rules	46
4.2. Score Reporting Methods	47
4.3. Report Summary.....	48
4.3.1. Report Verification.....	50
Section 5: Constraint-Based Engine.....	52
5.1. Overview.....	52
5.2. Engine Simulations and Evaluation	53
5.2.1. Evaluation Criteria	54
5.2.2. Blueprint Constraint Accuracy	55
5.2.3. Item Exposure Rates.....	60
5.2.4. Score Precision and Reliability	62
5.3. Engine Simulations: Science Field Test	73
Section 6: Psychometric Analyses	78
6.1. Number of Students Included in the Analyses.....	78
6.2. Classical Item Analyses	79
6.2.1. Item Difficulty (P-value)	79
6.2.2. Item Discrimination (Item-Total Correlation)	80
6.2.3. Item Suppression	82
6.3. Differential Item Functioning (DIF).....	83
6.3.1. Logistic Regression (LR) DIF Method.....	83
6.3.2. Mantel-Haenszel (MH) DIF Methods	85
6.3.3. DIF Results	86
6.4. IRT Calibration.....	91
6.4.1. Summary IRT Item Statistics	92
6.5. Stability Check (ELA and Mathematics)	93
6.6. Science Measurement Model.....	95
6.7. Scaling.....	96
Section 7: Standard Setting.....	99
7.1. ELA and Mathematics.....	99
7.1.1. Overview	99
7.1.2. Meeting Process	100
7.1.3. ALD Revision	100

7.1.4. ID Matching Method	101
7.2. Science	101
7.2.1. Extended Angoff methodology	102
7.2.2. Meeting Process	102
7.3. Final Results	102
Section 8: Test Results	104
8.1. Demographics and Accommodations	104
8.2. Administration Mode (Online vs. Paper-Pencil)	110
8.3. Testing Time	111
8.4. Achievement Level Distributions	113
8.5. Descriptive Statistics of Scale Scores	113
8.6. Reporting Category Correlations	114
8.7. Correlations with MAP Growth	118
Section 9: Reliability	120
9.1. Marginal Reliability	120
9.2. Conditional Standard Error of Measurement (CSEM)	121
9.3. Classification Accuracy	123
9.4. Reliability for Fixed Forms (Science)	125
Section 10: Validity	127
10.1. Intended Purposes and Uses of Test Scores	127
10.2. Sources of Validity Evidence	128
10.3. Evidentiary Validity Framework	129
10.4. Interpretive Argument Claims	132
10.5. NSCAS Validity Argument	133
References	135
Appendix A: Data Review Cheat Sheet	138
Appendix B: Summary <i>P</i> -values by Item Type	143
Appendix C: Summary Item-Total Correlations by Item Type	153
Appendix D: Achievement Level Distributions & Scale Score Descriptive Statistics by Demographics	163
Appendix E: Marginal Reliability by Demographics	170
Appendix F: Scatterplots for Scale Score CSEM	177

List of Tables

Table 1.1. Schedule of Major Events for the Spring 2021 Administration	15
Table 2.1. NSCAS Growth in 2021–2022	19
Table 2.2. Number of Items and Points Per Test	20
Table 2.3. Online Item Types	22
Table 2.4. Task Development Results—Science	30
Table 2.5. Data Review Flagging Criteria—Multiple-Choice Items	34
Table 2.6. Data Review Flagging Criteria—Non-Multiple-Choice Items	35
Table 2.7. Data Review Results	35

Table 3.1. User Roles and Responsibilities	37
Table 3.2. Test Administration Workshop Dates and Participation	38
Table 3.4. Accommodations and Universal Features	39
Table 3.5. Partner Support Communication Options	45
Table 3.6. Number of NSCAS Cases to Partner Support in 2021–2022	45
Table 4.1. Attemptedness Rules for Scoring	46
Table 4.2. MLE Scoring.....	46
Table 4.3. Score range (LOSS and HOSS) for NSCAS scale score and estimated RIT score...	47
Table 4.4. Achievement Level Descriptions.....	47
Table 4.5. Reporting Categories	48
Table 4.6. Non-Tested Codes (NTCs).....	49
Table 5.1. Blueprint Constraint by Reporting Category—Winter Simulations.....	55
Table 5.2. Blueprint Constraint by Reporting Category—Winter Engine Evaluation	56
Table 5.3. Blueprint Constraint by Reporting Category—Spring Simulations.....	57
Table 5.4. Blueprint Constraint by Reporting Category—Spring Engine Evaluation	59
Table 5.5. Item Exposure Rates—Winter Simulations	60
Table 5.6. Item Exposure Rates—Winter Engine Evaluation.....	61
Table 5.7. Item Exposure Rates—Spring Simulations	61
Table 5.8. Item Exposure Rates—Spring Engine Evaluation.....	62
Table 5.9. Mean Bias of the NSCAS Ability Estimation (True– Estimated)—Winter Simulations	63
Table 5.10. Mean Bias of the NSCAS Ability Estimation (True–Estimated)—Spring Simulations	64
Table 5.11. Score Precision and Reliability, Items Contributed to NSCAS—Winter Simulations	66
Table 5.12. Score Precision and Reliability, Items Contributed to NSCAS—Winter Engine Evaluation.....	67
Table 5.13. Score Precision and Reliability, Items Contributed to NSCAS—Spring Simulations	68
Table 5.14. Score Precision and Reliability, Items Contributed to NSCAS—Spring Engine Evaluation.....	70
Table 5.15. SEM by Deciles for NSCAS Scores—Winter Simulations.....	71
Table 5.16. SEM by Deciles for NSCAS Scores—Winter Engine Evaluation.....	72
Table 5.17. SEM by Deciles for NSCAS Scores—Spring Simulations.....	72
Table 5.18. SEM by Deciles for NSCAS Scores—Spring Engine Evaluation.....	73
Table 5.19. General Population Demographic Distribution	74
Table 5.20. Demographic Distribution by Form—Grade 5 (Simulation)	74
Table 5.21. Demographic Distribution by Form—Grade 8 (Simulation)	75
Table 5.22. Demographic Distribution by Form—Grade 5 (Engine Evaluation)	75
Table 5.23. Demographic Distribution by Form—Grade 8 (Engine Evaluation)	76
Table 6.1. Number of Students Included in the Psychometric Analyses	78
Table 6.2. Summary <i>P</i> -values—Operational Items.....	79
Table 6.3. Summary <i>P</i> -values—Field Test Items	80
Table 6.4. Summary Item-Total Correlations—Operational Items.....	81
Table 6.5. Summary Item-Total Correlations—Field Test Items	81
Table 6.6. Flagging Criteria for MC Items.....	82
Table 6.7. Flagging Criteria for Partial-Credit Items.....	82
Table 6.8. 2021 NSCAS Items to be Suppressed.....	82

Table 6.9. Focal and Reference Groups for Gender- and Ethnicity-Based DIF	83
Table 6.10. LR DIF Categories.....	85
Table 6.11. MH DIF Categories for Dichotomous Items	86
Table 6.12. MH DIF Categories for Polytomous Items.....	86
Table 6.13. LR DIF Results—Field Test Items (ELA/Mathematics)	87
Table 6.14. LR UIDIF Results—Field Test Items (ELA/Mathematics).....	89
Table 6.15. MH DIF Results—Operational Items (Science).....	91
Table 6.16. MH DIF Results—Field Test Items (Science)	91
Table 6.17. Summary IRT Item Statistics—Operational Items.....	92
Table 6.18. Summary IRT Item Statistics—Field Test Items.....	92
Table 6.19. Scale Score Difference Between Pre-equated and Post-equated score	94
Table 6.20. Achievement Level Distributions.....	94
Table 6.21. Score Range (LOSS and HOSS) and Assigned Score	97
Table 6.22. Conversion of Theta to Scale Scores (ELA/Mathematics)	98
Table 6.23. Conversion of Theta to Scale Scores (Science).....	98
Table 7.1. Final Approved Cut Scores and Impact Data.....	103
Table 8.1. Number of Students Tested by Demographics and Accommodations—Grade 3	104
Table 8.2. Number of Students Tested by Demographics and Accommodations—Grade 4	105
Table 8.3. Number of Students Tested by Demographics and Accommodations—Grade 5	106
Table 8.4. Number of Students Tested by Demographics and Accommodations—Grade 6	107
Table 8.5. Number of Students Tested by Demographics and Accommodations—Grade 7	108
Table 8.6. Number of Students Tested by Demographics and Accommodations—Grade 8	109
Table 8.7. Number of Students Tested by Administration Mode.....	110
Table 8.8. Testing Time in Minutes—ELA	111
Table 8.9. Testing Time in Minutes—Mathematics.....	112
Table 8.10. Testing Time in Minutes—Science	112
Table 8.11. Achievement Level Distributions.....	113
Table 8.12. Scale Score Descriptive Statistics	114
Table 8.13. Reporting Category Correlations—Grade 3.....	115
Table 8.14. Reporting Category Correlations—Grade 4.....	115
Table 8.15. Reporting Category Correlations—Grade 5.....	115
Table 8.16. Reporting Category Correlations—Grade 6.....	116
Table 8.17. Reporting Category Correlations—Grade 7.....	116
Table 8.18. Reporting Category Correlations—Grade 8.....	116
Table 8.19. Reporting Category Disattenuated Correlations—Grade 3.....	117
Table 8.20. Reporting Category Disattenuated Correlations—Grade 4.....	117
Table 8.21. Reporting Category Disattenuated Correlations—Grade 5.....	117
Table 8.22. Reporting Category Disattenuated Correlations—Grade 6.....	118
Table 8.23. Reporting Category Disattenuated Correlations—Grade 7.....	118
Table 8.24. Reporting Category Disattenuated Correlations—Grade 8.....	118
Table 8.25. Correlation and Descriptive Statistics of NSCAS and MAP Growth Scores	119
Table 9.1. Marginal Reliability of Scale Scores—ELA	121
Table 9.2. Marginal Reliability of Scale Scores—Mathematics.....	121
Table 9.3. Marginal Reliability of Scale Scores—Science	121
Table 9.4. Marginal Reliability: Variance	121
Table 9.5. CSEMs at the Proficient Cut Scores.....	122

Table 9.6. Mean CSEMs by Deciles.....	123
Table 9.7. Classification Accuracy by Achievement Level—ELA.....	124
Table 9.8. Classification Accuracy by Achievement Level—Mathematics.....	124
Table 9.9. Classification Accuracy by Achievement Level and Reporting Category—Science.....	125
Table 9.10. Cronbach’s Alpha (Internal Consistency) by Demographics for Science Fixed Forms	126
Table 10.1. Sources of Validity Evidence for Each NSCAS Test Purpose	129
Table 10.2. Sources of Validity Evidence based on Test Content	129
Table 10.3. Sources of Validity Evidence based on Response Process	130
Table 10.4. Sources of Validity Evidence based on Internal Structure.....	131
Table 10.5. Sources of Validity Evidence based on Other Variables	132
Table 10.6. Interpretive Argument Claims, Evidence to Support the Essential Validity Elements	132

List of Figures

Figure 1.1. Principled Test Design Process to Support Test Score Interpretations and Uses	18
Figure 2.1. Test Development Process	19
Figure 2.6. Range ALD Example: ELA Grade 3	25
Figure 3.1. NSCAS Growth Platform Student Login Screen	36
Figure 5.1. Adaptive Engine Overview	52
Figure 5.2. Shadow Test Approach.....	53

List of Abbreviations

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

ALD	achievement level descriptor
CCC	Crosscutting Concept
CCR	College and Career Readiness
DCI.....	Disciplinary Core Idea
DIF	differential item functioning
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
FT.....	field test
HL	horizontal linking
ID	Item-Descriptor
ISR.....	Individual Student Report
IEP	Individualized Education Plan

IRT	item response theory
IWW	item writer workshop
LOSS	lowest obtainable scale score
MC	multiple-choice
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
PP	paper-pencil
RAEL.....	Recently Arrived Limited English Proficient
SD	standard deviation
SEM	standard error of measurement
SEP.....	Science and Engineering Practice
SFTP	Secure File Transfer Protocol
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 2021–2022 Nebraska Student-Centered Assessment System (NSCAS) Growth in 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 (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Below is a high-level summary of each section in the technical report.

Section 1: Introduction

In Winter 2021–2022, the NSCAS assessments are administered in ELA and mathematics in Grades 3–8. In Spring 2021–2022, the NSCAS assessments are administered in English language arts (ELA) and mathematics in Grades 3–8 and in science in Grades 5 and 8. The purposes of the NSCAS assessments are to measure and report Nebraska students' depth of achievement regarding the Nebraska College and Career Ready Standards; to report if student achievement is sufficient academic proficiency to be on track for achieving college readiness; to measure students' annual progress toward college and career readiness; to inform teachers how student thinking differs along different areas of the scale as represented by the range achievement level descriptors (RALDs) 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

The Nebraska 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 design of the NSCAS assessments is based on a principled approach to test design in which the evidence needed to draw a conclusion about where a student is in their learning of content is made explicit in the RALDs and items are developed according to those evidence pieces. To fully represent the constructs being assessed by 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 2022 NSCAS testing window was scheduled from March 21 to April 29, 2022. The tests were administered online 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. All tests were scored with maximum likelihood estimation (MLE) scoring. All steps of scoring went through a quality control process. Score reports were prepared at the individual student, school, district, and state levels.

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 blueprint while providing the maximum information about the student based on the student's momentary ability estimate: the item selection for multiple feasible student-specific plans (SSPs), followed by choosing the complete SSP that maximizes guideline adherence and information. Pre-administration simulations and a post-administration evaluation study were 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. For ELA and mathematics, the stability of the NCSAS scale was evaluated and linking with MAP Growth was updated after the Spring 2022 administration. For science, a study was conducted to find the best measurement model for the Nebraska science assessments.

Section 7: Standard Setting

For ELA and mathematics, no standard setting was held in 2021–2022. 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. Standard setting took place for the new NSCAS science assessment following the first operational administration in Spring 2022.

Section 8: Test Results

More than 20,000 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. Most students completed the ELA test in 20–120 minutes, the mathematics test in 20–100 minutes, and the science test in 10–70 minutes. The percentage of students at Developing is 47–57%, 50–58%, and 29–37% for ELA, mathematics, and science, respectively. Correlation coefficients between MAP Growth and NSCAS scores for students who took both tests in Spring 2021 were calculated. 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.

Section 9: Reliability

The reliability/precision of the 2022 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 80, which is typically considered the minimally acceptable level of reliability. The overall CSEM 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 assessments. This section revisits phases and summarizes relevant evidence and a rationale in support of any test score interpretations and intended uses based on the *Standards for Educational and Psychological Testing* (AERA et al., 2014). 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).

While NSCAS assessments offer the additional benefit of reporting category scores that indicate directions for gaining further instructional information through the interim system or classroom observation, scores based on NSCAS are equally reliable and valid as the traditional end of year assessment due to the following factors. First, NSCAS assessments go through the same rigorous psychometric analyses such as test reliability, classification accuracy, CSEMs, test information, DIF, and convergent validity check, and the analysis results we have so far strongly support the reliability and validity claim of NSCAS assessments. In addition, the test development process ensures validity of the intended test score interpretations provided through the Reporting ALDs and scale score. Last but not the least, as stated in the *Standards* (AERA et al., 2014, pp. 14-15), NSCAS assessments are aligned to grade-level content and their test scores are suitable for use in accountability systems, as a result of a robust development process of table of specifications (TOS), passage and item specifications, and achievement level descriptor (ALD).

Section 1: Introduction

The purpose of this technical report is to summarize the design, development, administration, technical processes, and results of the Nebraska Student-Centered Assessment System (NSCAS) Growth assessments to support test users in evaluating the intended purposes, uses, and interpretations of the test scores. For 2021–2022, the through year model was used in English language arts (ELA) and mathematics for Grades 3–8, which were administered for Winter and Spring. Spring assessments include science for Grades 5 and 8. 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 (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, and the federal peer review requirements USDE (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 assessment, developed specifically for Nebraska and aligned to the state content area standards, is the assessment system’s criterion-referenced measure designed for the Nebraska student population in grades 3–8.

The NSCAS assessments were administered online. They included a variety of item types, including multiple-choice and technology-enhanced items. Student scores were reported as composite 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 test were placed into one of the following achievement levels based on their final test scores:

- Developing
- On Track
- Advanced

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 through Spring 2021. The Spring tests also included previously and 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. Content development for the new three-dimensional science assessment began in Summer 2018 with the pilot occurring in March 2019. A full-scale field test was also administered in Spring 2021 to gain feedback from Nebraska students on newly developed performance tasks. The new science assessments that were

aligned to the Nebraska College and Career Ready Standards for Science (NCCRS-S; NDE, 2017) were administered in Spring 2022.

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 79-760.03¹ 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 was 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 2022 was the fourth administration of the NSCAS 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 NCCRS-S was piloted in March 2019, with a full-scale field test administered in Spring 2021. Due to the COVID-19 pandemic, the Spring 2020 NSCAS administration was cancelled, delaying the operational timeline from an operational launch in Spring 2021 to Spring 2022.

To ensure a successful transition to a through-year assessment that capitalizes on the benefits of MAP Growth while also meeting the state requirements for identifying proficiency, a link was established between the NSCAS and MAP Growth scales.

1.3. Schedule of Major Events

Table 1.1 presents the major events that occurred for the 2022 NSCAS assessments, including the new science assessment. NDE involves educators throughout the development process to produce customized items and provide an invaluable professional development opportunity, including item/task writing and review meetings and achievement level descriptor (ALD) reviews.

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

Table 1.1. Schedule of Major Events for the Spring 2021 Administration

Event	Date(s)
Technical Advisory Committee Meeting	January 31, 2022
Test Administration Training	February 16–March 2, 2022
Operational Test Window	March 21–April 29, 2022
Make-up test window	May 2–May 6, 2022
District review preliminary data and submit updates	May 15–May 19, 2022
ELA Range ALD Workshop	June 6–June 10, 2022
Science task writing workshop	July 18–July 22, 2022
ELA Alignment Study Workshop	July 25–July 29, 2022
Delivery of Individual Student Reports (ISR)	September 2, 2022
Science content and bias review committee	September 13–September 15, 2022
Data Review with NDE (ELA, Mathematics, and Science)	September 2022

1.4. Building a Validity Argument

The NSCAS assessments have been developed based on a principled approach to test design that centers around range achievement level descriptors (RALDs) 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 RALDs and items are developed according to those evidence pieces (Huff et al., 2016; Egan et al., 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 et al., 2016). The purposes of a test design centered in RALDs 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

RALDs 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 RALDs 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.

NDE took a balanced approach to the development process of the NSCAS assessments. Beginning with Policy ALDs, which are high-level expectations of student achievement within each achievement level across grades, NWEA, with input from Nebraska educators, developed Range ALDs which define within-standard learning progressions describing the knowledge and skills students at each achievement level can likely 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.

Range ALD progressions were added to the item specification in the item pool and used to support field test item development. After the test blueprint was finalized, the updated item pool was used to run simulations of the CAT engine in preparation for the Student Test Event or Fixed Form assessments.

Following the test administration, cut score for the achievement levels are defined during a Cut Score Workshop or Standard Setting. Using evidence from the test scale and the adopted final cut scores, finalized versions of the Range ALDs were created and linked to the Reporting and Policy ALDs. Content interpretations were finalized after the standard setting and are used to support item specifications to ensure a stable, comparable construct over time.

With a principled approach to test design, RALDs may be viewed as the score interpretation, or the construct interpretive argument described by Kane (2013). For RALDs to be the foundation of test score interpretation, they should reflect more complex knowledge, skills, and abilities (KSAs) as the achievement levels increase (Schneider et al., 2013). As such, NDE developed RALDs 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 RALDs, 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. RALDs were developed in an iterative manner based on feedback from educators (Plake et al., 2010) with the final RALDs providing the interpretive argument regarding what test scores mean. By developing RALDs 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.

1.5. 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 the Nebraska College and Career Ready Standards
2. To report if student achievement is sufficient academic proficiency 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

Ultimately, how test scores are used is determined by Nebraska educators. However, some intended uses of the NSCAS test results include the following:

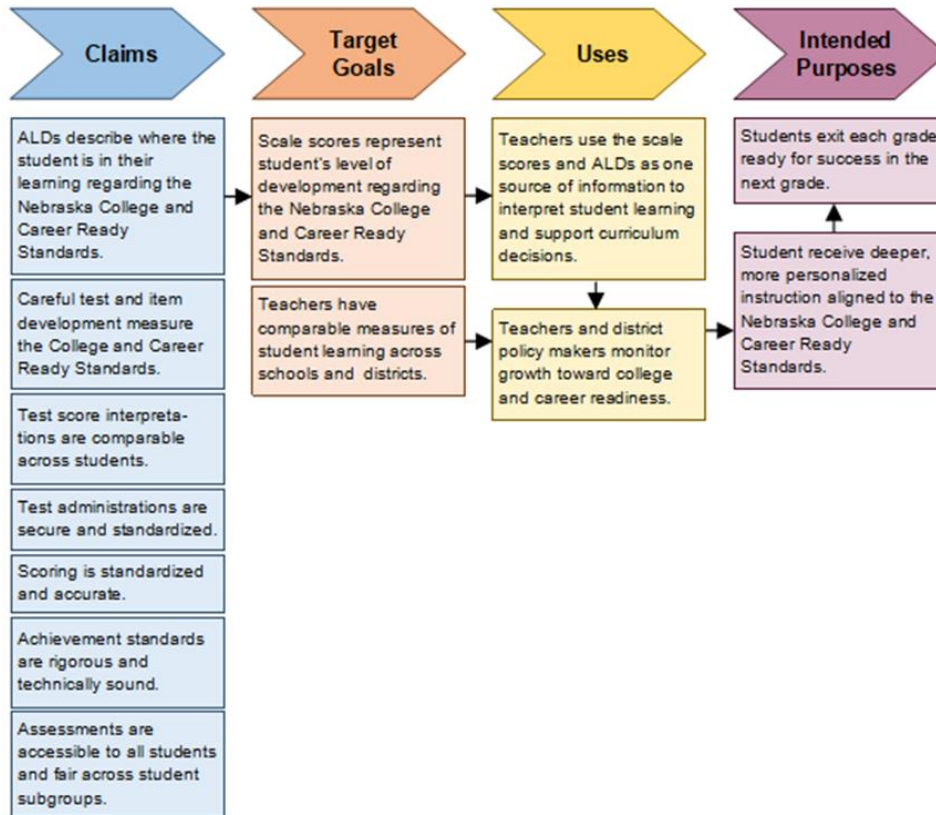
- To supplement teachers' observations and classroom assessment data
- To improve the decisions teachers make about sequencing instructional goals, designing instructional materials, and selecting instructional approaches for groups and individuals
- To identify individuals for summer school and other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

1.6. 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 1.1 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?

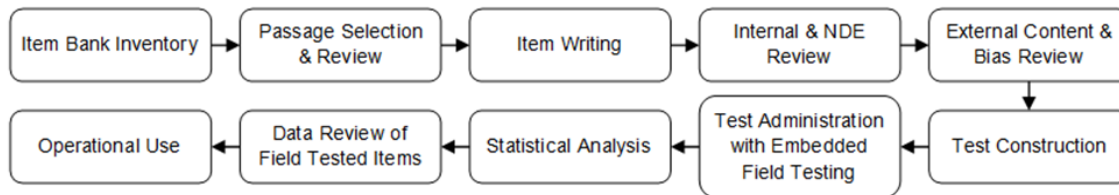
Figure 1.1. Principled Test Design Process to Support Test Score Interpretations and Uses



Section 2: Test Design and Development

This section describes the test design and development processes for the 2021–2022 NSCAS assessments. 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 RALDs. Development to support building of a bank to sufficiently support adaptive testing continued for 2022 to have enough content available to populate field test slots in the Spring 2021–2022 assessments. Previously, items were written by educators in an item writing workshop (IWW) and by independent contractors. Passages were also developed by contractors and reviewed by Nebraska educators. 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. Content development for the new three-dimensional science assessment began in Summer 2018 with the pilot occurring in March 2019, followed by the full-scale field test in the Spring 2021.

Figure 2.1. Test Development Process



2.1. Test Designs

Table 2.1 summarizes the versions of the NSCAS Growth assessments available for 2022. Table 2.2 presents the number of items and points possible.

Starting from 2021–2022 Winter, the through year model was used to narrow the blueprint in the second part of the test and thereby change the focus to become more diagnostic for students. That is, there are operational and diagnostic sections on the test.

The operational test is slightly longer in Spring 2021–2022, having a total of 45 items (i.e., 30 operational items, 8 diagnostic items, and 7 field test items), while the Winter test had a total of 40 items with 27 operational items and 13 diagnostic items.

Table 2.1. NSCAS Growth in 2021–2022

Content Area & Grade(s)	Available Assessments*				
	Online	PP	Spanish Online	Spanish PP	Breach
Winter					
ELA 3–8	Adaptive (40 total per grade, 27 OP items and 13 DO items)	One form per grade (40 OP)	Fixed (translation of PP form)	Same form as Spanish online	NA
Mathematics 3–8	Adaptive (40 total per grade, 27 OP items and 13 DO items)	One form per grade (40 OP)	Fixed (translation of PP form)	Same form as Spanish online	NA

Content Area & Grade(s)	Available Assessments*				
	Online	PP	Spanish Online	Spanish PP	Breach
Spring					
ELA 3–8	Adaptive (45 total per grade, 30 OP items, 8 DO items, and 7 FT items)	One form per grade (40 OP)	Fixed (translation of PP form)	Same form as Spanish online	Winter PP form
Mathematics 3–8	Adaptive (45 total per grade, 30 OP items, 8 DO items, and 7 FT items)	One form per grade (40 OP)	Fixed (translation of PP form)	Same form as Spanish online	Winter PP form
Science 5	20 forms (21 OP items and 4-6 FT items per form)	One form per grade (Form E, 21 OP and 4 FT items)	Fixed (translation of PP form)	Same form as Spanish online	NA
Science 8	20 forms (27 OP items and 4-7 FT items per form)	One form per grade (Form M, 27 OP and 5 FT)	Fixed (translation of PP form)	Same form as Spanish online	NA

*OP = operational. PP = paper=pencil. FT = field test.

Table 2.2. Number of Items and Points Per Test

Grade	Adaptive						Fixed					
	Total Items	NSCAS Scores		RIT Scores		FT	Total Items	NSCAS Scores		RIT Scores		FT
		Items	#Points	Items	#Points	Items	Items	Items	#Points	Items	#Points	Items
ELA (Winter)												
3	40	27	30 - 33	33	35 - 41	0	40	40	45	33	36	0
4	40	27	30 - 33	33	35 - 41	0	40	40	45	33	34	0
5	40	27	29 - 33	33	35 - 41	0	40	40	46	33	36	0
6	40	27	29 - 33	33	35 - 41	0	40	40	45	33	34	0
7	40	27	29 - 33	33	35 - 41	0	40	40	45	33	34	0
8	40	27	30 - 33	33	35 - 41	0	40	40	45	33	34	0
Mathematics (Winter)												
3	40	27	31	40	44	0	40	40	43	40	43	0
4	40	27	31	40	44	0	40	40	44	40	44	0
5	40	27	31	40	44	0	40	40	44	40	44	0
6	40	27	31	40	44	0	40	40	44	40	44	0
7	40	27	31	40	44	0	40	40	43	40	43	0
8	40	27	31	40	44	0	40	40	44	40	44	0
ELA (Spring)												
3	45	30	34 - 36	30	32 - 36	7	45	45	50	34	37	0
4	45	30	33 - 36	30	33 - 36	7	45	45	52	34	39	0
5	45	30	32 - 36	30	32 - 36	7	45	45	50	34	37	0
6	45	30	33 - 36	30	32 - 36	7	45	45	50	34	38	0
7	45	30	32 - 36	30	32 - 38	7	45	45	50	35	38	0
8	45	30	32 - 36	30	32 - 36	7	45	45	51	34	38	0
Mathematics (Spring)												
3	45	30	34 - 38	38	42 - 46	7	45	45	49	45	49	0
4	45	30	36 - 38	38	44 - 46	7	45	45	49	45	49	0
5	45	30	35 - 39	38	43 - 47	7	45	45	49	45	49	0
6	45	30	34 - 38	38	42 - 46	7	45	45	49	45	49	0
7	45	30	35 - 38	38	43 - 46	7	45	45	49	45	49	0
8	45	30	34 - 38	38	42 - 46	7	45	45	49	45	49	0
Science (Spring)												
5	25-27	21	22	NA	NA	4-6	21	17	18	NA	NA	4
8	31-34	27	33	NA	NA	4-7	21	16	20	NA	NA	5

*FT = field test.

2.2. Academic Content Standards

As stated in Nebraska Revised Statute 79-760.01² that was effective as of August 30, 2015:³

“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 that were implemented in the Spring 2019 pilot administration and will be implemented in the full-scale field test in Spring 2021.

2.3. Blueprints

The 2022 NSCAS blueprints for ELA and mathematics are embedded in the Table of Specifications (TOS) that indicate the range of test items included for each standards indicator. The adaptive test is constrained to make sure each student receives items within the identified ranges. The 2022 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. The ELA TOS for each grade is available online at <https://www.education.ne.gov/assessment/nscas-general-summative-assessment/nscas-english-language-arts-ela/>. The mathematics TOS for each grade is available online at <https://www.education.ne.gov/assessment/nscas-general-summative-assessment/nscas-mathematics/>. The blueprint for the new science assessment is available online at <https://www.education.ne.gov/wp-content/uploads/2022/08/NE-Science-Public-Blueprint-Final.pdf>. This document provides an expectation of the frequency of the DCIs, SEPs, and CCCs from the NCCRS-S. Each element from the DCIs, SEPs, and CCCs is assigned a frequency (i.e., frequent, infrequent, rare) that indicates how often the element will be assessed.

2.4. Item Types

Table 2.3 presents the item types available for the online ELA and mathematics adaptive tests. Tasks field tested in science include phenomena and a set of items (i.e., prompts) using that phenomena that may include all of the available item types.

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

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

Table 2.3. Online Item Types

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 RALDs, increases in cognitive processing complexity (e.g., DOK, difficulty, context) 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 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, in addition to other factors that may affect difficulty such as supports in the item. 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 RALDs.

To ensure that the NSCAS assessments include a deep pool of items that span a full range of cognitive levels and skills, each item in ELA and mathematics 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
- DOK 2: Skill & Concepts

- DOK 3: Strategic Thinking

Items at DOK 2 and 3 require conceptual and/or 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 are adaptive, the overall distribution of DOK for any given test event varies based on individual student achievement and other factors. In February 2018, the state adopted the policy that Developing items could be at or below the 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 RALDs and the review of 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 were developed in 2017 and reviewed by NWEA in 2018. Range ALDs for mathematics were developed in 2018, including an educator review in Spring 2018. Both ELA and mathematics Range ALDs were refined during the July 2018 standard setting and cut score review meetings. Range ALDs have also been generated for the new science assessment aligned to the NCCRS-S, beginning with an ALD workshop in May 2019.

2.6.2.1. ELA and Mathematics

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 again, 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 version of the ELA ALDs used for the Spring 2022 assessment.

In 2022, NWEA worked with NDE to update the ELA Range ALDs to the newly adopted 2021 ELA standards. NWEA first provided NDE with a draft version of the ELA Range ALDs aligned to the new ELA standards. NDE reviewed and provided feedback, which NWEA implemented. Then, Nebraska ELA educators provided feedback during a five-day, virtual Range ALD workshop held June 6–10, 2022. NWEA implemented the educators' feedback and provided a final version to NDE for their review and approval. NDE signed off on this document, which is 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 2.2 presents example Range ALDs for ELA Grade 3 for the 2014 standards that were assessed in Spring 2022. 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 2.2. Range ALD Example: 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
Reading Vocabulary					
	LA 3.1	Reading: Students will learn and apply reading skills and strategies to comprehend text.			
	LA 3.1.5	Vocabulary: 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.

2.6.2.2. Science

Before task development began in Summer 2019 for the new science assessment, it was essential to first develop the ALDs that correspond to the Developing, On Track, and CCR Benchmark achievement levels to guide development. The science Range ALDs are intended to describe students' increasingly advanced three-dimensional reasoning on tasks that require students to apply and integrate SEPs and CCCs within and among the disciplines of science. The science ALDs are available online at https://www.education.ne.gov/wp-content/uploads/2022/08/NSCAS-Science-Summative-Achievement-Level-Descriptors-ALDs-Final_8.17.2022.pdf.

The NCCRS-S may be thought of as the broad content learning goals for students at each grade level that are intended to cue instruction in ways that emphasize active scientific reasoning, but there is complexity regarding how the standards are intended to be interpreted, taught, and assessed. Indicators found in the NCCRS-S are meant only to provide examples of ways the three-dimensional standards could be integrated on an assessment. Assessment

tasks centered in the NCCRS-S are intended to measure a novel indicator based on the intersection of the grade-level DCI, CCC, and SEP through a task-based claim (i.e., students are applying SEPs to make sense of task phenomena using the intended DCIs and CCCs). Because a task-based claim represents a novel indicator, indicators can and likely will vary across alternate test forms of the state assessment. The ALDs must do two things:

1. Be specific enough to describe increasingly advanced three-dimensional reasoning and the required evidence the assessment must have that is common across alternate tasks and alternate forms of the assessment.
2. Be sufficiently generalized so that they may subsume novel indicators that change across time and potentially students.

To accommodate these needs, NDE has determined that specific science content claims (i.e., DCIs) should not be the focus of the ALDs. Instead, the grade-level content articulated in the DCIs becomes the foundation for measuring complex integration of scientific reasoning (i.e., SEPs and CCCs) and setting up phenomena that can change across alternate test forms and potentially students. Therefore, Range ALDs must reflect the progression of proficiency claims regarding how SEPs and CCCs become more sophisticated as each achievement level increases. In particular, in a three-dimensional assessment that emphasizes active scientific reasoning, the on-grade content must be extended in some way to a different phenomenon or problem so that NDE can learn about student abilities in “reasoning like a scientist.”

The DCI dimension will be embedded into the phenomena-based tasks so that the ALDs represent the three dimensions, which is represented by a consistent header in the ALDs that addresses the phenomena. For each SEP, each achievement level will need to describe the evidence NDE expects to collect to infer that a student is in that achievement level. For example, the evidence for the On Track achievement level should articulate more advanced, explicit student behaviors compared to those articulated in the Developing achievement level.

Range ALDs define the expected differences in scientific reasoning, which is useful to teachers because it aligns the evidence to be collected for each achievement level with NDE’s vision for student performance in terms of mastery of the dimensions of the NCCRS-S. Dimensional progressions are described in *A Framework for K–12 Science Education* (National Research Council, 2012), a guiding document to the NCCRS-S and to the science ALD development process. Given that NDE expects to integrate these dimensions within tasks, the dimensions cannot be viewed as independent. One dimension can influence the complexity of another dimension and therefore the difficulty of prompts along the reporting scale. Therefore, dimensions need to be integrated in the ALDs consistently to describe differences in student achievement. This also means that SEPs and CCCs need to be integrated consistently, even though the phenomena and problems used to measure those skills can vary.

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

Not applicable for the 2020–2021 administration.

2.8. Item Development

Item development for the 2021–2022 assessment administration was not required for math and ELA due to the shortened pilot. Items field tested in 2021–2022 had already been developed in prior years. Science summative task development occurred during Summer 2021 in two separate item writing workshops.

To support educators, the content teams created a variety of deliverables to support educators returning to the classroom, regardless of virtual or in-person status.

Content Specialists built pre-assessments focusing on essential work of the grade as determined by NDE in math and ELA Grades 3–8. To further support educators, the content teams created annotations for items within the item sampler related to the Range ALDs. The team selected a subset of those items to show how educators could adapt existing items to the additional Range ALD levels. The intent was to help educators adapt materials they already have rather than needing to search/buy additional materials. This work was provided to NDE in November of 2020.

The team also created an item release in paper format in both English and Spanish that was also available in large print and Braille. This could be used in addition to the item samplers to support learning within the classroom. These can be found on the NDE website listed as classroom assessments.

The science team also attended the formative science workshop from June 14–June 17, 2021, to observe the development process. Information learned was implemented in development for the 2021–2022 assessment administration.

2.8.1. Item Specifications

While there was no new item development for ELA and math in 2021–2022, previous item development ensured that each item on the NSCAS assessments should align to one standard and should follow best practices for creating test items. The RALDs 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 incorporate information from each source into a single file to provide a high-level overview for creating NSCAS 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. 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
- Range ALDs

2.8.2. ELA and Mathematics

Not applicable for the 2021–2022 administration.

2.8.3. Science

Nebraska teachers were recruited by NDE and brought together from May 26–June 2, 2021, and from July 6– July 12, 2021, for phenomena writing workshops. A total of 34 teachers participated across both workshops, ten in each grade per workshop. Table 2.4 presents the number of tasks developed at these workshops. Each task included 4–8 prompts.

Table 2.4. Task Development Results—Science

Grade	#Tasks Written	#Prompts Completed
5	17	77
8	20	98

The writers were guided in the vision of the new NSCAS science assessment and began the development process by identifying a phenomenon that met NDE’s criteria (e.g., it is observable, accessible, engaging, and explainable using grade-level appropriate science core ideas). Writers then thought about the steps needed for students to make sense of the phenomenon and identified SEPs and CCCs students would use in the sense-making process. A task was built by introducing the phenomenon in a scenario that was bimodal (e.g., it had text and graphics) followed by prompts that were minimally two-dimensional. When additional information was needed, it was presented with another mini-scenario. Each task had at least one three-dimensional prompt. The newly developed tasks and prompts were further refined by a task review committee that met from September 13–15, 2021, and consisted of NDE staff, NWEA staff, and 22 educators recruited by NDE who were not involved in writing the tasks. The tasks and prompts were reviewed for content and bias concerns.

2.8.4. Item Retirement

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. 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 in prior developments for the items used in this administration, 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 of the items tagged to the blueprint features 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 RALDs 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 RALDs 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 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 RALDs 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 RALDs within an achievement level and ask independent judges if items align to a specific RALD within an achievement level. This can be done during external content reviews with educators. Next, with the documented RALD 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–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 blueprint. The workshop was conducted virtually. The results of the study contribute to the validity evidence to support the use of 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 blueprint 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 (ELA and Mathematics)

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 2021–2022 ELA and mathematics paper-pencil forms were created based on the blueprint and statistical guidelines. The online 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).

2.12.1. Fixed-Forms

The ELA and mathematics fixed forms were created based on the blueprint and fixed-form construction specifications that included the following statistical guidelines:

- Absolute test characteristic curve (TCC) difference < 0.05
- 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 also considered the following.

- 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

Item selection was an iterative process between the psychometrics and content teams before being sent to NDE for review and approval.

2.12.2. MAP Growth Item Selection

For the through-year model, MAP Growth items were added to the item pool for the diagnostic purpose. NWEA Content team reviewed the MAP Growth items and selected the items that were aligned to NSCAS standards, conformed to NSCAS item specifications, and could contribute towards the test blueprint. Because a link was established between NSCAS ELA and MAP Growth Reading, only MAP Growth Reading items are considered. That is, MAP Growth Language Usage items were not included. Further, to include the NSCAS-like items, MAP Growth Reading items were included if they are associated with passages, and mathematics items were included if their calculator use is aligned with that of NSCAS. Specifically, reading items were removed if they were not associated with any passage or if any passages had less than three items because all NSCAS Reading Vocabulary and Reading Comprehension items are associated with passages.

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 2022 field tested items occurred virtually in September/October 2022 between NDE and NWEA. Table 2.5 and Table 2.6 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.⁴ Participants were provided a spreadsheet with the statistics for each item, as well as a data review "cheat sheet" provided in Appendix A. Table 2.7 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 2022 testing window, the number of field test items included for Data Review, the number of rejected field test items, and the number of accepted field test items.

Table 2.5. 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
IRT Difficulty or Step parameters are extremely High	≥ 4.25	Probability of getting an item correct may require extremely high ability
item fit statistics	< 0.7 or > 1.3	poor fit
<i>p</i> -value	< 0.20 or > 0.9	very difficult item
<i>p</i> -value for distractors	Distractor % $>$ Key %	More students chose a distractor than the key
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

⁴ The summaries of item analyses are included in Section 6: Psychometric Analyses of this technical report.

Table 2.6. 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
IRT Difficulty or Step parameters are extremely High	≥ 4.25	Probability of getting an item correct may require extremely high ability
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 2.7. Data Review Results

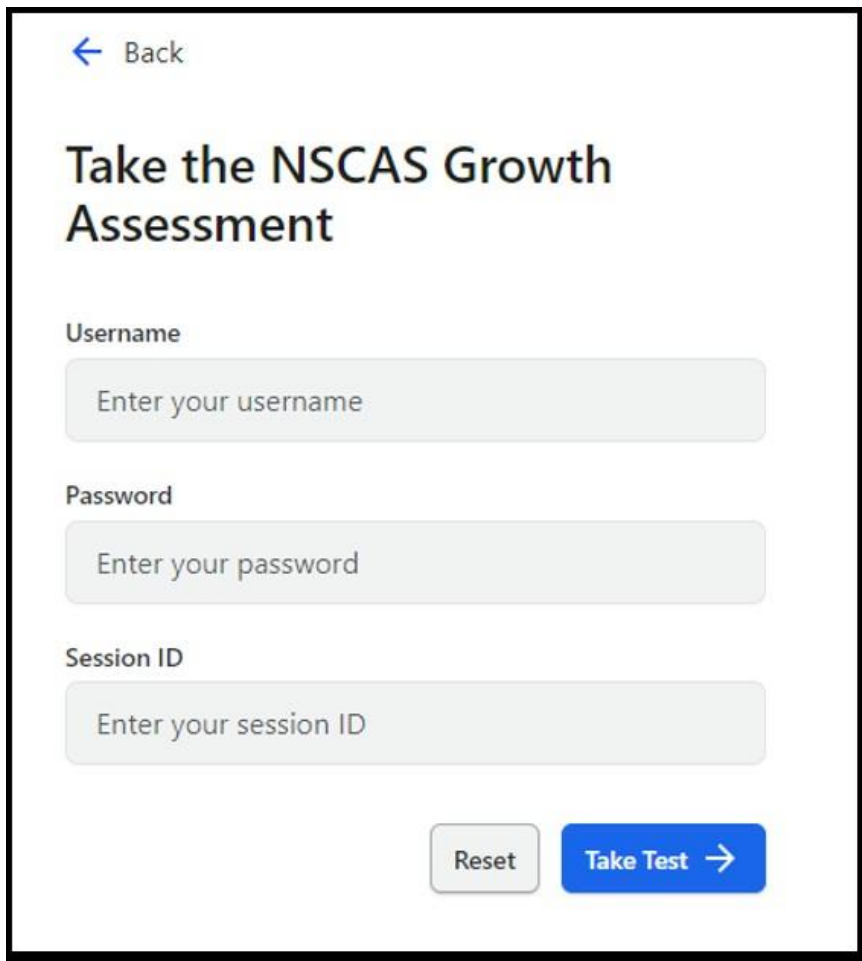
Grade	#FT Items in the Pool	#FT Items Administered	Data Review				#Total Accepted Items
			#Included	#Accepted	#Rejected /DNU	#Revise /ReFT	
ELA							
3	257	257	163	46	33	15	209
4	218	218	149	50	14	5	199
5	240	240	147	55	21	17	202
6	238	238	152	52	21	13	204
7	230	230	153	48	17	12	201
8	225	225	131	72	8	14	203
Mathematics							
3	112	112	64	30	0	18	94
4	54	54	40	10	1	3	50
5	78	78	50	14	3	11	64
6	253	253	178	33	6	36	211
7	121	121	54	22	3	42	76
8	89	89	37	16	1	35	53
Science							
5	90	90	73	6	0	11	79
8	96	96	69	4	0	23	73

Section 3: Test Administration and Security

The Spring 2022 NSCAS testing window was from March 21–April 29, 2022, and the make-up testing window was from May 2–May 6, 2022. The tests were to be untimed and administered online via the NSCAS Growth Platform. 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 NSCAS Growth Platform 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 NSCAS Growth Platform login screen. NSCAS Growth Platform works with the NWEA secure lockdown testing browser to administer the assessments, which is required for NSCAS 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. NSCAS Growth Platform Student Login Screen



← Back

Take the NSCAS Growth Assessment

Username

Password

Session ID

Reset Take Test →

The NSCAS 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. The System and Technology Guide has system requirements (p.16).⁵

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 Coordinator	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 Coordinator	Served as single points of contact at the schools for the District Assessment Coordinators 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 Coordinator might be the same person depending on the district's decisions.
Proctor	Responsible for administering the tests to students.

District Assessment Coordinators 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 Coordinators 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 the following:

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

3.2. Administration Training

In addition to district- and school-held training, NWEA, in collaboration with NDE, held two trainings for district leaders in advance of testing. The Fall 2020 regional workshops were a half-

⁵ <https://cdn.nwea.org/docs/NE/SystemTechnologyGuide.pdf>

day, virtual workshop held across multiple regions of the state. Information on the Spring administration including test sessions, accessibility, and student rostering was presented. The three test administration workshops in February 2022 were two-hour virtual sessions that provided important information on the NSCAS assessments.

Table 3.2 presents the dates and number of participants based on the registration numbers for the test administration workshop. Training presentations are available online.⁶

Table 3.2. Test Administration Workshop Dates and Participation

Date	#Participants
Feb 15, 2022	105
Feb 16, 2022	77
Feb 18, 2022	42
March 2, 2022	31

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://nwea.force.com/nweaconnection/s/nebraska-practice-tests?language=en_US. 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.

The Item Type Samplers were not adaptive. For ELA and mathematics, the Item Type Sampler has 20 items for each respective grade in a content area. The Science Item Type Sampler has 13 questions for grade 5 and 12 questions for grade 8. They were also untimed, although the estimated test-taking time for each was 40 minutes. Unlike the actual 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 assessments. They also allowed other stakeholders such as parents and administrators to experience the 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 assessment.

⁶ https://www.youtube.com/watch?v=POh_P9Tcptshhttps://cdn.education.ne.gov/wp-content/uploads/2020/10/Regional-Workshop-2020-2021-Publishing.pptx
<https://vimeo.com/user84717829/review/515870657/f69712e944>

3.4. Accommodations and Accessibility Features

Table 3.3 presents the accessibility supports available for the Spring 2022 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, 2019).

- 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 3.3. Accommodations and Universal Features

Support	Description
Embedded Accommodations	
Text-to-speech (TTS)	A student can use this feature to hear audio of the item content.
Non-Embedded Accommodations	
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.

Support	Description
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)
Embedded Universal Features	
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.
Non-Embedded Universal Features	
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.
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.

Support	Description
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 (plain or graph)	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 term 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

Testers are typically NWEA staff who are at least somewhat familiar with how the 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. Triage meetings take place to review all new reported entries 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.

NWEA staff review the data produced from the UAT to ensure it conforms to expectations for completed tests, tests assigned NTCs, incomplete tests, tests that were reset, and additional activities that occur during testing. User roles are tested for accessibility and functionality. Operational and score reports are reviewed to ensure they meet requirements.

3.6. Student Participation

All students with disabilities were expected to participate in the NSCAS. No student, including students with disabilities or required a paper assessment, 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 without accommodations.
- Students were tested on the NSCAS 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.
- 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 assessments. Districts should have reviewed the following guidelines before testing:

- In determining appropriate linguistic supports for students in the NSCAS system, districts should use the NSCAS Summative & Alternate Accessibility Manual (NDE, 2019).
- 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 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 assessments were required to complete and sign a Building Principal Security Agreement and return it to the Statewide Assessment Office by October 15, 2021. District Assessment Coordinators were required to complete and sign the District Assessment Coordinator Confidentiality of Information Agreement and return it to the Statewide Assessment Office by October 12, 2020. 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. Test Security

3.7.1.1. Physical Warehouse Security

All NWEA 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 NWEA headquarters was by badged-security access. 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 program:

- Test materials received from the printing subcontractors were stored in a room at NWEA headquarters prior to packaging and shipping to districts.

3.7.1.2. Secure Destruction of Test Materials

Printed materials for the Spring 2022 administration were not considered secure, therefore districts were authorized to destroy material locally.

3.7.1.3. Shipping Security

For district shipments, NWEA 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. Every box was assigned a unique UPS tracking number.

3.7.1.4. 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 print vendors followed secure procedures. Data files were exchanged through an SFTP site and the secure application program interface.

3.7.2. Caveon Test Security

Caveon LLC provided three services during the Spring of 2022 for the NSCAS assessment. Web patrol, data analysis, and real time incident management were provided. Caveon provided test security monitoring protocols according to the Nebraska Department of Education's test security specifications.

3.8. Partner Support

The NWEA Partner Support Services team provided implementation and technical support throughout the 2021–2022 school year for the NSCAS 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 3.4. Table 3.5 presents the number of cases presented to the Partner Support team by case type for the entire 2021–2022 school year for the NSCAS tests. More than half of the cases were related to testing (i.e., administration questions).

Table 3.4. Partner Support Communication Options

Phone Support	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.
Email Support	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.
Chat Support	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 3.5. Number of NSCAS Cases to Partner Support in 2021–2022

Case Type	#Cases	% of Total Cases
Student Mobility	28	N/A
Reports	117	11%
Navigation	225	21%
Setup and Management	392	37%
Testing	253	31%
Total	1,015	100.0

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. On average, each state team member participated in four hours of training related to Nebraska programs.

Section 4: Scoring and Reporting

The online ELA and mathematics assessments were administered adaptively via NWEA’s constraint-based engine, whereas the science assessments were administered as fixed-form. For science, each grade had 20 different forms, but the operational items are the same across forms. Also, all paper-pencil tests and all Spanish versions were administered as fixed-form.

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

*LOSS = lowest obtainable scale score. MLE = maximum likelihood estimation.

The attemptedness rule was decided based on the results of the standard error of measurement (SEM) that became relatively stable (i.e., SEM became less than 1.0 for students in the middle of true theta distribution) 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, as summarized in Table 4.2:

1. Students who took the adaptive assessment (i.e., ELA and mathematics online adaptive forms) received straight maximum likelihood estimation (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.

Table 4.2. MLE Scoring

Content Area	English Form		Spanish Form		Breach Form
	Online	Paper-pencil	Online	Paper-pencil	Paper-pencil
ELA/Mathematics	No Penalty	With Penalty	No Penalty	With Penalty	With Penalty
Science	With Penalty	With Penalty	With Penalty	With Penalty	With Penalty

4.2. Score Reporting Methods

Student performance on the NSCAS assessment is reported as a scale score and achievement level. Each content area is scaled separately. Therefore, the scale scores for one content area cannot be compared to another content area. For ELA and mathematics, NSCAS Growth reports also provide estimated RIT scores. Table 4.3. presents score range for both scores.

Table 4.3. Score range (LOSS and HOSS) for NSCAS scale score and estimated RIT score

Content Area	Grade	NSCAS Scale Score			Estimated RIT Score		
		LOSS	HOSS	Calculated LOSS*	LOSS	HOSS	Calculated LOSS*
ELA	3	2220	2840	2222	100	350	102
	4	2250	2850	2252	100	350	102
	5	2280	2860	2282	100	350	102
	6	2290	2870	2292	100	350	102
	7	2300	2880	2302	100	350	102
	8	2310	2890	2312	100	350	102
Mathematics	3	1000	1470	1002	100	350	102
	4	1010	1500	1012	100	350	102
	5	1020	1510	1022	100	350	102
	6	1030	1530	1032	100	350	102
	7	1040	1540	1042	100	350	102
	8	1050	1550	1052	100	350	102
Science	5	3000	3250	3002	–	–	–
	8	3000	3250	3002	–	--	□

* Calculated LOSS = Lowest calculated score for students with 10 or more OP items attempted.

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 content 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 4.4.

Table 4.4. Achievement Level Descriptions

Achievement Level	Description
ELA/Mathematics	
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.

Achievement Level	Description
ELA/Mathematics	
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.
Science	
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.
Advanced	Advanced learners demonstrate high levels of 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.

The reporting categories in Table 4.5 were to be used for scoring and reporting. Items were mapped to a reporting category based on the indicators. For science, reporting category scores were not provided in 2022.

Table 4.5. Reporting Categories

Content Area	Reporting Categories
ELA	<ul style="list-style-type: none"> • Reading Vocabulary • Reading Comprehension • Writing Skills
Mathematics	<ul style="list-style-type: none"> • Number • Algebra • Geometry • Data

4.3. Report Summary

The following reports were prepared for the 2022 NSCAS test administration. Examples of the reports and additional information can be found in the Interpretive Guide.⁷

- State Level
 - Student Score Data File
 - Organization Report—State level

⁷ https://connection.nwea.org/s/nebraska?language=en_US

- State Demographic Report
- Region
 - Organization Report—Region level
- District Level
 - Student Score Data File
 - Organization Report
 - District Demographic Report
- School Level
 - Organization Report—School level
 - School Roster
 - School Demographic Report
- Class/Group Level
 - Class/Group Roster
- Student Level
 - Dynamic Student Report
 - Student Growth Report
 - Individual Student Report (ISR) English
 - Individual Student Report (ISR) Spanish

ISRs show a student’s performance on the NSCAS Growth tests. Content areas are combined to produce a single ISR report for a student. ISRs are available through the NSCAS Growth platform and shipped to the districts. Some ISRs are shipped to their new fall enrollment district while others shipped to their reportable district. If a non-tested code (NTC) is applied to a content area, the student’s achievement level scores are reported as affected by the NTC, as defined in Table 4.6. If a student has an NTC of INV, PAR, STR, or UTT assigned to their test, the automatically assigned score displays with a score of the lowest scale score for that grade and content area.

Table 4.6. Non-Tested Codes (NTCs)

Code	Name/Description	Include in reports	Scoring
ALT	Alternate Assessment: Student took the NSCAS Alternate assessment and is not included in results from this testing vendor.	FALSE	No Score Provided
COV	COVID-19 Waiver: Student not tested because of an ongoing and continued concern about exposure to COVID-19.	TRUE	No Score Provided
EMW	Emergency Medical Waiver: Student was not tested because of an approved Emergency Medical Waiver.	TRUE	No Score Provided
EXP	Exception: Student exempt from testing due to certain circumstances, such as student requiring unavailable accommodation; student is attending an out-of-state facility; or testing irregularities.	FALSE	Score not included in Reports or Calculations
FTE	Not Full Time: Full-Time Equivalency is less than 51% so the student is excluded from testing.	FALSE	Score not included in Reports or Calculations

Code	Name/Description	Include in reports	Scoring
INV	Invalid: Student's assessment was invalidated, such as security breach or student refuses to finish test.	TRUE	Score as LOSS
NCE	Not Currently Enrolled: Student was not enrolled in the district/school during testing window.	FALSE	Score not included in Reports or Calculations
OTH	Other: Student was not tested for reasons not covered by other descriptions. For example, occurrence of a natural disaster.	TRUE	Score Suppressed
PAR	Parent Refusal: Student was not tested because of a formal request from parent or guardian.	TRUE	Score as LOSS
RMV	Removal: Student left the district before the test window; student is a full-time home-schooled student; or there are duplicate student records.	FALSE	Score not included in Reports or Calculations
STR	Student Refusal: Student was not tested due to student refusal to participate.	TRUE	Score as LOSS
UTT	Unable To Test: District was unable to test the student during the testing windows due to excessive absences or suspension-expulsion.	TRUE	Score as LOSS

4.3.1. Report Verification

The NSCAS report quality assurance (QA) process consisted of validating the data and reports using the scoring specifications, reporting specifications, mockups, layouts, scale score, and cut information.

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 segments 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, achievement levels)
- Historical charts and data footnotes
- NTC behavior
- Not enough items (NEI) behavior
- Sorting (sort order of the report)
- Naming conventions reports, files, and folders
- Similar data is the same across all reports

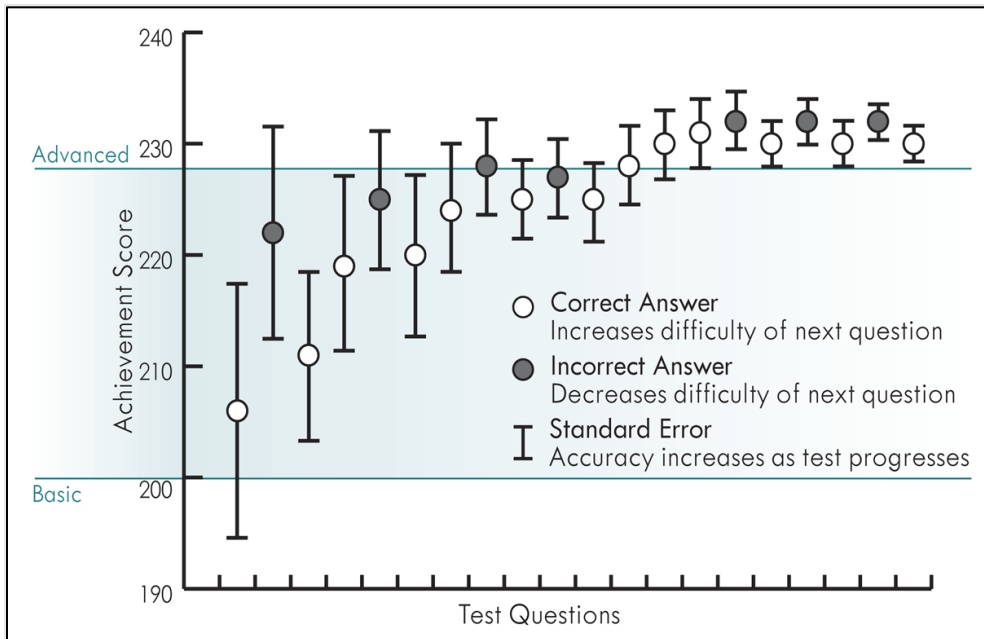
- Summation of data
- User Interface functionality

Section 5: Constraint-Based Engine

5.1. Overview

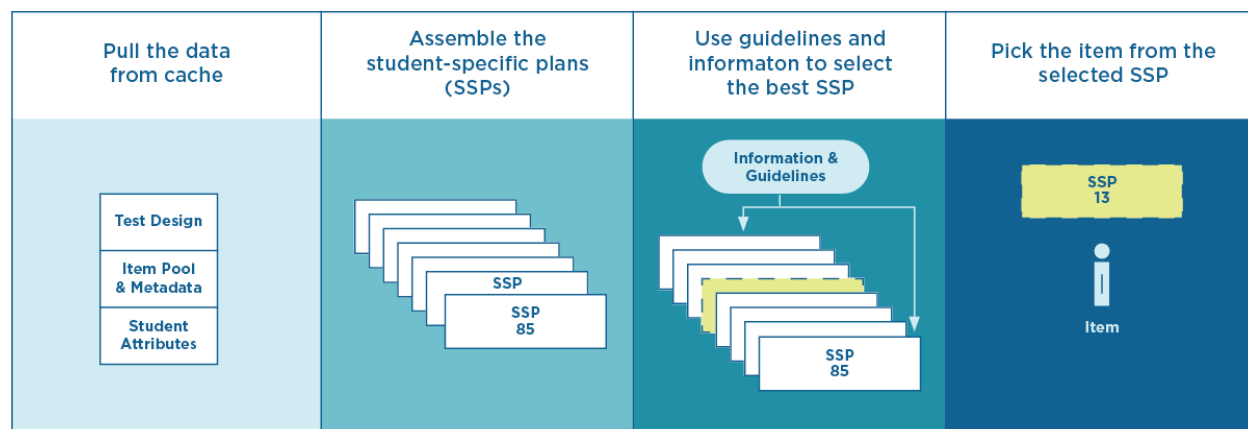
An adaptive assessment administers items to match the ability level of the student. Students receive different items based on item difficulty and their ability levels. For example, students with lower ability levels (based on their answers to previous items) receive easier items compared to students with higher ability levels who receive harder items as the test progresses. The constraint-based engine (CBE) uses the TOS and a student's momentary theta (θ) 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" constraints and nice-to-have" guidelines. The CBE has two stages of consideration as it selects the items necessary to conform to the test blueprint while providing the maximum information about the student based on the student's momentary ability estimate. The student-specific plan (SSP), similar to the shadow test approach (Van der Linden & Reese, 1998), selects items based on the required aspects of the test blueprint and the student's momentary theta, as shown in Figure 5.1. Item selection for the SSP occurs through a process of choosing multiple feasible SSPs, then choosing the complete SSP that best maximizes guideline adherence and information. Only after the best SSP has been chosen are items ordered (NWEA, 2020).

Figure 5.2. Shadow Test Approach



The following updates made for Winter:

- The through year model was used to narrow the blueprint in the second part of the test and thereby change the focus to become more diagnostic for students.
- There are operational and diagnostic sections on the test.
- Item exposure was controlled by assigning a weight to an item based on the number of times the item is seen by students.

The following updates made for Spring:

- The operational test is slightly longer in Spring 2021–2022, having a total of 45 items (i.e., 30 operational items, 8 diagnostic items, and 7 field test items), while the Winter test had a total of 40 items with 27 operational items and 13 diagnostic items.
- NWEA set population exposure targets to represent major demographic groups such as gender (female vs. male) and ethnic group (American Indian, Asian, Black, Hispanic vs. White). With a lot of field test items included (i.e., more than 200 field test items for many grades and content areas), the population exposure controls resulted in fewer items being flagged than the pseudo-random assignment that was applied in 2020.
- Longitudinal item exposure is applied. Thus, the engine would prefer not to expose items if student saw it within last 365 days.
- Each student's score from Winter simulation was used as an initial ability estimate.
- When a prior score was not available, a historical population value was used for the initial ability estimate. Specifically, the mean minus one standard deviation (SD) from 2021 NSCAS is used. In previous simulations, the midpoint between lowest obtainable scale score (LOSS) and highest obtainable scale score (HOSS) on the theta scale had been used for student's initial ability estimate in running simulations. NWEA showed the impact of initial theta on the final score estimates would be minor and students can start the test with easier items.

5.2. Engine Simulations and Evaluation

Pre-administration engine simulations and 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 operational testing window to evaluate the CBE’s item selection algorithm and estimation of student ability based on the TOS. The simulation tool used the operational CBE, 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, 2021a, 2022a).

After the testing window closed, a post-administration evaluation study was then conducted to determine whether the constraint-based engine performed as expected. Detailed information regarding all results of the post-administration evaluation study can be found in the full report (NWEA, 2022b, 2022c).

Overall, the CBE performed as it should based on the blueprint (i.e., TOS) constraints. The reporting category points had a 100% match. 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 all 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 θ_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 θ 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 θ_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:

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

where σ is defined as:

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

5.2.2. Blueprint Constraint Accuracy

Table 5.1 through Table 5.4 present the blueprint constraint results at the reporting category level for the Winter pre-administration simulation study, the Winter post-administration engine evaluation study, the Spring pre-administration simulation study, and the Spring post-administration engine evaluation study, respectively. The number of items at the reporting category level resulted in a 100% match for all grades based on the blueprint, with marginal deviation in the number of points based on the availability and selection of polytomously-scored items in Reading Comprehension.

Table 5.1. Blueprint Constraint by Reporting Category—Winter Simulations

Grade	Reporting Category	#Items			#Points		
		Min.	Max.	%Match	Min.	Max.	%Match
ELA							
3	Reading Vocabulary	6	6	100.0	6	7	100.0
	Reading Comprehension	14	14	100.0	16	17	100.0
	Writing Skills	7	7	100.0	9	9	100.0
4	Reading Vocabulary	6	6	100.0	6	7	100.0
	Reading Comprehension	14	14	100.0	16	17	88.5
	Writing Skills	7	7	100.0	9	9	100.0
5	Reading Vocabulary	6	6	100.0	6	7	100.0
	Reading Comprehension	14	14	100.0	16	17	90.0
	Writing Skills	7	7	100.0	9	9	100.0
6	Reading Vocabulary	6	6	100.0	6	7	100.0
	Reading Comprehension	14	14	100.0	16	17	93.9
	Writing Skills	7	7	100.0	9	9	100.0
7	Reading Vocabulary	6	6	100.0	6	7	100.0
	Reading Comprehension	14	14	100.0	16	17	78.0
	Writing Skills	7	7	100.0	9	9	100.0

Grade	Reporting Category	#Items			#Points		
		Min.	Max.	%Match	Min.	Max.	%Match
8	Reading Vocabulary	6	6	100.0	6	7	100.0
	Reading Comprehension	14	14	100.0	16	17	98.8
	Writing Skills	7	7	100.0	9	9	100.0
Mathematics							
3	Number	10	10	100.0	11	11	100.0
	Algebra	5	5	100.0	6	6	100.0
	Geometry	7	7	100.0	8	8	100.0
	Data	5	5	100.0	6	6	100.0
4	Number	10	10	100.0	11	11	100.0
	Algebra	6	6	100.0	7	7	100.0
	Geometry	6	6	100.0	7	7	100.0
	Data	5	5	100.0	6	6	100.0
5	Number	10	10	100.0	11	11	100.0
	Algebra	6	6	100.0	7	7	100.0
	Geometry	6	6	100.0	7	7	100.0
	Data	5	5	100.0	6	6	100.0
6	Number	7	7	100.0	8	8	100.0
	Algebra	10	10	100.0	11	11	100.0
	Geometry	5	5	100.0	6	6	100.0
	Data	5	5	100.0	6	6	100.0
7	Number	6	6	100.0	7	7	100.0
	Algebra	10	10	100.0	11	11	100.0
	Geometry	6	6	100.0	7	7	100.0
	Data	5	5	100.0	6	6	100.0
8	Number	7	7	100.0	8	8	100.0
	Algebra	7	7	100.0	8	8	100.0
	Geometry	8	8	100.0	9	9	100.0
	Data	5	5	100.0	6	6	100.0

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

Grade	Reporting Category	#Items			#Points		
		Min.	Max.	%Match	Min.	Max.	%Match
ELA							
3	Reading Vocabulary	6	6	100.0	6	7	100.0
	Reading Comprehension	14	14	100.0	16	17	100.0
	Writing Skills	7	7	100.0	9	9	100.0
4	Reading Vocabulary	6	6	100.0	6	7	100.0
	Reading Comprehension	14	14	100.0	16	17	86.3
	Writing Skills	7	7	100.0	9	9	100.0
5	Reading Vocabulary	6	6	100.0	6	7	100.0
	Reading Comprehension	14	14	100.0	16	17	87.5
	Writing Skills	7	7	100.0	9	9	100.0
6	Reading Vocabulary	6	6	100.0	6	7	100.0

Grade	Reporting Category	#Items			#Points		
		Min.	Max.	%Match	Min.	Max.	%Match
	Reading Comprehension	14	14	100.0	16	17	93.0
	Writing Skills	7	7	100.0	9	9	100.0
7	Reading Vocabulary	6	6	100.0	6	7	100.0
	Reading Comprehension	14	14	100.0	16	17	77.5
	Writing Skills	7	7	100.0	9	9	100.0
8	Reading Vocabulary	6	6	100.0	6	7	100.0
	Reading Comprehension	14	14	100.0	16	17	98.8
	Writing Skills	7	7	100.0	9	9	100.0
Mathematics							
3	Number	10	10	100.0	11	11	100.0
	Algebra	5	5	100.0	6	6	100.0
	Geometry	7	7	100.0	8	8	100.0
	Data	5	5	100.0	6	6	100.0
4	Number	10	10	100.0	11	11	100.0
	Algebra	6	6	100.0	7	7	100.0
	Geometry	6	6	100.0	7	7	100.0
	Data	5	5	100.0	6	6	100.0
5	Number	10	10	100.0	11	11	100.0
	Algebra	6	6	100.0	7	7	100.0
	Geometry	6	6	100.0	7	7	100.0
	Data	5	5	100.0	6	6	100.0
6	Number	7	7	100.0	8	8	100.0
	Algebra	10	10	100.0	11	11	100.0
	Geometry	5	5	100.0	6	6	100.0
	Data	5	5	100.0	6	6	100.0
7	Number	6	6	100.0	7	7	100.0
	Algebra	10	10	100.0	11	11	100.0
	Geometry	6	6	100.0	7	7	100.0
	Data	5	5	100.0	6	6	100.0
8	Number	7	7	100.0	8	8	100.0
	Algebra	7	7	100.0	8	8	100.0
	Geometry	8	8	100.0	9	9	100.0
	Data	5	5	100.0	6	6	100.0

Table 5.3. Blueprint Constraint by Reporting Category—Spring Simulations

Grade	Reporting Category	#Items			#Points		
		Min.	Max.	%Match	Min.	Max.	%Match
ELA							
3	Reading Vocabulary	7	7	100.0	7	8	100.0
	Reading Comprehension	15	15	100.0	17	18	96.3
	Writing Skills	8	8	100.0	10	10	100.0
4	Reading Vocabulary	7	7	100.0	7	8	100.0
	Reading Comprehension	15	15	100.0	17	18	98.0

Grade	Reporting Category	#Items			#Points		
		Min.	Max.	%Match	Min.	Max.	%Match
	Writing Skills	8	8	100.0	10	10	100.0
5	Reading Vocabulary	7	7	100.0	7	8	100.0
	Reading Comprehension	15	15	100.0	17	18	94.4
	Writing Skills	8	8	100.0	10	10	100.0
6	Reading Vocabulary	7	7	100.0	7	8	100.0
	Reading Comprehension	15	15	100.0	17	18	95.8
	Writing Skills	8	8	100.0	10	10	100.0
7	Reading Vocabulary	7	7	100.0	7	8	100.0
	Reading Comprehension	15	15	100.0	17	18	86.9
	Writing Skills	8	8	100.0	10	10	100.0
8	Reading Vocabulary	7	7	100.0	7	8	100.0
	Reading Comprehension	15	15	100.0	17	18	91.3
	Writing Skills	8	8	100.0	10	10	100.0
Mathematics							
3	Number	12	12	100.0	13	14	100.0
	Algebra	5	5	100.0	6	7	100.0
	Geometry	8	8	100.0	9	10	100.0
	Data	5	5	100.0	6	7	100.0
4	Number	11	11	100.0	12	13	100.0
	Algebra	7	7	100.0	8	9	100.0
	Geometry	7	7	100.0	8	9	100.0
	Data	5	5	100.0	6	7	100.0
5	Number	11	11	100.0	12	14	100.0
	Algebra	7	7	100.0	8	10	100.0
	Geometry	7	7	100.0	8	10	100.0
	Data	5	5	100.0	6	7	100.0
6	Number	8	8	100.0	9	10	100.0
	Algebra	10	10	100.0	11	12	100.0
	Geometry	6	6	100.0	7	8	100.0
	Data	6	6	100.0	7	8	100.0
7	Number	7	7	100.0	8	9	100.0
	Algebra	10	10	100.0	11	12	100.0
	Geometry	6	6	100.0	7	8	100.0
	Data	7	7	100.0	8	9	100.0
8	Number	8	8	100.0	9	10	100.0
	Algebra	8	8	100.0	9	10	100.0
	Geometry	8	8	100.0	9	10	100.0
	Data	6	6	100.0	7	8	100.0

Table 5.4. Blueprint Constraint by Reporting Category—Spring Engine Evaluation

Grade	Reporting Category	#Items			#Points		
		Min.	Max.	%Match	Min.	Max.	%Match
ELA							
3	Reading Vocabulary	7	7	100.0	7	8	100.0
	Reading Comprehension	15	15	100.0	17	18	96.5
	Writing Skills	8	8	100.0	10	10	100.0
4	Reading Vocabulary	7	7	100.0	7	8	100.0
	Reading Comprehension	15	15	100.0	17	18	98.4
	Writing Skills	8	8	100.0	10	10	100.0
5	Reading Vocabulary	7	7	100.0	7	8	100.0
	Reading Comprehension	15	15	100.0	17	18	94.2
	Writing Skills	8	8	100.0	10	10	100.0
6	Reading Vocabulary	7	7	100.0	7	8	100.0
	Reading Comprehension	15	15	100.0	17	18	95.5
	Writing Skills	8	8	100.0	10	10	100.0
7	Reading Vocabulary	7	7	100.0	7	8	100.0
	Reading Comprehension	15	15	100.0	17	18	86.9
	Writing Skills	8	8	100.0	10	10	100.0
8	Reading Vocabulary	7	7	100.0	7	8	100.0
	Reading Comprehension	15	15	100.0	17	18	90.8
	Writing Skills	8	8	100.0	10	10	100.0
Mathematics							
3	Number	12	12	100.0	13	14	100.0
	Algebra	5	5	100.0	6	7	100.0
	Geometry	8	8	100.0	9	10	100.0
	Data	5	5	100.0	6	7	100.0
4	Number	11	11	100.0	12	13	100.0
	Algebra	7	7	100.0	8	9	100.0
	Geometry	7	7	100.0	8	9	100.0
	Data	5	5	100.0	6	7	100.0
5	Number	11	11	100.0	12	14	100.0
	Algebra	7	7	100.0	8	10	100.0
	Geometry	7	7	100.0	8	10	100.0
	Data	5	5	100.0	6	7	100.0
6	Number	8	8	100.0	9	10	100.0
	Algebra	10	10	100.0	11	12	100.0
	Geometry	6	6	100.0	7	8	100.0
	Data	6	6	100.0	7	8	100.0
7	Number	7	7	100.0	8	9	100.0
	Algebra	10	10	100.0	11	12	100.0
	Geometry	6	6	100.0	7	8	100.0
	Data	7	7	100.0	8	9	100.0
8	Number	8	8	100.0	9	10	100.0

Grade	Reporting Category	#Items			#Points		
		Min.	Max.	%Match	Min.	Max.	%Match
	Algebra	8	8	100.0	9	10	100.0
	Geometry	8	8	100.0	9	10	100.0
	Data	6	6	100.0	7	8	100.0

5.2.3. Item Exposure Rates

Table 5.5 through Table 5.8 present the item exposure rates from the Winter pre-administration simulation study, the Winter post-administration engine evaluation study, the Spring pre-administration simulation study, and the Spring 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 the tables, Table 5.5 “Total” is the total number of items in the operational item pool. “Unused” shows the number and percentage of unused items that were never administered to students.

For the 2021–2022 administration, item exposure is being controlled by an update to a feature in the engine that assigns a weight to an item based on the number of times the item is seen by students. As the weight increases, that item is no longer preferred in the item selection SSP (student-specific plan). This feature does not prevent the item from being seen by students if it is the best item in the pool to meet the requirements for that student. Rather, this feature prefers the engine to look at additional items in the pool that might meet the requirements for the student, as well as the item that has already been exposed. The results show that this updated feature, which has been applied since Spring 2021, combined with new test design (i.e., including diagnostic items of adjacent grade), resulted in increased item pool usage, especially for ELA, compared with historical simulation results.

Table 5.5. Item Exposure Rates—Winter Simulations

Grade	#Items				Exposure Rate									
					0–20%		21–40%		41–60%		61–80%		81–99%	
	Total	Used	Unused	Unused %	N	%	N	%	N	%	N	%	N	%
ELA														
3	547	442	105	19.20	409	92.53	23	5.20	4	0.90	2	0.45	2	0.45
4	539	534	5	0.93	512	95.88	11	2.06	11	2.06	–	–	–	–
5	472	391	81	17.16	361	92.33	18	4.60	4	1.02	4	1.02	4	1.02
6	470	451	19	4.04	424	94.01	27	5.99	–	–	–	–	–	–
7	458	370	88	19.21	325	87.84	35	9.46	4	1.08	3	0.81	2	0.54
8	525	443	82	15.62	419	94.58	13	2.93	8	1.81	1	0.23	–	–
Mathematics														
3	479	477	2	0.42	470	98.53	7	1.47	–	–	–	–	–	–
4	357	357	0	0.00	334	93.56	23	6.44	–	–	–	–	–	–
5	379	376	3	0.79	367	97.61	9	2.39	–	–	–	–	–	–
6	473	470	3	0.63	464	98.72	6	1.28	–	–	–	–	–	–
7	416	412	4	0.96	403	97.82	8	1.94	1	0.24	–	–	–	–

Grade	#Items				Exposure Rate											
					0–20%		21–40%		41–60%		61–80%		81–99%		100%	
	Total	Used	Unused	Unused %	N	%	N	%	N	%	N	%	N	%		
8	356	356	0	0.00	339	95.22	16	4.49	1	0.28	–	–	–	–	–	–

Table 5.6. Item Exposure Rates—Winter Engine Evaluation

Grade	#Items				Exposure Rate											
					0–20%		21–40%		41–60%		61–80%		81–99%		100%	
	Total	Used	Unused	Unused %	N	%	N	%	N	%	N	%	N	%		
ELA																
3	539	438	101	18.74	407	92.92	21	4.79	4	0.91	2	0.46	2	0.46	2	0.46
4	555	532	23	4.14	509	95.68	12	2.26	11	2.07	–	–	–	–	–	–
5	482	386	96	19.92	356	92.23	18	4.66	4	1.04	5	1.30	3	0.78	–	–
6	473	453	20	4.23	425	93.82	27	5.96	1	0.22	–	–	–	–	–	–
7	459	364	95	20.70	319	87.64	34	9.34	5	1.37	3	0.82	1	0.27	2	0.55
8	508	435	73	14.37	409	94.02	16	3.68	6	1.38	2	0.46	–	–	2	0.46
Mathematics																
3	479	476	3	0.63	468	98.32	8	1.68	–	–	–	–	–	–	–	–
4	357	357	0	0.00	328	91.88	29	8.12	–	–	–	–	–	–	–	–
5	379	375	4	1.06	370	98.67	5	1.33	–	–	–	–	–	–	–	–
6	473	465	8	1.69	457	98.28	8	1.72	–	–	–	–	–	–	–	–
7	416	408	8	1.92	397	97.30	10	2.45	–	–	1	0.25	–	–	–	–
8	356	356	0	0.00	341	95.79	11	3.09	4	1.12	–	–	–	–	–	–

Table 5.7. Item Exposure Rates—Spring Simulations

Grade	#Items				Exposure Rate											
					0–20%		21–40%		41–60%		61–80%		81–99%		100%	
	Total	Used	Unused	Unused %	N	%	N	%	N	%	N	%	N	%		
ELA																
3	539	517	22	4.08	480	92.84	32	6.19	4	0.77	1	0.19	–	–	–	–
4	555	549	6	1.08	536	97.63	9	1.64	3	0.55	1	0.18	–	–	–	–
5	482	476	6	1.24	451	94.75	23	4.83	2	0.42	–	–	–	–	–	–
6	473	427	46	9.73	390	91.33	27	6.32	3	0.70	1	0.23	3	0.70	3	0.70
7	459	427	32	6.97	379	88.76	41	9.60	6	1.41	1	0.23	–	–	–	–
8	508	472	36	7.09	439	93.01	26	5.51	5	1.06	1	0.21	1	0.21	–	–
Mathematics																
3	714	708	6	0.84	696	98.31	12	1.69	–	–	–	–	–	–	–	–
4	520	508	12	2.31	493	97.05	15	2.95	–	–	–	–	–	–	–	–
5	564	556	8	1.42	537	96.58	18	3.24	1	0.18	–	–	–	–	–	–
6	708	691	17	2.40	677	97.97	13	1.88	1	0.14	–	–	–	–	–	–
7	617	605	12	1.94	585	96.69	16	2.64	4	0.66	–	–	–	–	–	–
8	528	522	6	1.14	507	97.13	13	2.49	2	0.38	–	–	–	–	–	–

Table 5.8. Item Exposure Rates—Spring Engine Evaluation

Grade	#Items				Exposure Rate									
					0–20%		21–40%		41–60%		61–80%		81–99%	
	Total	Used	Unused	Unused %	N	%	N	%	N	%	N	%	N	%
ELA														
3	539	519	20	3.71	482	92.87	32	6.17	5	0.96	–	–	–	–
4	555	549	6	1.08	538	98.00	7	1.28	3	0.55	1	0.18	–	–
5	482	471	11	2.28	448	95.12	20	4.25	2	0.42	1	0.21	–	–
6	473	436	37	7.82	401	91.97	26	5.96	2	0.46	1	0.23	3	0.69
7	459	432	27	5.88	388	89.81	39	9.03	3	0.69	2	0.46	–	–
8	508	475	33	6.50	439	92.42	29	6.11	5	1.05	1	0.21	1	0.21
Math														
3	714	705	9	1.26	691	98.01	14	1.99	–	–	–	–	–	–
4	520	511	9	1.73	496	97.06	15	2.94	–	–	–	–	–	–
5	564	556	8	1.42	534	96.04	22	3.96	–	–	–	–	–	–
6	708	700	8	1.13	686	98.00	13	1.86	1	0.14	–	–	–	–
7	617	608	9	1.46	592	97.37	13	2.14	3	0.49	–	–	–	–
8	528	522	6	1.14	507	97.13	13	2.49	2	0.38	–	–	–	–

5.2.4. Score Precision and Reliability

The studies provided precision ability estimations that showed how well the constraint-based engine recovered students’ true ability based on the item pool. It 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 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. Generally, it is expected that about 5% are outside the 95% confidence interval and about 1% are outside the 99% confidence interval.

Table 5.9 and Table 5.10 present the results of the precision ability estimation from the Winter and Spring simulations, respectively. Because this study did not involve an actual test administration, the constraint-based engine is not scoring student responses but is instead simulating whether a student got items correct or incorrect based on the student’s ability. Because a student’s true theta is known, the engine should be able to recover the student’s theta after administering all the items. This is the estimated theta. The null hypothesis is that there is no difference between true and estimated theta.

For the overall scores across all students, the mean biases are small (i.e., less than or equal to 0.03 in magnitude), for both ELA and mathematics, and the p -value for the z-test supports the null-hypothesis that there is not a significant difference between the simulated students' true and final estimated thetas. For some reporting category scores across all students, the mean biases are larger and the p -value for the z-test results are not supporting the null hypothesis. This is because the number of items is much smaller at the reporting category level and the large sample size could increase the likelihood of significant p -values. The MSE is also relatively small, showing that the constraint-based engine typically recovered a value near the student's true theta.

Table 5.9. Mean Bias of the NSCAS Ability Estimation (True– Estimated)—Winter Simulations

Grade	Reporting Category	Bias		P-value for Z-test	MSE	95% Coverage	99% Coverage
		Mean	SE				
ELA							
3	Reading Vocabulary	0.02	0.01	0.08	0.91	1.66	0.10
	Reading Comprehension	0.03	0.01	0.02	0.34	4.08	0.46
	Writing Skills	0.02	0.01	0.23	0.59	2.70	0.13
	Overall	0.01	0.00	0.32	0.16	4.45	0.73
4	Reading Vocabulary	-0.05	0.01	0.00	0.84	1.68	0.14
	Reading Comprehension	0.00	0.01	0.68	0.33	4.08	0.64
	Writing Skills	0.00	0.01	0.80	0.59	2.05	0.11
	Overall	0.00	0.00	0.86	0.15	5.10	0.88
5	Reading Vocabulary	-0.02	0.01	0.05	0.86	1.76	0.13
	Reading Comprehension	0.01	0.01	0.25	0.32	3.84	0.54
	Writing Skills	-0.01	0.01	0.63	0.57	2.16	0.10
	Overall	0.00	0.00	0.80	0.15	5.04	1.23
6	Reading Vocabulary	-0.03	0.01	0.01	0.92	1.44	0.08
	Reading Comprehension	0.00	0.01	0.94	0.30	4.36	0.56
	Writing Skills	-0.01	0.01	0.61	0.56	1.74	0.05
	Overall	0.00	0.00	0.88	0.15	5.23	1.04
7	Reading Vocabulary	-0.01	0.01	0.35	0.83	1.53	0.04
	Reading Comprehension	0.01	0.01	0.48	0.33	3.34	0.40
	Writing Skills	0.00	0.01	0.96	0.58	1.96	0.11
	Overall	0.00	0.00	0.81	0.14	4.18	0.68
8	Reading Vocabulary	-0.05	0.01	0.00	0.87	1.29	0.04
	Reading Comprehension	-0.01	0.01	0.23	0.29	3.78	0.30
	Writing Skills	0.04	0.01	0.00	0.56	1.78	0.06
	Overall	0.00	0.00	0.80	0.14	4.01	0.60
Mathematics							
3	Number	-0.02	0.01	0.31	0.46	3.70	0.41
	Algebra	-0.05	0.01	0.00	0.93	1.23	0.08
	Geometry	-0.04	0.01	0.02	0.63	2.24	0.13
	Data	-0.01	0.01	0.59	0.91	1.28	0.06
	Overall	0.00	0.00	0.82	0.15	5.03	1.04
4	Number	-0.02	0.01	0.12	0.42	2.05	0.05
	Algebra	0.00	0.01	0.86	0.72	1.11	0.04

Grade	Reporting Category	Bias		P-value for Z-test	MSE	95% Coverage	99% Coverage
		Mean	SE				
	Geometry	-0.01	0.01	0.43	0.71	0.71	0.03
	Data	0.07	0.01	0.00	0.98	0.75	0.04
	Overall	0.00	0.00	0.96	0.13	3.01	0.39
5	Number	0.01	0.01	0.56	0.39	1.89	0.06
	Algebra	0.06	0.01	0.00	0.75	1.30	0.06
	Geometry	-0.12	0.01	0.00	0.80	0.78	0.03
	Data	-0.08	0.01	0.00	0.93	0.69	0.01
	Overall	-0.02	0.00	0.21	0.13	2.61	0.19
6	Number	-0.01	0.01	0.58	0.59	1.10	0.04
	Algebra	-0.01	0.01	0.70	0.42	2.46	0.14
	Geometry	0.02	0.01	0.29	0.86	0.61	
	Data	0.06	0.01	0.00	0.91	1.04	0.01
	Overall	0.01	0.00	0.58	0.12	2.93	0.33
7	Number	0.02	0.01	0.12	0.73	0.89	
	Algebra	0.04	0.01	0.01	0.40	2.03	0.16
	Geometry	0.14	0.01	0.00	0.79	1.10	0.06
	Data	0.02	0.01	0.20	0.88	0.53	0.01
	Overall	0.04	0.00	0.01	0.12	2.45	0.34
8	Number	0.05	0.01	0.00	0.62	1.24	0.03
	Algebra	0.05	0.01	0.00	0.59	1.28	0.05
	Geometry	0.05	0.01	0.00	0.52	1.59	
	Data	-0.12	0.01	0.00	1.00	0.64	0.01
	Overall	0.02	0.00	0.14	0.12	2.13	0.21

Table 5.10. Mean Bias of the NSCAS Ability Estimation (True–Estimated)—Spring Simulations

Grade	Reporting Category	Bias		P-value for Z-test	MSE	95% Coverage	99% Coverage
		Mean	SE				
ELA							
3	Reading Vocabulary	-0.03	0.01	0.00	0.65	1.98	0.16
	Reading Comprehension	0.00	0.00	0.79	0.27	4.29	0.63
	Writing Skills	0.02	0.00	0.02	0.53	2.80	0.10
	Overall	0.00	0.00	0.84	0.13	4.94	0.98
4	Reading Vocabulary	-0.06	0.01	0.00	0.77	1.91	0.16
	Reading Comprehension	-0.02	0.00	0.02	0.29	4.61	0.70
	Writing Skills	-0.01	0.00	0.06	0.52	2.90	0.14
	Overall	-0.01	0.00	0.26	0.14	5.32	0.99
5	Reading Vocabulary	-0.07	0.01	0.00	0.80	2.07	0.18
	Reading Comprehension	-0.01	0.00	0.12	0.29	4.36	0.75
	Writing Skills	0.00	0.00	0.54	0.50	2.44	0.10
	Overall	-0.01	0.00	0.36	0.13	5.44	0.96
6	Reading Vocabulary	-0.05	0.01	0.00	0.79	1.98	0.10
	Reading Comprehension	-0.01	0.00	0.10	0.29	4.08	0.59
	Writing Skills	0.01	0.00	0.11	0.48	2.33	0.11
	Overall	0.00	0.00	0.75	0.13	5.03	0.95

Grade	Reporting Category	Bias		P-value for Z-test	MSE	95% Coverage	99% Coverage
		Mean	SE				
7	Reading Vocabulary	-0.06	0.01	0.00	0.69	1.69	0.13
	Reading Comprehension	0.00	0.00	0.55	0.25	3.92	0.57
	Writing Skills	0.03	0.00	0.00	0.52	2.17	0.08
	Overall	0.00	0.00	0.99	0.12	4.44	0.80
8	Reading Vocabulary	-0.06	0.01	0.00	0.74	1.54	0.10
	Reading Comprehension	0.01	0.00	0.20	0.26	3.77	0.51
	Writing Skills	0.04	0.00	0.00	0.50	2.18	0.08
	Overall	0.00	0.00	0.46	0.12	4.05	0.73
Mathematics							
3	Number	-0.01	0.00	0.13	0.34	3.93	0.45
	Algebra	-0.05	0.01	0.00	0.84	1.27	0.09
	Geometry	-0.03	0.00	0.00	0.52	2.90	0.23
	Data	-0.01	0.01	0.10	0.80	1.37	0.04
	Overall	-0.01	0.00	0.28	0.12	5.28	1.10
4	Number	-0.02	0.00	0.05	0.32	2.59	0.18
	Algebra	-0.01	0.00	0.28	0.54	1.68	0.05
	Geometry	0.00	0.00	0.89	0.54	1.53	0.05
	Data	0.04	0.01	0.00	0.81	0.87	0.01
	Overall	0.00	0.00	0.96	0.10	2.92	0.36
5	Number	0.01	0.00	0.30	0.30	2.33	0.17
	Algebra	-0.01	0.00	0.09	0.57	1.70	0.03
	Geometry	-0.04	0.00	0.00	0.47	1.37	0.02
	Data	-0.10	0.01	0.00	0.83	0.61	0.01
	Overall	-0.02	0.00	0.04	0.09	2.39	0.36
6	Number	-0.01	0.00	0.43	0.44	1.78	0.04
	Algebra	0.01	0.00	0.44	0.38	2.51	0.22
	Geometry	0.00	0.01	0.63	0.61	1.21	0.02
	Data	0.08	0.01	0.00	0.67	1.22	0.02
	Overall	0.01	0.00	0.23	0.09	2.73	0.34
7	Number	0.03	0.00	0.00	0.57	1.47	0.03
	Algebra	0.03	0.00	0.00	0.35	2.35	0.10
	Geometry	0.10	0.01	0.00	0.67	0.80	0.01
	Data	0.01	0.00	0.10	0.55	1.60	0.03
	Overall	0.03	0.00	0.00	0.09	2.55	0.28
8	Number	0.06	0.00	0.00	0.49	1.58	0.08
	Algebra	0.06	0.00	0.00	0.48	1.46	0.03
	Geometry	0.03	0.00	0.00	0.47	1.80	0.05
	Data	0.01	0.01	0.09	0.63	1.10	0.02
	Overall	0.03	0.00	0.00	0.09	2.28	0.27

Table 5.11 through Table 5.14 present the score precision and reliability estimates for the Winter pre-administration simulation study, the Winter post-administration engine evaluation study, the Spring pre-administration simulation study, and the Spring post-administration engine evaluation study, respectively. Tables include the average number of items administered, the

standard deviation (SD) of the estimated theta, the mean SEM, the RMSE, and a marginal reliability coefficient. The SD, mean SEM, and RMSE are relatively small, and the range of the marginal reliability for the overall scores is close to or higher than 0.90. These results indicate that, overall, the score precision is reasonable: The overall mean SEM values were approximately 0.40 while the reliability estimates are consistent with the guidelines for reliability in a graduation test (Phillips & Camara, 2006). The reliability for the overall scores shows higher reliability estimates compared to that of reporting category scores, which can be expected as more items are contributing to the overall scores.

Table 5.11. Score Precision and Reliability, Items Contributed to NSCAS—Winter Simulations

Grade		Average #Items	SD of Estimated Theta	Mean SEM	RMSE	Reliability
ELA						
3	Reading Vocabulary	6	1.52	0.99	1.01	0.55
	Reading Comprehension	14	1.31	0.58	0.58	0.80
	Writing Skills	7	1.39	0.75	0.76	0.70
	Overall	27	1.22	0.40	0.40	0.89
4	Reading Vocabulary	6	1.45	0.98	1.03	0.50
	Reading Comprehension	14	1.26	0.56	0.57	0.80
	Writing Skills	7	1.36	0.74	0.75	0.69
	Overall	27	1.17	0.38	0.39	0.89
5	Reading Vocabulary	6	1.43	0.94	0.98	0.53
	Reading Comprehension	14	1.24	0.55	0.56	0.79
	Writing Skills	7	1.33	0.73	0.75	0.68
	Overall	27	1.14	0.38	0.38	0.89
6	Reading Vocabulary	6	1.43	0.94	0.96	0.55
	Reading Comprehension	14	1.18	0.53	0.54	0.79
	Writing Skills	7	1.28	0.71	0.73	0.67
	Overall	27	1.09	0.37	0.37	0.89
7	Reading Vocabulary	6	1.38	0.89	0.92	0.56
	Reading Comprehension	14	1.16	0.56	0.57	0.76
	Writing Skills	7	1.27	0.74	0.75	0.65
	Overall	27	1.06	0.38	0.38	0.87
8	Reading Vocabulary	6	1.36	0.95	0.98	0.48
	Reading Comprehension	14	1.15	0.53	0.54	0.78
	Writing Skills	7	1.26	0.73	0.74	0.65
	Overall	27	1.05	0.37	0.37	0.87
Mathematics						
3	Number	10	1.53	0.67	0.68	0.80
	Algebra	5	1.70	0.98	1.01	0.65
	Geometry	7	1.59	0.80	0.82	0.73
	Data	5	1.67	0.98	1.01	0.63
	Overall	27	1.42	0.39	0.39	0.93
4	Number	10	1.51	0.67	0.67	0.80
	Algebra	6	1.61	0.87	0.89	0.70
	Geometry	6	1.60	0.87	0.88	0.70
	Data	5	1.76	1.06	1.11	0.61

Grade		Average #Items	SD of Estimated Theta	Mean SEM	RMSE	Reliability
	Overall	27	1.41	0.39	0.39	0.92
5	Number	10	1.47	0.66	0.67	0.80
	Algebra	6	1.59	0.90	0.91	0.67
	Geometry	6	1.60	0.91	0.94	0.65
	Data	5	1.67	1.01	1.04	0.61
	Overall	27	1.38	0.39	0.39	0.92
6	Number	7	1.52	0.78	0.80	0.73
	Algebra	10	1.47	0.67	0.67	0.79
	Geometry	5	1.58	0.94	0.96	0.63
	Data	5	1.66	0.98	1.01	0.63
	Overall	27	1.35	0.38	0.38	0.92
7	Number	6	1.51	0.87	0.88	0.66
	Algebra	10	1.38	0.66	0.66	0.77
	Geometry	6	1.55	0.93	0.96	0.62
	Data	5	1.53	0.95	0.97	0.60
	Overall	27	1.27	0.38	0.38	0.91
8	Number	7	1.51	0.81	0.82	0.70
	Algebra	7	1.54	0.80	0.81	0.72
	Geometry	8	1.49	0.76	0.77	0.74
	Data	5	1.70	1.04	1.09	0.59
	Overall	27	1.34	0.39	0.39	0.92

Table 5.12. Score Precision and Reliability, Items Contributed to NSCAS—Winter Engine Evaluation

Grade		Average #Items	SD of Estimated Theta	Mean SEM	RMSE	Reliability
ELA						
3	Reading Vocabulary	6	1.46	0.99	1.02	0.51
	Reading Comprehension	14	1.11	0.57	0.57	0.74
	Writing Skills	7	1.26	0.77	0.78	0.61
	Overall	27	1.04	0.39	0.40	0.85
4	Reading Vocabulary	6	1.50	0.96	1.00	0.55
	Reading Comprehension	14	1.23	0.56	0.56	0.79
	Writing Skills	7	1.25	0.74	0.75	0.64
	Overall	27	1.11	0.38	0.38	0.88
5	Reading Vocabulary	6	1.45	0.94	0.97	0.55
	Reading Comprehension	14	1.12	0.54	0.55	0.76
	Writing Skills	7	1.20	0.74	0.75	0.60
	Overall	27	1.01	0.37	0.38	0.86
6	Reading Vocabulary	6	1.41	0.94	0.96	0.54
	Reading Comprehension	14	1.17	0.53	0.54	0.79
	Writing Skills	7	1.17	0.73	0.75	0.59
	Overall	27	1.03	0.37	0.37	0.87
7	Reading Vocabulary	6	1.40	0.90	0.92	0.57
	Reading Comprehension	14	1.10	0.55	0.56	0.74

Grade		Average #Items	SD of Estimated Theta	Mean SEM	RMSE	Reliability
	Writing Skills	7	1.18	0.73	0.75	0.60
	Overall	27	0.99	0.37	0.38	0.85
8	Reading Vocabulary	6	1.40	0.94	0.96	0.53
	Reading Comprehension	14	1.12	0.53	0.54	0.77
	Writing Skills	7	1.13	0.73	0.75	0.56
	Overall	27	0.99	0.37	0.37	0.86
Mathematics						
3	Number	10	1.37	0.68	0.68	0.75
	Algebra	5	1.59	0.99	1.01	0.59
	Geometry	7	1.29	0.78	0.79	0.63
	Data	5	1.61	1.01	1.04	0.58
	Overall	27	1.21	0.38	0.38	0.90
4	Number	10	1.38	0.69	0.70	0.74
	Algebra	6	1.57	0.91	0.93	0.65
	Geometry	6	1.53	0.93	0.96	0.61
	Data	5	1.43	1.05	1.10	0.41
	Overall	27	1.20	0.40	0.40	0.89
5	Number	10	1.31	0.67	0.68	0.73
	Algebra	6	1.34	0.93	0.96	0.49
	Geometry	6	1.45	0.89	0.91	0.61
	Data	5	1.59	1.01	1.05	0.56
	Overall	27	1.12	0.38	0.39	0.88
6	Number	7	1.45	0.81	0.83	0.67
	Algebra	10	1.40	0.69	0.69	0.75
	Geometry	5	1.47	0.98	1.01	0.52
	Data	5	1.32	1.06	1.10	0.30
	Overall	27	1.14	0.38	0.38	0.89
7	Number	6	1.43	0.89	0.91	0.59
	Algebra	10	1.35	0.67	0.68	0.74
	Geometry	6	1.26	0.98	1.01	0.35
	Data	5	1.42	0.99	1.01	0.49
	Overall	27	1.08	0.39	0.39	0.87
8	Number	7	1.44	0.87	0.89	0.62
	Algebra	7	1.46	0.82	0.84	0.67
	Geometry	8	1.43	0.82	0.84	0.65
	Data	5	1.51	1.02	1.06	0.50
	Overall	27	1.15	0.39	0.40	0.88

Table 5.13. Score Precision and Reliability, Items Contributed to NSCAS—Spring Simulations

Grade		Average #Items	SD of Estimated Theta	Mean SEM	RMSE	Reliability
ELA						
3	Reading Vocabulary	7	1.42	0.82	0.84	0.64
	Reading Comprehension	15	1.27	0.51	0.52	0.83
	Writing Skills	8	1.37	0.71	0.72	0.72

Grade		Average #Items	SD of Estimated Theta	Mean SEM	RMSE	Reliability
	Overall	30	1.20	0.36	0.36	0.91
4	Reading Vocabulary	7	1.42	0.92	0.97	0.54
	Reading Comprehension	15	1.23	0.52	0.53	0.81
	Writing Skills	8	1.32	0.70	0.70	0.71
	Overall	30	1.15	0.36	0.37	0.90
5	Reading Vocabulary	7	1.41	0.91	0.95	0.55
	Reading Comprehension	15	1.19	0.52	0.52	0.81
	Writing Skills	8	1.29	0.68	0.69	0.72
	Overall	30	1.12	0.36	0.36	0.90
6	Reading Vocabulary	7	1.39	0.87	0.90	0.58
	Reading Comprehension	15	1.17	0.52	0.53	0.79
	Writing Skills	8	1.25	0.66	0.68	0.71
	Overall	30	1.08	0.35	0.35	0.89
7	Reading Vocabulary	7	1.32	0.82	0.85	0.58
	Reading Comprehension	15	1.12	0.50	0.50	0.80
	Writing Skills	8	1.26	0.70	0.72	0.68
	Overall	30	1.05	0.35	0.35	0.89
8	Reading Vocabulary	7	1.33	0.88	0.91	0.53
	Reading Comprehension	15	1.11	0.51	0.51	0.79
	Writing Skills	8	1.22	0.70	0.71	0.67
	Overall	30	1.04	0.35	0.36	0.88
Mathematics						
3	Number	12	1.48	0.58	0.58	0.85
	Algebra	5	1.67	0.90	0.93	0.69
	Geometry	8	1.56	0.71	0.72	0.79
	Data	5	1.66	0.89	0.92	0.69
	Overall	30	1.41	0.34	0.34	0.94
4	Number	11	1.46	0.59	0.59	0.84
	Algebra	7	1.54	0.73	0.75	0.76
	Geometry	7	1.53	0.73	0.75	0.76
	Data	5	1.62	0.92	0.94	0.67
	Overall	30	1.37	0.34	0.34	0.94
5	Number	11	1.45	0.57	0.57	0.84
	Algebra	7	1.51	0.76	0.77	0.74
	Geometry	7	1.48	0.68	0.70	0.78
	Data	5	1.62	0.93	0.96	0.65
	Overall	30	1.34	0.33	0.33	0.94
6	Number	8	1.46	0.67	0.67	0.79
	Algebra	10	1.45	0.63	0.64	0.81
	Geometry	6	1.51	0.79	0.80	0.72
	Data	6	1.59	0.83	0.85	0.71
	Overall	30	1.34	0.34	0.34	0.94
7	Number	7	1.45	0.76	0.77	0.72
	Algebra	10	1.36	0.62	0.62	0.79

Grade		Average #Items	SD of Estimated Theta	Mean SEM	RMSE	Reliability
	Geometry	6	1.54	0.82	0.85	0.70
	Data	7	1.41	0.75	0.76	0.71
	Overall	30	1.26	0.34	0.34	0.93
8	Number	8	1.47	0.73	0.74	0.75
	Algebra	8	1.51	0.69	0.70	0.78
	Geometry	8	1.48	0.71	0.72	0.77
	Data	6	1.52	0.80	0.82	0.71
	Overall	30	1.33	0.34	0.34	0.93

Table 5.14. Score Precision and Reliability, Items Contributed to NSCAS—Spring Engine Evaluation

Grade		Average #Items	SD of Estimated Theta	Mean SEM	RMSE	Reliability
ELA						
3	Reading Vocabulary	7	1.54	0.83	0.85	0.69
	Reading Comprehension	15	1.36	0.52	0.52	0.85
	Writing Skills	8	1.23	0.71	0.72	0.66
	Overall	30	1.22	0.36	0.36	0.91
4	Reading Vocabulary	7	1.57	0.94	0.99	0.60
	Reading Comprehension	15	1.33	0.52	0.53	0.84
	Writing Skills	8	1.23	0.69	0.70	0.68
	Overall	30	1.19	0.36	0.36	0.91
5	Reading Vocabulary	7	1.47	0.91	0.94	0.59
	Reading Comprehension	15	1.25	0.52	0.52	0.83
	Writing Skills	8	1.21	0.68	0.69	0.67
	Overall	30	1.13	0.36	0.36	0.90
6	Reading Vocabulary	7	1.39	0.88	0.91	0.57
	Reading Comprehension	15	1.17	0.52	0.53	0.80
	Writing Skills	8	1.21	0.67	0.69	0.68
	Overall	30	1.04	0.35	0.35	0.89
7	Reading Vocabulary	7	1.33	0.81	0.83	0.61
	Reading Comprehension	15	1.25	0.51	0.51	0.83
	Writing Skills	8	1.12	0.69	0.70	0.61
	Overall	30	1.07	0.35	0.35	0.89
8	Reading Vocabulary	7	1.35	0.87	0.90	0.56
	Reading Comprehension	15	1.13	0.51	0.51	0.79
	Writing Skills	8	1.12	0.69	0.70	0.60
	Overall	30	1.01	0.35	0.36	0.88
Mathematics						
3	Number	12	1.66	0.59	0.59	0.87
	Algebra	5	1.77	0.91	0.94	0.72
	Geometry	8	1.56	0.71	0.72	0.78
	Data	5	1.71	0.90	0.93	0.70
	Overall	30	1.48	0.34	0.35	0.95
4	Number	11	1.63	0.60	0.61	0.86

Grade		Average #Items	SD of Estimated Theta	Mean SEM	RMSE	Reliability
	Algebra	7	1.69	0.75	0.77	0.79
	Geometry	7	1.60	0.76	0.78	0.77
	Data	5	1.64	0.93	0.96	0.66
	Overall	30	1.42	0.34	0.34	0.94
5	Number	11	1.52	0.58	0.59	0.85
	Algebra	7	1.66	0.78	0.80	0.77
	Geometry	7	1.53	0.71	0.73	0.77
	Data	5	1.73	0.95	0.99	0.67
	Overall	30	1.37	0.33	0.33	0.94
6	Number	8	1.47	0.68	0.69	0.78
	Algebra	10	1.57	0.64	0.65	0.83
	Geometry	6	1.62	0.81	0.83	0.73
	Data	6	1.58	0.85	0.88	0.69
	Overall	30	1.34	0.34	0.34	0.94
7	Number	7	1.54	0.77	0.79	0.74
	Algebra	10	1.50	0.63	0.64	0.82
	Geometry	6	1.53	0.83	0.86	0.69
	Data	7	1.58	0.77	0.79	0.75
	Overall	30	1.33	0.34	0.34	0.93
8	Number	8	1.57	0.75	0.76	0.76
	Algebra	8	1.64	0.71	0.72	0.80
	Geometry	8	1.60	0.73	0.74	0.79
	Data	6	1.62	0.82	0.84	0.73
	Overall	30	1.39	0.35	0.35	0.94

Table 5.15 through Table 5.19 present the average SEM by decile of the overall proficiency score, including the overall student ability distribution, for the Winter pre-administration simulation study, the Winter post-administration engine evaluation study, the Spring pre-administration simulation study, and the Spring post-administration engine evaluation study, respectively. A decile is similar to a percentile rank, with 10 ranks related to the 10th, 20th . . . 90th, 100th percentile ranks. 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 (i.e., less than 0.40).

Table 5.15. SEM by Deciles for NSCAS Scores—Winter 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	
ELA											
3	0.48	0.42	0.40	0.38	0.38	0.37	0.37	0.37	0.38	0.40	0.40
4	0.42	0.38	0.37	0.36	0.35	0.36	0.36	0.37	0.39	0.47	0.38
5	0.44	0.40	0.38	0.37	0.36	0.35	0.35	0.35	0.37	0.43	0.38
6	0.42	0.37	0.35	0.35	0.34	0.34	0.35	0.35	0.37	0.42	0.37
7	0.44	0.39	0.37	0.36	0.36	0.35	0.35	0.36	0.37	0.43	0.38
8	0.42	0.38	0.36	0.35	0.35	0.35	0.35	0.35	0.37	0.43	0.37
Mathematics											

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	
3	0.41	0.38	0.38	0.37	0.37	0.37	0.37	0.37	0.38	0.46	0.39
4	0.45	0.40	0.38	0.37	0.37	0.37	0.37	0.38	0.38	0.41	0.39
5	0.42	0.39	0.38	0.37	0.37	0.37	0.37	0.37	0.39	0.46	0.39
6	0.43	0.38	0.38	0.37	0.37	0.37	0.37	0.37	0.37	0.38	0.38
7	0.45	0.39	0.38	0.38	0.37	0.37	0.37	0.36	0.37	0.38	0.38
8	0.45	0.40	0.39	0.38	0.38	0.37	0.37	0.37	0.37	0.40	0.39

Table 5.16. SEM by Deciles for NSCAS Scores—Winter 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	
ELA											
3	0.49	0.42	0.40	0.39	0.38	0.37	0.37	0.37	0.37	0.38	0.39
4	0.45	0.40	0.38	0.36	0.35	0.35	0.35	0.36	0.37	0.42	0.38
5	0.44	0.40	0.38	0.37	0.36	0.35	0.34	0.34	0.35	0.40	0.37
6	0.46	0.39	0.37	0.35	0.34	0.34	0.34	0.35	0.36	0.38	0.37
7	0.46	0.40	0.38	0.37	0.36	0.35	0.35	0.35	0.36	0.39	0.38
8	0.45	0.39	0.37	0.36	0.35	0.35	0.35	0.35	0.35	0.38	0.37
Mathematics											
3	0.42	0.40	0.39	0.38	0.37	0.37	0.37	0.37	0.37	0.39	0.38
4	0.47	0.42	0.41	0.40	0.39	0.37	0.37	0.37	0.37	0.38	0.40
5	0.44	0.40	0.39	0.38	0.38	0.37	0.37	0.37	0.36	0.39	0.38
6	0.45	0.40	0.39	0.38	0.38	0.37	0.37	0.36	0.36	0.37	0.38
7	0.45	0.40	0.39	0.39	0.38	0.38	0.37	0.37	0.36	0.37	0.39
8	0.46	0.42	0.40	0.39	0.39	0.38	0.38	0.38	0.37	0.39	0.39

Table 5.17. SEM by Deciles for NSCAS Scores—Spring 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	
ELA											
3	0.40	0.36	0.35	0.34	0.34	0.34	0.34	0.34	0.35	0.39	0.36
4	0.37	0.35	0.35	0.34	0.34	0.34	0.35	0.36	0.38	0.45	0.36
5	0.38	0.35	0.35	0.35	0.34	0.34	0.34	0.34	0.36	0.42	0.36
6	0.39	0.35	0.34	0.33	0.33	0.33	0.34	0.34	0.36	0.41	0.35
7	0.39	0.35	0.34	0.34	0.33	0.33	0.33	0.33	0.34	0.39	0.35
8	0.39	0.36	0.35	0.35	0.34	0.34	0.34	0.34	0.35	0.39	0.35
Mathematics											
3	0.36	0.34	0.33	0.33	0.33	0.33	0.33	0.34	0.35	0.39	0.34
4	0.38	0.34	0.32	0.32	0.32	0.33	0.33	0.33	0.34	0.37	0.34
5	0.35	0.32	0.32	0.32	0.32	0.31	0.31	0.32	0.33	0.38	0.33
6	0.38	0.34	0.33	0.33	0.33	0.33	0.33	0.32	0.32	0.34	0.34
7	0.40	0.36	0.34	0.33	0.33	0.32	0.32	0.32	0.33	0.34	0.34
8	0.40	0.36	0.35	0.34	0.33	0.33	0.32	0.32	0.33	0.35	0.34

Table 5.18. SEM by Deciles for NSCAS Scores—Spring 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	
ELA											
3	0.43	0.37	0.34	0.34	0.34	0.34	0.34	0.34	0.35	0.37	0.36
4	0.39	0.36	0.35	0.34	0.34	0.34	0.35	0.36	0.37	0.42	0.36
5	0.39	0.36	0.35	0.35	0.34	0.34	0.34	0.34	0.36	0.40	0.36
6	0.41	0.35	0.34	0.33	0.33	0.33	0.33	0.34	0.35	0.38	0.35
7	0.42	0.36	0.35	0.34	0.33	0.33	0.32	0.32	0.33	0.36	0.35
8	0.42	0.37	0.35	0.34	0.34	0.34	0.34	0.34	0.34	0.36	0.35
Mathematics											
3	0.36	0.34	0.33	0.32	0.32	0.33	0.33	0.34	0.35	0.40	0.34
4	0.37	0.34	0.32	0.32	0.32	0.33	0.33	0.33	0.34	0.37	0.34
5	0.35	0.32	0.32	0.32	0.32	0.31	0.31	0.32	0.33	0.38	0.33
6	0.38	0.34	0.33	0.33	0.33	0.33	0.33	0.33	0.32	0.34	0.34
7	0.40	0.36	0.35	0.33	0.33	0.32	0.32	0.32	0.33	0.35	0.34
8	0.41	0.37	0.35	0.34	0.33	0.33	0.32	0.32	0.33	0.35	0.35

5.3. Engine Simulations: Science Field Test

Spring 2022 science are operational field tests (i.e., all items were re-calibrated following the 2022 administration). The number of items and points possible are reported in Table 2.2. Spring 2022 Science tests can be summarized as following:

- English online forms
 - Each grade has 20 different forms, but the operational items are the same across forms.
 - Grade 5 has 5 sets, including 4 operational sets and 1 field test set. Each field test set includes 4 to 6 items.
 - Grade 8 has 5 sets, including 5 operational sets and 1 field test set. Each field test set includes 4 to 7 items.
 - The overall test score is based on 21 operational items (worth 22 points) in grade 5 and 27 operational items (worth 33 points) in grade 8.
- Paper-pencil and Spanish forms
 - Each grade has 4 sets, including 3 operational sets and 1 field test set. Field test set includes 4 and 7 items for grades 5 and 8, respectively.
 - The overall test score is based on 17 operational items (worth 18 points) in grade 5 and 16 operational items (worth 20 points) in grade 8.

A simulation study and an engine evaluation check were conducted to provide evidence that the NWEA constraint-based engine can properly administer the fixed forms as intended for the newly developed NSCAS science assessment for Grades 5 and 8. Because science assessments are fixed forms with small number of operational items, simulation and engine evaluation focused on whether each form was delivered to a representative sample of Nebraska students.

Approximately 23,000 students per grade were included in the simulation study sample. The true values of student ability (θ , or θ) were drawn from a normal distribution with a mean of

0.0 and a standard deviation of 1.0. The student sample was simulated to have similar demographic characteristics to Nebraska’s general student population based on the roster file. The student sample also had similar demographic characteristics to Nebraska’s general population based on the 2022 roster file, as shown in Table 5.19.

Tables 5.20 through 5.23 present the number and percentage of simulated students who received each form by gender and ethnicity for grades 5 and 8, respectively. Each form was delivered to a representative sample of Nebraska students, demonstrating that the proportions set in the engine population exposure control are representative of the Nebraska general student population in terms of gender and ethnicity. Thus, it can be reasonably assumed that each field test task and its prompts were also delivered to a representative sample of Nebraska students. These results suggest that the population exposure control function of the constraint-based engine works well.

Table 5.19. General Population Demographic Distribution

Grade	Gender				Ethnicity										Total N
	Female		Male		American Indian		Asian		Black		Hispanic		White		
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	
Nebraska General Population															
5	11,066	48.7	11,674	51.3	270	1.2	685	3.0	1,415	6.2	4,566	20.1	14,702	64.7	22,740
8	11,686	48.5	12,407	51.5	303	1.3	682	2.8	1,603	6.7	4,881	20.3	15,510	64.4	24,093
Simulation Student Sample															
5	11,616	48.6	12,284	51.4	290	1.2	685	2.9	1,687	7.1	4,576	19.1	15,651	65.5	23,900
8	11,178	48.5	11,848	51.5	278	1.2	615	2.7	1,588	6.9	4,269	18.5	15,405	66.9	23,026

Table 5.20. Demographic Distribution by Form—Grade 5 (Simulation)

Form	Gender				Ethnicity										Total N
	Female		Male		American Indian		Asian		Black		Hispanic		White		
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	
(All)	11,616	48.6	12,284	51.4	290	1.2	685	2.9	1,687	7.1	4,576	19.1	15,651	65.5	23,900
A	593	49.2	612	50.8	14	1.2	36	3.0	82	6.8	253	21.0	770	63.9	1,205
B	591	49.1	613	50.9	19	1.6	35	2.9	85	7.1	231	19.2	789	65.5	1,204
C	565	47.0	637	53.0	20	1.7	37	3.1	87	7.2	218	18.1	788	65.6	1,202
D	554	46.9	627	53.1	14	1.2	35	3.0	85	7.2	201	17.0	791	67.0	1,181
E	572	47.7	627	52.3	16	1.3	32	2.7	88	7.3	247	20.6	772	64.4	1,199
F	591	49.2	611	50.8	13	1.1	38	3.2	86	7.2	232	19.3	783	65.1	1,202
G	595	49.4	609	50.6	16	1.3	31	2.6	83	6.9	239	19.9	783	65.0	1,204
H	565	47.4	627	52.6	13	1.1	32	2.7	89	7.5	225	18.9	794	66.6	1,192
I	583	49.3	600	50.7	9	0.8	29	2.5	80	6.8	229	19.4	792	66.9	1,183
J	541	46.0	636	54.0	11	0.9	34	2.9	81	6.9	224	19.0	789	67.0	1,177
K	593	49.9	595	50.1	15	1.3	31	2.6	81	6.8	228	19.2	778	65.5	1,188
L	597	49.7	605	50.3	16	1.3	36	3.0	89	7.4	235	19.6	780	64.9	1,202
M	577	48.4	614	51.6	11	0.9	34	2.9	82	6.9	220	18.5	799	67.1	1,191
N	566	47.7	621	52.3	11	0.9	39	3.3	83	7.0	220	18.5	788	66.4	1,187
O	606	51.0	582	49.0	9	0.8	37	3.1	84	7.1	233	19.6	774	65.2	1,188

Form	Gender				Ethnicity								Total N		
	Female		Male		American Indian		Asian		Black		Hispanic			White	
	N	%	N	%	N	%	N	%	N	%	N	%		N	%
P	573	47.9	623	52.1	23	1.9	34	2.8	90	7.5	232	19.4	769	64.3	1,196
Q	564	47.7	618	52.3	15	1.3	33	2.8	80	6.8	227	19.2	757	64.0	1,182
R	603	50.1	600	49.9	16	1.3	35	2.9	84	7.0	244	20.3	776	64.5	1,203
S	587	49.7	594	50.3	18	1.5	32	2.7	82	6.9	212	18.0	801	67.8	1,181
T	600	48.7	633	51.3	11	0.9	35	2.8	86	7.0	226	18.3	778	63.1	1,233

Table 5.21. Demographic Distribution by Form—Grade 8 (Simulation)

Form	Gender				Ethnicity								Total N		
	Female		Male		American Indian		Asian		Black		Hispanic			White	
	N	%	N	%	N	%	N	%	N	%	N	%		N	%
(All)	11,178	48.5	11,848	51.5	278	1.2	615	2.7	1,588	6.9	4,269	18.5	15,405	66.9	23,026
A	543	47.4	603	52.6	14	1.2	28	2.4	79	6.9	219	19.1	771	67.3	1,146
B	584	49.7	590	50.3	12	1.0	32	2.7	84	7.2	203	17.3	796	67.8	1,174
C	576	48.8	605	51.2	12	1.0	30	2.5	80	6.8	223	18.9	796	67.4	1,181
D	550	48.6	582	51.4	14	1.2	32	2.8	80	7.1	213	18.8	747	66.0	1,132
E	554	48.2	596	51.8	15	1.3	31	2.7	78	6.8	219	19.0	767	66.7	1,150
F	590	50.0	590	50.0	10	0.8	38	3.2	81	6.9	213	18.1	784	66.4	1,180
G	562	49.0	585	51.0	15	1.3	32	2.8	77	6.7	204	17.8	776	67.7	1,147
H	554	48.9	580	51.1	7	0.6	34	3.0	71	6.3	213	18.8	757	66.8	1,134
I	560	48.6	593	51.4	10	0.9	31	2.7	84	7.3	218	18.9	770	66.8	1,153
J	552	47.8	603	52.2	13	1.1	32	2.8	88	7.6	206	17.8	771	66.8	1,155
K	540	47.0	609	53.0	24	2.1	30	2.6	75	6.5	224	19.5	767	66.8	1,149
L	526	45.7	624	54.3	18	1.6	30	2.6	86	7.5	205	17.8	770	67.0	1,150
M	568	50.2	563	49.8	12	1.1	27	2.4	72	6.4	205	18.1	779	68.9	1,131
N	530	47.9	577	52.1	15	1.4	30	2.7	83	7.5	192	17.3	743	67.1	1,107
O	566	49.0	590	51.0	12	1.0	33	2.9	89	7.7	217	18.8	757	65.5	1,156
P	540	47.5	598	52.5	11	1.0	31	2.7	73	6.4	220	19.3	766	67.3	1,138
Q	564	48.7	595	51.3	12	1.0	25	2.2	80	6.9	212	18.3	792	68.3	1,159
R	554	48.6	585	51.4	17	1.5	28	2.5	69	6.1	238	20.9	752	66.0	1,139
S	566	49.6	576	50.4	18	1.6	29	2.5	71	6.2	214	18.7	762	66.7	1,142
T	599	49.8	604	50.2	17	1.4	32	2.7	88	7.3	211	17.5	782	65.0	1,203

Table 5.22. Demographic Distribution by Form—Grade 5 (Engine Evaluation)

Form	Gender				Ethnicity								Total N		
	Female		Male		American Indian		Asian		Black		Hispanic			White	
	N	%	N	%	N	%	N	%	N	%	N	%		N	%
(All)	10,775	48.6	11,409	51.4	259	1.2	687	3.1	1,383	6.2	4,430	20.0	14,343	64.7	22,184
A	535	48.6	566	51.4	15	1.4	34	3.1	64	5.8	205	18.6	730	66.3	1,101
B	527	46.8	598	53.2	17	1.5	40	3.6	73	6.5	213	18.9	725	64.4	1,125
C	568	51.3	540	48.7	10	0.9	34	3.1	70	6.3	230	20.8	699	63.1	1,108

Form	Gender				Ethnicity										Total N
	Female		Male		American Indian		Asian		Black		Hispanic		White		
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	
D	503	46.0	591	54.0	9	0.8	32	2.9	63	5.8	207	18.9	735	67.2	1,094
E	553	50.4	544	49.6	12	1.1	35	3.2	67	6.1	231	21.1	714	65.1	1,097
F	518	47.1	581	52.9	7	0.6	33	3.0	67	6.1	217	19.7	724	65.9	1,099
G	526	47.4	584	52.6	11	1.0	35	3.2	70	6.3	235	21.2	708	63.8	1,110
H	542	48.0	586	52.0	23	2.0	32	2.8	71	6.3	242	21.5	717	63.6	1,128
I	516	47.1	580	52.9	14	1.3	31	2.8	74	6.8	220	20.1	716	65.3	1,096
J	513	46.4	593	53.6	9	0.8	32	2.9	67	6.1	234	21.2	709	64.1	1,106
K	573	52.6	516	47.4	8	0.7	33	3.0	67	6.2	225	20.7	705	64.7	1,089
L	543	48.4	579	51.6	12	1.1	39	3.5	66	5.9	209	18.6	745	66.4	1,122
M	518	47.3	576	52.7	12	1.1	34	3.1	69	6.3	212	19.4	714	65.3	1,094
N	530	48.1	571	51.9	15	1.4	35	3.2	73	6.6	216	19.6	713	64.8	1,101
O	528	47.9	575	52.1	15	1.4	35	3.2	70	6.3	213	19.3	711	64.5	1,103
P	557	49.2	576	50.8	16	1.4	37	3.3	74	6.5	221	19.5	728	64.3	1,133
Q	567	51.5	535	48.5	16	1.5	35	3.2	74	6.7	216	19.6	711	64.5	1,102
R	554	50.0	554	50.0	13	1.2	33	3.0	73	6.6	229	20.7	717	64.7	1,108
S	560	49.6	568	50.4	16	1.4	36	3.2	65	5.8	236	20.9	716	63.5	1,128
T	544	47.7	596	52.3	9	0.8	32	2.8	66	5.8	219	19.2	706	61.9	1,140

Table 5.23. Demographic Distribution by Form—Grade 8 (Engine Evaluation)

Form	Gender				Ethnicity										Total N
	Female		Male		American Indian		Asian		Black		Hispanic		White		
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	
(All)	11,409	48.4	12,142	51.6	286	1.2	674	2.9	1,544	6.6	4,791	20.3	15,197	64.5	23,551
A	561	48.0	608	52.0	15	1.3	34	2.9	73	6.2	227	19.4	760	65.0	1,169
B	598	49.8	603	50.2	17	1.4	35	2.9	87	7.2	252	21.0	755	62.9	1,201
C	566	48.3	605	51.7	13	1.1	34	2.9	72	6.1	261	22.3	749	64.0	1,171
D	592	50.0	592	50.0	18	1.5	31	2.6	79	6.7	256	21.6	754	63.7	1,184
E	563	47.4	626	52.6	12	1.0	35	2.9	85	7.1	246	20.7	763	64.2	1,189
F	561	47.4	623	52.6	17	1.4	38	3.2	77	6.5	240	20.3	756	63.9	1,184
G	540	45.4	649	54.6	13	1.1	39	3.3	78	6.6	234	19.7	782	65.8	1,189
H	610	51.0	585	49.0	11	0.9	32	2.7	71	5.9	244	20.4	788	65.9	1,195
I	582	48.4	620	51.6	10	0.8	32	2.7	88	7.3	245	20.4	773	64.3	1,202
J	593	50.0	593	50.0	9	0.8	35	3.0	90	7.6	253	21.3	736	62.1	1,186
K	558	48.4	596	51.6	14	1.2	34	2.9	69	6.0	249	21.6	744	64.5	1,154
L	587	50.0	587	50.0	14	1.2	30	2.6	90	7.7	218	18.6	773	65.8	1,174
M	596	51.9	552	48.1	17	1.5	30	2.6	68	5.9	225	19.6	758	66.0	1,148
N	579	49.0	602	51.0	18	1.5	37	3.1	76	6.4	243	20.6	752	63.7	1,181
O	563	48.5	597	51.5	16	1.4	35	3.0	72	6.2	226	19.5	763	65.8	1,160
P	540	46.2	628	53.8	14	1.2	28	2.4	73	6.3	252	21.6	744	63.7	1,168
Q	563	48.5	597	51.5	12	1.0	33	2.8	81	7.0	212	18.3	768	66.2	1,160
R	543	46.3	630	53.7	15	1.3	32	2.7	73	6.2	223	19.0	780	66.5	1,173

Form	Gender				Ethnicity										Total N
	Female		Male		American Indian		Asian		Black		Hispanic		White		
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	
S	542	46.3	629	53.7	13	1.1	32	2.7	77	6.6	236	20.2	764	65.2	1,171
T	572	48.0	620	52.0	18	1.5	38	3.2	65	5.5	249	20.9	735	61.7	1,192

Section 6: Psychometric Analyses

During the Spring 2022 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 Growth 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. 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

6.1. Number of Students 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, and IRT calibration). As in the previous technical reports since 2018, 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	TB-151	22,759
	4	TB-152	22,906
	5	TB-153	22,751
	6	TB-154	23,413
	7	TB-155	23,888
	8	TB-156	23,912
Mathematics	3	TB-157	22,739
	4	TB-158	22,878
	5	TB-159	22,723
	6	TB-160	23,382
	7	TB-161	23,842
	8	TB-162	23,862
Science	5	TB-163	22,726
	8	TB-164	23,851

6.2. Classical Item Analyses

This section summarizes the p -values and item-total correlations for operational and field test items. Appendix B and Appendix C provide the classical item-level statistics. 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 included for analyses.

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 B provides the summary p -value statistics by item type.

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
ELA															
3	564	0.501	0.162	0	1	5	5	25	118	157	122	78	24	18	12
4	647	0.535	0.160	0	1	8	4	20	79	145	189	119	49	18	16
5	561	0.516	0.163	0	1	8	11	23	74	146	144	89	46	12	8
6	542	0.501	0.180	0	1	14	5	39	81	144	116	67	52	14	10
7	539	0.517	0.183	0	1	15	3	25	76	145	122	80	43	11	19
8	553	0.509	0.175	0	1	4	8	32	102	152	120	64	35	14	22
Mathematics															
3	705	0.537	0.109	0	0.886	1	3	5	59	168	305	113	42	9	0
4	511	0.511	0.100	0.146	1	0	2	7	50	176	201	57	13	4	1
5	556	0.531	0.111	0.193	0.875	0	1	12	43	160	209	91	33	7	0
6	700	0.507	0.127	0	1	8	3	15	70	262	215	86	29	8	4
7	608	0.464	0.107	0.151	0.857	0	5	25	133	242	144	45	12	2	0
8	522	0.467	0.106	0	0.826	1	5	17	95	220	137	32	13	2	0
Science															
5	21	0.617	0.111	0.475	0.844	0	0	0	0	3	7	5	5	1	0
8	27	0.405	0.179	0.139	0.795	0	3	6	5	7	2	2	2	0	0

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
ELA															
3	257	0.475	0.157	0.141	0.889	0	6	30	49	66	51	36	14	5	0
4	218	0.534	0.155	0.169	0.899	0	4	11	26	50	58	37	19	13	0
5	240	0.505	0.158	0.137	0.913	0	5	22	35	58	50	41	23	5	1
6	238	0.487	0.166	0.083	0.891	1	10	18	47	55	43	41	16	7	0
7	230	0.524	0.146	0.100	0.877	0	1	13	34	50	65	40	16	11	0
8	225	0.542	0.154	0.167	0.892	0	1	11	28	55	47	47	23	13	0
Mathematics															
3	112	0.516	0.229	0.047	0.97	3	8	15	9	13	25	10	16	9	4
4	54	0.478	0.205	0.110	0.916	0	4	11	7	6	7	12	4	2	1
5	78	0.451	0.195	0.019	0.882	3	5	11	10	19	11	12	4	3	0
6	253	0.412	0.214	0.003	0.945	12	25	45	63	28	26	19	23	9	3
7	121	0.323	0.191	0.034	0.764	21	18	17	22	22	9	8	4	0	0
8	89	0.279	0.179	0.015	0.678	20	14	16	14	13	7	5	0	0	0
Science															
5	90	0.565	0.211	0.094	0.911	1	4	6	10	13	13	16	11	14	2
8	96	0.447	0.217	0.033	0.858	8	8	11	10	17	17	10	14	1	0

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 C 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.

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
ELA												
3	564	0.297	0.159	-0.559	1	44	69	163	169	78	30	11
4	647	0.303	0.166	-1	1	81	40	119	246	124	32	5
5	561	0.284	0.194	-0.974	0.967	72	48	135	168	110	23	5
6	542	0.276	0.213	-1	1	93	53	121	157	80	20	18
7	539	0.277	0.212	-1	1	86	58	107	155	104	20	9
8	553	0.27	0.204	-1	1	87	49	119	180	103	10	5
Mathematics												
3	705	0.372	0.084	0	0.754	4	14	113	307	225	41	1
4	511	0.364	0.093	0	0.804	3	22	82	221	162	20	1
5	556	0.362	0.092	0	0.979	4	15	109	229	173	25	1
6	700	0.356	0.097	-0.073	0.852	16	13	135	319	186	29	2
7	608	0.362	0.098	-0.468	0.774	4	26	110	258	172	35	3
8	522	0.36	0.088	0	1	3	14	104	240	142	17	2
Science												
5	21	0.472	0.091	0.254	0.614	0	0	1	3	8	8	1
8	27	0.439	0.109	0.211	0.652	0	0	3	6	9	7	2

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
ELA												
3	257	0.331	0.148	-0.18	0.61	22	26	53	61	63	31	1
4	218	0.341	0.134	-0.083	0.572	14	20	40	58	65	21	0
5	240	0.31	0.133	-0.196	0.565	17	33	55	70	50	15	0
6	238	0.324	0.141	-0.234	0.69	17	24	52	66	60	17	2
7	230	0.357	0.122	-0.097	0.577	7	16	48	68	68	23	0
8	225	0.341	0.118	-0.001	0.617	7	25	46	69	63	14	1
Mathematics												
3	112	0.396	0.121	0.062	0.662	2	8	14	30	37	18	3
4	54	0.44	0.122	0.075	0.671	1	2	4	9	20	16	2
5	78	0.447	0.132	0.138	0.672	0	5	5	16	24	17	11
6	253	0.396	0.129	-0.11	0.643	7	12	31	66	82	51	4
7	121	0.392	0.127	0.06	0.615	1	8	29	19	37	24	3
8	89	0.352	0.137	-0.073	0.62	3	12	11	26	25	11	1
Science												
5	90	0.346	0.113	0.032	0.559	4	4	20	31	25	6	0
8	96	0.337	0.117	0.011	0.608	4	6	25	30	27	3	1

6.2.3. Item Suppression

Table 6.6 and Table 6.7 present the 2022 flagging criteria for multiple-choice (MC) and non-MC operational items, respectively, which are the same as the 2021 criteria. Based on the item analysis conducted using the Spring 2022 administration results and removing items with n-counts less than 100 (statistics for items with $N < 100$ are considered to be unstable), 344 MC items and 35 non-MC items were identified for content and psychometric review.

Table 6.6. Flagging Criteria for MC Items

Flag Type*	Criterion	Indication
low item-total	< 0.20	poorly discriminating 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 recommends suppressing four items (two ELA and two mathematics items) from the 2022 scoring and removing them from the future item pool, as shown in Table 6.8. Following NDE approval, these suppressed items were not included for all subsequent analyses and score reporting. There was no suppression for science operational items.

Table 6.8. 2021 NSCAS Items to be Suppressed

Grade	Item Code	Item Role*	Item Type	Standard (Indicator)	Max. #Points	NWEA Recommendations	
						2022 Scoring	2023 Pool & Later
ELA							
5	VR431908	OP	Choice	LA 5.1.6.b	1	Suppress	Remove
7	VR431913	OP	Choice	LA 7.1.6.c	1	Suppress	Remove
Math							
3	VR406038	OP	Choice	MA 3.3.3.g	1	Suppress	Remove
5	VR446697	OP	Composite	MA 5.2.3.a	2	Suppress	Remove

*OP = operational. DO = diagnostic operational.

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.

Because fairness is a fundamental validity issue, it is essential that items be reviewed and assessed for DIF. Many methods for assessing DIF have been used and compared in conventional paper-pencil non-adaptive tests. However, DIF detection may be more important for CAT than it is for traditional paper-pencil non-adaptive tests with two reasons (Zwick, Thayer, & Wingersky, 1994): First, items with DIF may be more consequential for the examinees because fewer items are administered in a CAT. Second, several potential sources of DIF may be introduced, such as differential computer familiarity, facility, and anxiety. The difficulty of DIF analysis in the CAT is introduced by the fact that different sets of items are administered to different examinees. Therefore, the logistic regression (LR) procedure was applied to ELA and mathematics items that were administered in CAT, while the Mantel-Haenszel (MH) procedure was used to science items that were administered as a fixed form.

6.3.1. Logistic Regression (LR) DIF Method

The LR DIF procedure models item responses (for both dichotomous and polytomous items) as a function of group memberships, ability estimates, and their interaction. Testing for the presence of DIF based on logistic regression provides a model-based approach to identify

uniform and non-uniform DIF. DIF is classified as uniform if the effect is constant. That is, uniform DIF exists when the difference in the probabilities of a correct answer for the two groups is the same at all ability levels. DIF is classified as non-uniform if the effect varies conditional on the ability level. That is, non-uniform DIF exists if the interaction between item response function and group membership is disordinal.

The LR procedure compares the following three models (Fu & Monfils, 2016; Swaminathan & Rogers, 1990; Zumbo, 1999):

$$\text{Model 1: } \textit{logit}(P) = \beta_0 + \beta_1X + \beta_2E$$

$$\text{Model 2: } \textit{logit}(P) = \beta_0 + \beta_1X + \beta_2G + \beta_3E$$

$$\text{Model 3: } \textit{logit}(P) = \beta_0 + \beta_1X + \beta_2G + \beta_3XG + \beta_4E$$

Where:

- P is the probability of a test taker answering an item incorrectly (for a dichotomous item) and the probability of getting an item score or lower (for a polytomous item),
- X is the criterion variable,
- G is group membership,
- E is a vector including additional explanatory variables, and
- β are the associated regression parameters for model k .

For both dichotomous and polytomous items, Models 1, 2, and 3 are also referred to as a no DIF model, a uniform DIF model, and a nonuniform DIF model, respectively. The group estimates (β_2) are related with uniform DIF, and the interaction estimates (β_3) are associated with nonuniform DIF. *Proc Logistic* procedure in SAS was used in estimating the LR DIF. Note that for a dichotomously scored item the target probability that the LR estimates is the probability of answering an item incorrectly, which is different from the probability as answering an item correctly that many people may be accustomed to. Similarly, the target probability in the regression model for a polytomously scored item is the probability of obtaining an item score or below, to be consistent with that for a dichotomously scored item.

The item shows DIF if the modeled fit statistic is improved when group and interaction are added to the model, in order. To test the presence of nonuniform DIF, Model 2 and Model 3 are compared, using the likelihood ratio test with 1 degree of freedom (df) in chi-square distribution: $x^2 = [-2 \ln L(\text{Model2})] - [-2 \ln L(\text{Model3})]$.

Similarly, to test the presence of uniform DIF, Model 1 and Model 2 are compared, using the likelihood ratio test with 1 df:

$$x^2 = [-2 \ln L(\text{Model1})] - [-2 \ln L(\text{Model2})]$$

To test overall DIF (uniform DIF or nonuniform DIF), Model 1 and Model 3 are compared, using the likelihood ratio test with 2 df:

$$x^2 = [-2 \ln L(\text{Model1})] - [-2 \ln L(\text{Model3})]$$

The effect size is also used to avoid practically trivial but statistically significant results (French & Maller, 2007). Effect size is indicated by the difference of the Nagelkerke R^2 between two models (Gómez-Benito, Hidalgo, & Padilla, 2009). Table 6.10 presents the DIF classification rule for the LR DIF procedure used for NSCAS. This rule was confirmed to be consistent to the MH DIF classification rule for dichotomous items used by ETS (Fu & Monfils, 2016).

Table 6.10. LR DIF Categories

DIF Category	Level of DIF	Definition *
A	Negligible	χ^2 test is not significant at 0.05 level or $\Delta R^2 < 0.035$
B	Moderate	χ^2 test is significant at 0.05 level and $0.035 \leq \Delta R^2 < 0.070$
C	Strong	χ^2 test is significant at 0.05 level and $\Delta R^2 \geq 0.070$

* ΔR^2 is the Nagelkerke R^2 difference between two models.

6.3.2. Mantel-Haenszel (MH) DIF Methods

The 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:

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

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:

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

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:

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

where:

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

The SMD is divided by the total item group standard deviation to get a measure of the effect size. Table 6.11 and Table 6.12 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 6.11. MH DIF Categories for Dichotomous Items

DIF Category	Level of DIF	Definition
A	Negligible	• $MH \chi^2$ test is not significant at 0.05 level or $ MH \text{ D-DIF} < 1.0$.
B	Moderate	• $MH \chi^2$ test is not significant at 0.05 level and $1.0 \leq MH \text{ D-DIF} < 1.5$
C	Strong	• $MH \chi^2$ test is not significant at 0.05 level and $ MH \text{ D-DIF} \geq 1.5$.

* $|MH \text{ D-DIF}|$ = Absolute value of the Mantel-Haenszel delta difference.

Table 6.12. MH DIF Categories for Polytomous Items

DIF Category	Level of DIF	Definition
A	Negligible	$MH \chi^2$ test is not significant at 0.05 level or $ SMD/SD \leq 0.17$
B	Moderate	$MH \chi^2$ test is not significant at 0.05 level and $0.17 < SMD/SD \leq 0.25$
C	Strong	$MH \chi^2$ test is not significant at 0.05 level and $ SMD/SD > 0.25$

* SMD = Standardized mean difference. SD = Standard deviation.

6.3.3. DIF Results

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 100, which is reduced from 250 due to increased number of field test items. 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.

Tables 6.13 and 6.14 present the number of field test items assigned to each LR DIF category for DIF and NUIDIF, respectively, for ELA and mathematics. Considering that Rasch model is applied (i.e., the same slope is assumed for all items), UIDIF results are not reported. For the Spring 2021–2022 administration, item exposure was being controlled by the engine feature that assigns a weight to an item based on the number of the times the item is seen by students. The feature resulted in increased item pool usage, which is one of the desired properties that adaptive testing can achieve. However, it reduced the number of operational items meeting the minimum student counts required for DIF analyses, because all operational items were selected adaptively, while field test items were controlled to be distributed to have required students counts and to be administered across demographics. Thus, the DIF results for field test items in ELA and mathematics are reported.

Table 6.15. MH DIF Results—Operational Items (Science)

Table 6.15 and Table 6.16 present the number of items assigned to each MH DIF category for science operational and field test items, respectively. As shown in the tables, most items were categorized as DIF Category A (negligible DIF).

Table 6.13. LR DIF Results—Field Test Items (ELA/Mathematics)

Grade	Focal Group	#Items by DIF Category							C+	C-
		Total	A	B	B+	B-	C			
ELA										
3	Female	257	256	1	–	–	–	–	–	–
	Black or African American	–	–	–	–	–	–	–	–	–
	Hispanic	256	252	4	–	–	–	–	–	–
	Asian	–	–	–	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–	–	–	–
4	Female	218	215	1	–	1	1	–	–	–
	Black or African American	–	–	–	–	–	–	–	–	–
	Hispanic	218	217	1	–	–	–	–	–	–
	Asian	–	–	–	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–	–	–	–
5	Female	240	236	4	–	–	–	–	–	–
	Black or African American	–	–	–	–	–	–	–	–	–
	Hispanic	217	215	1	–	1	–	–	–	–
	Asian	–	–	–	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–	–	–	–
6	Female	238	236	2	–	–	–	–	–	–
	Black or African American	–	–	–	–	–	–	–	–	–
	Hispanic	238	238	–	–	–	–	–	–	–
	Asian	–	–	–	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–	–	–	–
7	Female	230	229	1	–	–	–	–	–	–
	Black or African American	–	–	–	–	–	–	–	–	–
	Hispanic	230	227	3	–	–	–	–	–	–
	Asian	–	–	–	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–	–	–	–
8	Female	225	223	–	1	1	–	–	–	–
	Black or African American	–	–	–	–	–	–	–	–	–
	Hispanic	211	211	–	–	–	–	–	–	–
	Asian	–	–	–	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–	–	–	–
Mathematics										
3	Female	112	112	–	–	–	–	–	–	–
	Black or African American	30	30	–	–	–	–	–	–	–
	Hispanic	112	111	1	–	–	–	–	–	–
	Asian	–	–	–	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–	–	–	–
4	Female	54	54	0–	–	–	–	–	–	–
	Black or African American	54	53	1	–	–	–	–	–	–

Grade	Focal Group	#Items by DIF Category					C+	C-
		Total	A	B	B+	B-		
ELA								
	Hispanic	54	54	-	-	-	-	-
	Asian	-	-	-	-	-	-	-
	Two or More Races	-	-	-	-	-	-	-
5	Female	78	76	2	-	-	-	-
	Black or African American	67	66	1	-	-	-	-
	Hispanic	78	78	-	-	-	-	-
	Asian	-	-	-	-	-	-	-
	Two or More Races	-	-	-	-	-	-	-
6	Female	253	251	1	-	1	-	-
	Black or African American	-	-	-	-	-	-	-
	Hispanic	253	212	31	-	7	3	-
	Asian	-	-	-	-	-	-	-
	Two or More Races	-	-	-	-	-	-	-
7	Female	121	120	1	-	-	-	-
	Black or African American	32	32	-	-	-	-	-
	Hispanic	121	119	2	-	-	-	-
	Asian	-	-	-	-	-	-	-
	Two or More Races	-	-	-	-	-	-	-
8	Female	89	88	-	-	1	-	-
	Black or African American	82	82	-	-	-	-	-
	Hispanic	89	89	-	-	-	-	-
	Asian	-	-	-	-	-	-	-
	Two or More Races	-	-	-	-	-	-	-

Table 6.14. LR UIDIF Results—Field Test Items (ELA/Mathematics)

Grade	Focal Group	#Items by DIF Category					
		Total	A	B+	B-	C+	C-
ELA							
3	Female	257	257	–	–	–	–
	Black or African American	–	–	–	–	–	–
	Hispanic	256	255	–	1	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
4	Female	218	216	–	2	–	–
	Black or African American	–	–	–	–	–	–
	Hispanic	218	217	–	1	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
5	Female	240	239	1	–	–	–
	Black or African American	–	–	–	–	–	–
	Hispanic	217	216	–	1	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
6	Female	238	237	–	1	–	–
	Black or African American	–	–	–	–	–	–
	Hispanic	238	238	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
7	Female	230	229	–	1	–	–
	Black or African American	–	–	–	–	–	–
	Hispanic	230	230	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
8	Female	225	223	1	1	–	–
	Black or African American	–	–	–	–	–	–
	Hispanic	211	211	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–

Grade	Focal Group	#Items by DIF Category					
		Total	A	B+	B-	C+	C-
Mathematics							
3	Female	112	112	–	–	–	–
	Black or African American	30	30	–	–	–	–
	Hispanic	112	111	–	1	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
4	Female	54	54	–	–	–	–
	Black or African American	54	53	–	1	–	–
	Hispanic	54	54	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
5	Female	78	76	–	2	–	–
	Black or African American	67	66	–	1	–	–
	Hispanic	78	78	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
6	Female	253	251	–	2	–	–
	Black or African American	–	–	–	–	–	–
	Hispanic	253	224	–	27	–	2
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
7	Female	121	120	–	1	–	–
	Black or African American	32	32	–	–	–	–
	Hispanic	121	120	–	1	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
8	Female	89	88	–	1	–	–
	Black or African American	82	82	–	–	–	–
	Hispanic	89	89	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–

Table 6.15. MH DIF Results—Operational Items (Science)

Grade	Focal Group	#Items by DIF Category					
		Total	A	B+	B-	C+	C-
Science							
5	Female	21	21	–	–	–	–
	Black or African American	21	20	–	1	–	–
	Hispanic	21	21	–	–	–	–
	Asian	21	20	–	–	–	1
	Two or More Races	21	21	–	–	–	–
8	Female	27	27	–	–	–	–
	Black or African American	27	27	–	–	–	–
	Hispanic	27	27	–	–	–	–
	Asian	27	25	1	1	–	–
	Two or More Races	27	27	–	–	–	–

Table 6.16. MH DIF Results—Field Test Items (Science)

Grade	Focal Group	#Items by DIF Category					
		Total	A	B+	B-	C+	C-
Science							
5	Female	90	75	1	11	1	2
	Black or African American	–	–	–	–	–	–
	Hispanic	–	–	–	–	–	–
	Asian	–	–	–	–	–	–
	Two or More Races	–	–	–	–	–	–
8	Female	96	90	–	4	–	2
	Black or African American	–	–	–	–	–	–
	Hispanic	23	19	1	3	–	–
	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 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 θ responding correctly to item i is as follows, where θ_j and b_i are the person and item parameters, respectively:

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

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

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

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 θ_j and b_i are the person and item parameters, respectively.

6.4.1. Summary IRT Item Statistics

Table 6.17 and

Table 6.18 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 6.17. Summary IRT Item Statistics—Operational Items

Grade	#Items	#Parameters	Mean	SD	Min.	Max.	Range (Max. – Min.)
ELA							
3	564	564	-0.682	1.208	-3.517	3.912	7.429
4	647	647	-0.554	1.215	-3.638	5.456	9.094
5	561	561	-0.257	1.196	-2.969	4.307	7.276
6	542	542	-0.096	1.146	-3.088	2.988	6.076
7	539	539	-0.007	1.039	-2.719	3.626	6.345
8	553	553	0.097	1.101	-2.442	4.211	6.653
Mathematics							
3	705	705	-0.633	1.431	-4.731	6.297	11.027
4	511	511	0.319	1.317	-2.879	5.079	7.958
5	556	556	0.298	1.334	-4.154	5.256	9.41
6	700	700	0.628	1.428	-3.653	5.355	9.007
7	608	608	1.209	1.387	-2.944	6.018	8.962
8	522	522	1.458	1.441	-2.403	5.542	7.945
Science							
5	21	21	0.101	0.926	-1.517	3.029	4.546
8	27	27	-0.053	1.073	-2.182	1.689	3.871

Table 6.18. Summary IRT Item Statistics—Field Test Items

Grade	#Items	#Parameters	Mean	SD	Min.	Max.	Range (Max. – Min.)
ELA							
3	257	341	-0.335	3.45	-42.212	42.778	84.99
4	218	295	-0.302	1.223	-7.570	2.304	9.873
5	240	341	0.047	4.805	-43.008	43.607	86.615
6	238	331	0.428	3.806	-46.212	45.344	91.556
7	230	338	0.308	3.517	-43.356	43.518	86.874
8	225	337	0.446	1.202	-5.650	5.636	11.286
Mathematics							
3	112	136	-0.232	1.645	-4.724	4.272	8.997
4	54	62	0.448	1.466	-2.878	3.425	6.303

Grade	#Items	#Parameters	Mean	SD	Min.	Max.	Range (Max. – Min.)
5	78	97	0.935	1.386	-2.063	5.501	7.564
6	253	294	1.321	1.522	-2.615	7.579	10.194
7	121	149	1.968	1.58	-0.965	5.536	6.501
8	89	108	2.51	1.496	-0.04	6.148	6.189
Science							
5	90	92	0.249	1.293	-2.189	3.637	5.826
8	96	102	-0.196	1.307	-2.664	3.408	6.072

6.5. Stability Check (ELA and Mathematics)

To evaluate the stability of the NCSAS scale, NWEA conducted horizontal equating for each grade. The first step was to identify items that were showing evidence of drift and update them. This is intended to minimize disruption to student test scores, considering that the provisional student scores are available before the final ones are reported. The Robust Z post-equating check procedure was used to identify items that show significant difficulty changes from the bank values, which has previously been used for Nebraska assessments.

Using all on-grade operational items administered to 50 or more students, the item difficulty equivalence was checked by comparing the current item calibration (i.e., pre-equating) with a new unanchored calibration of the 2022 data (i.e., post-equating) using Winsteps 4.8.0.0 (Linacre, 2021). The evaluations were conducted for each grade and content area 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. 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 ± 1.645 , was used. Items that exceeded the Robust Z critical value were flagged for large change. The next step was to obtain post-equated estimates by calibrating all flagged NSCAS operational items for each grade, while fixing the rest of the not-flagged item parameters. Then, post-equated student theta scores were obtained using the pre-equated estimates for not-flagged items and the post-equated estimates for flagged items. In comparing scores, descriptive statistics were compared between pre- and post-equated results for both NSCAS scores. Specifically, the mean, standard deviation, and percentage of students at each achievement level were computed for each grade. Effect sizes were also calculated. To determine whether the pre-equated results are plausible, the comparison criteria included the following:

- The difference in achievement level distributions between pre- and post-equated scores within 5%
- The absolute effect size between pre- and post-equated scores less than or equal to 0.1

The criterion of a 5% difference for the achievement level distributions has been used for NSCAS consistency checks in 2019. Also, it is within the bounds of the misclassification rate from the simulation reports for 2018 through 2021. The 5% criterion is also often used for large-scale state assessment programs. The effect size criterion of 0.1 is equivalent to the sampling criterion recommended by the External Experts Advisory (EEA) panel (i.e., one tenth of 1 standard deviation was used to approve the vertical linking set samples). Considering that the effect size of 0.2 is typically referred to as a small difference, the use of a more conservative criterion of 0.1 is proposed.

Table 6.19 presents the scale score descriptive statistics for pre-equated and post-equated scores and the difference. The effect sizes are all within NWEA’s criterion of 0.1 in absolute value. Table 6.20 presents the percentage of students in each achievement level for each year and the difference. The percentage difference is within the criterion (i.e., less than 5%), except for Grade 5 mathematics On Track (5.4%). However, the percentage difference for proficient or not (Developing or not Developing) is within the criterion for all grades and subjects.

The two sets of scoring results indicate that the current test scores are stable to maintain. Therefore, NWEA recommends using the current parameters for 2022 Spring scoring (i.e., keeping the 2022 Spring NSCAS scores unchanged) and updating the item parameters for future administrations and other psychometric analyses. NDE approved NWEA recommendations.

Table 6.19. Scale Score Difference Between Pre-equated and Post-equated score

Content	Grade	N	Pre-equated		Post-equated		Mean Difference (Post–Pre)	Effect Size
			Mean	SD	Mean	SD		
ELA	3	22,541	2465.52	88.30	2467.68	93.24	2.16	0.02
	4	22,703	2495.63	86.49	2494.32	94.92	-1.31	-0.01
	5	22,650	2515.79	81.77	2517.02	89.44	1.23	0.01
	6	23,222	2523.40	75.51	2525.01	79.15	1.61	0.02
	7	23,574	2531.19	77.29	2529.83	78.76	-1.36	-0.02
	8	23,668	2546.24	73.41	2546.83	74.16	0.59	0.01
Math	3	21,853	1187.84	81.54	1190.84	91.10	3.00	0.03
	4	21,811	1215.84	78.15	1217.08	87.96	1.24	0.01
	5	21,757	1231.11	75.48	1234.16	83.27	3.05	0.04
	6	22,702	1238.80	73.65	1238.43	81.83	-0.37	0.00
	7	23,215	1240.92	73.05	1240.21	82.43	-0.71	-0.01
	8	23,370	1250.89	76.58	1249.75	84.68	-1.14	-0.01

Table 6.20. Achievement Level Distributions

Content	Grade	N	Pre-Equated			Post- Equated			Difference (Post–Pre)		
			%Dev	%OT	%CCR	%Dev	%OT	%CCR	%Dev	%OT	%CCR
ELA	3	22,541	49.90	36.10	14.00	50.30	32.10	17.60	0.40	-4.00	3.60
	4	22,703	47.10	38.30	14.60	47.10	36.70	16.20	0.00	-1.60	1.60
	5	22,650	52.50	33.00	14.50	51.80	30.40	17.80	-0.70	-2.60	3.30
	6	23,222	56.00	31.10	12.90	55.40	28.60	16.00	-0.60	-2.50	3.10
	7	23,574	57.60	34.70	7.70	58.50	33.10	8.40	0.90	-1.60	0.70
	8	23,668	53.70	36.40	9.90	54.50	34.00	11.50	0.80	-2.40	1.60
Math	3	21,853	49.90	38.90	11.20	49.60	35.20	15.20	-0.30	-3.70	4.00
	4	21,811	53.60	36.30	10.10	53.30	32.90	13.80	-0.30	-3.40	3.70
	5	21,757	51.20	39.90	8.90	52.00	34.60	13.40	0.80	-5.30	4.50
	6	22,702	54.10	37.40	8.50	54.40	35.00	10.60	0.30	-2.40	2.10
	7	23,215	55.70	35.70	8.70	55.70	33.90	10.40	0.00	-1.80	1.70
	8	23,370	58.20	34.40	7.30	58.40	32.60	9.00	0.20	-1.80	1.70

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

6.6. Science Measurement Model

The new science assessment is designed to measure three-dimensional science learning, incorporating elements of Science and Engineering Practices (SEPs), Crosscutting Concepts (CCCs), and Disciplinary Core Ideas (DCIs) from the NCCRS-S. The new assessment design is based on performance tasks and associated prompts that lead students into more complex thinking and a focus on doing science rather than knowing discrete science facts. A small-scale pilot test was administered in March 2019 to glean meaningful information about the tasks that were used to inform field test development in Summer 2019. A full-scale field test was conducted in Spring 2021 due to the administration cancellation in 2020.

Spring 2022 science are operational field tests (i.e., all items were re-calibrated following the 2022 administration). The number of items and points possible are reported in Table 2.2. Spring 2022 science tests can be summarized as following:

- English online forms
 - Each grade has 20 different forms, but the operational items are the same across forms.
 - Grade 5 has 5 sets, including 4 operational sets and 1 field test set. Each field test set includes 4 to 6 items.
 - Grade 8 has 5 sets, including 5 operational sets and 1 field test set. Each field test set includes 4 to 7 items. A total of 111 field test items were administered.
 - The overall test score is based on 21 operational items (worth 22 points) in grade 5 and 27 operational items (worth 33 points) in grade 8. A total of 123 field test items were administered.
- Paper-pencil and Spanish forms
 - Each grade has 4 sets, including 3 operational sets and 1 field test set. Field test set includes 4 and 7 items for grades 5 and 8, respectively.
 - The overall test score is based on 17 operational items (worth 18 points) in grade 5 and 16 operational items (worth 20 points) in grade 8.

Science simulation and engine evaluations are summarized in Section 5. Following the Spring 2022 administration, a study was conducted to find the best measurement model for the Nebraska science assessments, as a continuation of an NWEA Spring 2021 examination.

Unidimensionality is the most commonly violated assumption in the latent trait structure implied by the item response data. However, due to the robustness of the model, 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. There is not a single statistic to determine the dimensionality of assessment data. Instead, for the preponderance of evidence from multiple approaches when determining dimensionality, the following analyses were conducted:

- a. Correlation between DCI, SEP and CCC
- b. Principal Component Analysis (PCA)
- c. Parallel Analysis
- d. Velicer's minimum average partial

In Grade 5, all analyses suggest a unidimensional solution fit the data. In Grade 8, all analyses except the parallel analyses suggest a unidimensional solution. In both grades, the preponderance of the evidence points to a single-factor solution. In other words, the

dimensionality study confirmed that the unidimensional measurement model is sufficient to model Nebraska science assessment in order to monitor and report student learning progress in science. There is no reason to consider a multi-dimensional model. Thus, the following unidimensional IRT models were applied to fit the data:

- Rasch one-parameter logistic (1PL) for dichotomous items and partial credit model (PCM) for polytomous items
- Two-parameter logistic (2PL) for dichotomous items and general partial credit model (GPCM) for polytomous items

Like the 2021 analysis, the model fit shows that the 2PL and GPCM combination model has the better fit, based on fit statistics. While statistically the 2PL and GPCM combination model fit the data better than the 1PL and PCM combination, those differences are small. Based on this analysis, NWEA believes that the 1PL and PCM combination model approach not only fit the data well, but also provided more reasonable item difficulty parameters. Further, this combination method has a long-standing practice in Nebraska and is the methodology used to calibrate the items in the current NSCAS assessments for ELA and mathematics as well as previous Nebraska State Accountability (NeSA) for ELA, mathematics, and science. NDE decided to move forward with the 1PL+PCM model for the science exams. The summary IRT item statistics using the 1PL and PCM is included in Section 6.4.

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 σ_{G5} is the standard deviation of the Grade 5 theta score.

The theta estimate, θ , and associated θ -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:

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

θ -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):

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

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

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

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 6.21 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 6.21. 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
ELA					
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
Mathematics					
3	1000	1470	1000	1001	1002

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
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
Science					
5	3000	3250	3000	3001	3002
8	3000	3250	3000	3001	3002

*Cut scores were determined in 2018 for ELA and mathematics and in 2022 for science.

Table 6.22 and Table 6.23 summarize 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 6.22. Conversion of Theta to Scale Scores (ELA/Mathematics)

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
ELA									
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
Mathematics									
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

*Cut scores were determined in 2018 for ELA and mathematics

Table 6.23. Conversion of Theta to Scale Scores (Science)

Grade	Scale Score Ranges			Cut Scores		Conversion		Cuts (Theta)*	
	Developing	On Track	Advanced	On Track	Advanced	Slope <i>b</i>	Intercept <i>a</i>	On Track	Advanced
Science									
5	3000–3099	3100–3149	3150–3250	3100	3150	23.2612	3099.8511	0.0064	2.1559
8	3000–3099	3100–3149	3150–3250	3100	3150	27.5346	3121.8432	-0.7933	1.0226

**Cut scores were determined in 2022 for science.

Section 7: Standard Setting

For ELA and mathematics, no standard setting was held in 2021–2022. 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). Standard setting took place for the new NSCAS science assessment following the first operational administration in Spring 2022.

7.1. ELA and Mathematics

7.1.1. Overview

In 2016–2017, the NSCAS ELA assessment 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.

7.1.2. 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 ordered item book (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.

7.1.3. ALD Revision

The ID Matching method requires clear ALDs that describe the knowledge, skills, and abilities 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. 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.

7.1.4. 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.

7.2. Science

The Nebraska Department of Education and Early Development (NDE) has partnered with NWEA and ACS Ventures, LLC (ACS) to establish cut scores for the Nebraska Student-Centered Assessment System (NSCAS) science assessments administered in grades 5 and 8. The first operational administration of the new NSCAS Science assessments was in the Spring of 2022. As part of the implementation, it was necessary to establish cut scores for the NSCAS science assessments which: (a) reflect the current Nebraska state standards, (b) link students' scores on the assessments to the state's expectations for students in each performance level, and (c) are articulated between grades 5 and 8. On July 6 and 7, 2022, NWEA and ACS worked with subject matter expert panelists from Nebraska to formulate recommended cut scores.

The NSCAS science assessment is formed around tasks, which are collections of prompts focused on a single phenomenon or problem. Tasks may use elements from the DCIs, SEPs, and CCCs in any combination. Each test is a series of tasks with associated items. Grade 5 includes 4 tasks with 21 total items and grade 8 includes 8 tasks with 27 items. There are three types of items that may appear on a test form:

- Single multiple-choice: These items are presented with the stimulus and are scored as correct (1) or incorrect (0)
- Composite multiple-choice: These items are presented as a pair and are scored as both correct (1) or incorrect (0)
- Polytomous: These items may be hot text and gap match types in which the student needs to make multiple hot text selections or drag multiple objects into gaps. These are scored as correct (2), partially correct (1) or incorrect (0)

7.2.1. Extended Angoff methodology

Student performance on each test were classified into one of three achievement levels: *Developing*, *On-Track*, or *Advanced*. Therefore, two cut scores were required to interpret test scores—one to distinguish *On-Track* performance from *Developing* performance and one to distinguish *Advanced* performance from *On-Track* performance.

The Extended Angoff methodology (Hambleton & Plake, 1995) was used to establish the recommendations for two cut scores for each assessment. Panelists were asked to work individually to consider the knowledge, skills, and abilities measured by each task and subsumed items and compare these against the Threshold ALDs. Panelists made a judgment for each item within each task as to how well they believe the *Just Barely On-Track* and *Just Barely Advanced* students will perform (i.e., what score they will earn).

7.2.2. Meeting Process

The meetings included a general training, grade-level training, standard setting judgments, and vertical articulation. The panelists participated in a general training where they learned about the purpose of the standard setting activities, how they would first conceptually define their expectations for each performance level (develop threshold ALDs), and then how they would translate these conceptual expectations into performance expectations (cut scores). The process continued with each grade level panel preparing to make standard setting judgments. This preparation included developing threshold ALDs from the range ALDs and practicing the standard setting judgmental process with a set of field test items. Each panel then had the opportunity to use the threshold ALDs to make standard setting judgments on a single test form in a process that included two rounds of judgments with discussion and feedback in between. The two panels then reconvened to share their results and compare the estimated impact (% of students in each achievement level) from their recommended cut scores.

7.3. Final Results

For ELA and mathematics, the recommended cut scores were presented to the Nebraska State Board of Education on August 2, 2018. For science, the recommended cut scores were presented to the Nebraska State Board of Education on August 5, 2022. Table 7.1 presents the final approved cut scores that were used for subsequent scoring. That is, the cuts have been used starting from 2018 for ELA and mathematics, while new cuts have been applied for science in 2022.

Table 7.1. Final Approved Cut Scores and Impact Data

Grade	Cut Scores*	
	On Track	CCR
ELA		
3	2477	2557
4	2500	2582
5	2531	2599
6	2543	2603
7	2556	2630
8	2561	2632
Mathematics		
3	1190	1286
4	1222	1317
5	1236	1331
6	1244	1342
7	1247	1346
8	1264	1365
Science		
5	3100	3150
8	3100	3150

*Cut scores were determined in 2018 for ELA and mathematics and in 2022 for science

Section 8: Test Results

All students who took the online, paper-pencil, and Spanish forms of the Spring 2022 NSCAS Growth 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. 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 41% to 45% are eligible for FRL, 9–17% have LEP status, and 14–17% belong to at least one SPED category. For all three of these programs/categories, the participation rate is slightly lower for upper-grade students.

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

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
Total N-Count		22,779	100.00	22,777	100.00
Gender	Female	11,056	48.54	11,050	48.51
	Male	11,723	51.46	11,727	51.49
Ethnicity	AI/AN	289	1.27	288	1.26
	Asian	760	3.34	760	3.34
	Black or African American	1,483	6.51	1,482	6.51
	Hispanic	4,701	20.64	4,701	20.64
	NH/PI	35	0.15	35	0.15
	White	14,402	63.22	14,403	63.23
	Two or More Races	1,109	4.87	1,108	4.86
FRL	Yes	9,998	43.89	10,000	43.90
	No	12,781	56.11	12,777	56.10
LEP	Yes	3,982	17.48	3,979	17.47
	No	18,797	82.52	18,798	82.53

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
SPED	Yes	3,865	16.97	3,867	16.98
	No	18,914	83.03	18,910	83.02
Universal Features & Accommodations	Text to Speech	4,458	19.57	4,485	19.69
	Basic Calculator	–	–	809	3.55
	Read Aloud	177	0.78	151	0.66
	One-on-One Setting	1,062	4.66	1,068	4.69
	Bilingual Dictionary/Word List	6	0.03	12	0.05
	Language Translation	–	–	21	0.09
	Mathematical Supports	–	–	1,000	4.39
	Assistive Technology	10	0.04	9	0.04
	Specialized Presentation	3	0.01	3	0.01
	Scribe	48	0.21	39	0.17
	Paper-Pencil (PP)	3	0.01	5	0.02
	Spanish Online	16	0.07	33	0.14
	Spanish Paper-Pencil (PP)	1	–	–	–
	Braille**	–	–	–	–
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 8.2. Number of Students Tested by Demographics and Accommodations—Grade 4

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
Total N-Count		22,927	100.00	22,929	100.00
Gender	Female	11,163	48.69	11,164	48.69
	Male	11,764	51.31	11,765	51.31
Ethnicity	AI/AN	305	1.33	304	1.33
	Asian	746	3.25	745	3.25
	Black or African American	1,473	6.42	1,471	6.42
	Hispanic	4,594	20.04	4,592	20.03
	NH/PI	43	0.19	43	0.19
	White	14,661	63.95	14,669	63.98
	Two or More Races	1,105	4.82	1,105	4.82
FRL	Yes	9,769	42.61	9,771	42.61
	No	13,158	57.39	13,158	57.39
LEP	Yes	3,725	16.25	3,721	16.23
	No	19,202	83.75	19,208	83.77
SPED	Yes	3,900	17.01	3,900	17.01
	No	19,027	82.99	19,029	82.99

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
Universal Features & Accommodations	Text to Speech	4,242	18.50	4,230	18.45
	Basic Calculator	–	–	1,002	4.37
	Read Aloud	162	0.71	138	0.60
	One-on-One Setting	1,103	4.81	1,120	4.88
	Bilingual Dictionary/Word List	9	0.04	11	0.05
	Language Translation	–	–	19	0.08
	Mathematical Supports	–	–	1,181	5.15
	Assistive Technology	27	0.12	25	0.11
	Specialized Presentation	20	0.09	21	0.09
	Scribe	42	0.18	34	0.15
	Paper-Pencil (PP)	7	0.03	5	0.02
	Spanish Online	13	0.06	45	0.20
	Spanish Paper-Pencil (PP)	1	–	1	–
	Braille**	1	–	1	–
Large Print**	3	–	3	–	

*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 and Accommodations—Grade 5

Demographic Sub-Group*		ELA		Mathematics		Science	
		N	%	N	%	N	%
Total N-Count		22,774	100.00	22,775	100.00	22,769	100.00
Gender	Female	11,074	48.63	11,072	48.61	11,070	48.62
	Male	11,700	51.37	11,703	51.39	11,699	51.38
Ethnicity	AI/AN	269	1.18	268	1.18	266	1.17
	Asian	697	3.06	696	3.06	698	3.07
	Black or African American	1,411	6.20	1,409	6.19	1,412	6.20
	Hispanic	4,594	20.17	4,596	20.18	4,591	20.16
	NH/PI	42	0.18	42	0.18	42	0.18
	White	14,698	64.54	14,700	64.54	14,696	64.54
	Two or More Races	1,063	4.67	1,064	4.67	1,064	4.67
FRL	Yes	9,712	42.65	9,710	42.63	9,421	41.38
	No	13,062	57.35	13,065	57.37	13,348	58.62
LEP	Yes	3,467	15.22	3,466	15.22	3,473	15.25
	No	19,307	84.78	19,309	84.78	19,296	84.75
SPED	Yes	3,748	16.46	3,745	16.44	3,836	16.85
	No	19,026	83.54	19,030	83.56	18,933	83.15

Demographic Sub-Group*		ELA		Mathematics		Science	
		N	%	N	%	N	%
Universal Features & Accommodations	Text to Speech	3,986	17.50	3,980	17.48	3,933	17.27
	Basic Calculator	–	–	1,180	5.18	–	–
	Read Aloud	138	0.61	131	0.58	139	0.61
	One-on-One Setting	1,193	5.24	1,173	5.15	1,163	5.11
	Bilingual Dictionary/Word List	12	0.05	23	0.10	22	0.10
	Language Translation	–	–	16	0.07	16	0.07
	Mathematical Supports	–	–	1,312	5.76	–	–
	Assistive Technology	35	0.15	22	0.10	23	0.10
	Specialized Presentation	23	0.10	22	0.10	30	0.13
	Scribe	34	0.15	31	0.14	29	0.13
	Paper-Pencil (PP)	5	0.02	5	0.02	5	0.02
	Spanish Online	18	0.08	47	0.21	38	0.17
	Spanish Paper-Pencil (PP)	–	–	–	–	–	–
Braille**	3	–	3	–	3	–	
Large Print**	2	–	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 8.4. Number of Students Tested by Demographics and Accommodations—Grade 6

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
Total N-Count		23,436	100.00	23,426	100.00
Gender	Female	11,417	48.72	11,417	48.74
	Male	12,019	51.28	12,009	51.26
Ethnicity	AI/AN	295	1.26	295	1.26
	Asian	690	2.94	691	2.95
	Black or African American	1,574	6.72	1,568	6.69
	Hispanic	4,787	20.43	4,785	20.43
	NH/PI	46	0.20	45	0.19
	White	14,964	63.85	14,960	63.86
	Two or More Races	1,080	4.61	1,082	4.62
FRL	Yes	10,129	43.22	10,120	43.20
	No	13,307	56.78	13,306	56.80
LEP	Yes	3,286	14.02	3,284	14.02
	No	20,150	85.98	20,142	85.98
SPED	Yes	3,502	14.94	3,503	14.95
	No	19,934	85.06	19,923	85.05

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
Universal Features & Accommodations	Text to Speech	3,605	15.38	3,567	15.23
	Basic Calculator	–	–	1,608	6.86
	Read Aloud	140	0.60	126	0.54
	One-on-One Setting	872	3.72	868	3.71
	Bilingual Dictionary/Word List	13	0.06	33	0.14
	Language Translation	–	–	14	0.06
	Mathematical Supports	–	–	1,099	4.69
	Assistive Technology	30	0.13	24	0.10
	Specialized Presentation	15	0.06	13	0.06
	Scribe	30	0.13	26	0.11
	Paper-Pencil (PP)	8	0.03	3	0.01
	Spanish Online	13	0.06	39	0.17
	Spanish Paper-Pencil (PP)	2	0.01	2	0.01
	Braille**	2	–	2	–
Large Print**	1	–	1	–	

*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 and Accommodations—Grade 7

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
Total N-Count		23,928	100.00	23,910	100.00
Gender	Female	11,654	48.70	11,640	48.68
	Male	12,274	51.30	12,270	51.32
Ethnicity	AI/AN	303	1.27	304	1.27
	Asian	640	2.67	639	2.67
	Black or African American	1,566	6.54	1,560	6.52
	Hispanic	5,041	21.07	5,040	21.08
	NH/PI	35	0.15	35	0.15
	White	15,292	63.91	15,287	63.94
	Two or More Races	1,051	4.39	1,045	4.37
FRL	Yes	10,191	42.59	10,174	42.55
	No	13,737	57.41	13,736	57.45
LEP	Yes	2,721	11.37	2,719	11.37
	No	21,207	88.63	21,191	88.63
SPED	Yes	3,507	14.66	3,502	14.65
	No	20,421	85.34	20,408	85.35

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
Universal Features & Accommodations	Text to Speech	3,243	13.55	3,262	13.64
	Scientific Calculator	–	–	1,731	7.24
	Read Aloud	118	0.49	97	0.41
	One-on-One Setting	816	3.41	821	3.43
	Bilingual Dictionary/Word List	22	0.09	40	0.17
	Language Translation	–	–	10	0.04
	Mathematical Supports	–	–	939	3.93
	Assistive Technology	42	0.18	37	0.15
	Specialized Presentation	8	0.03	8	0.03
	Scribe	18	0.08	16	0.07
	Paper-Pencil (PP)	9	0.04	9	0.04
	Spanish Online	31	0.13	59	0.25
	Spanish Paper-Pencil (PP)	–	–	–	–
	Braille**	2	–	2	–
Large Print**	3	–	3	–	

*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 and Accommodations—Grade 8

Demographic Sub-Group*		ELA		Mathematics		Science	
		N	%	N	%	N	%
Total N-Count		23,975	100.00	23,954	100.00	23,944	100.00
Gender	Female	11,600	48.38	11,589	48.38	11,585	48.38
	Male	12,375	51.62	12,365	51.62	12,359	51.62
Ethnicity	AI/AN	289	1.21	287	1.20	286	1.19
	Asian	683	2.85	683	2.85	682	2.85
	Black or African American	1,584	6.61	1,574	6.57	1,578	6.59
	Hispanic	4,916	20.50	4,908	20.49	4,904	20.48
	NH/PI	34	0.14	34	0.14	34	0.14
	White	15,418	64.31	15,416	64.36	15,409	64.35
	Two or More Races	1,051	4.38	1,052	4.39	1,051	4.39
FRL	Yes	9,795	40.86	9,774	40.80	9,203	38.44
	No	14,180	59.14	14,180	59.20	14,741	61.56
LEP	Yes	2,234	9.32	2,230	9.31	2,229	9.31
	No	21,741	90.68	21,724	90.69	21,715	90.69
SPED	Yes	3,293	13.74	3,284	13.71	3,329	13.90
	No	20,682	86.26	20,670	86.29	20,615	86.10

Demographic Sub-Group*		ELA		Mathematics		Science	
		N	%	N	%	N	%
Universal Features & Accommodations	Text to Speech	2,949	12.30	2,943	12.29	2,911	12.16
	Scientific Calculator	–	–	1,753	7.32	–	–
	Read Aloud	116	0.48	104	0.43	107	0.45
	One-on-One Setting	825	3.44	826	3.45	818	3.42
	Bilingual Dictionary/Word List	12	0.05	33	0.14	32	0.13
	Language Translation	–	–	9	0.04	8	0.03
	Mathematical Supports	–	–	840	3.51	–	–
	Assistive Technology	10	0.04	8	0.03	8	0.03
	Specialized Presentation	5	0.02	5	0.02	8	0.03
	Scribe	20	0.08	18	0.08	17	0.07
	Paper-Pencil (PP)	15	0.06	17	0.07	15	0.06
	Spanish Online	45	0.19	72	0.30	75	0.31
	Spanish Paper-Pencil (PP)	3	0.01	3	0.01	3	0.01
	Braille**	4	0.01	4	0.01	4	0.01
Large Print**	2	0.01	2	0.01	2	0.01	

*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 2022 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	%
ELA				
3	22,752	22,749	3	0.0
4	22,884	22,877	7	0.0
5	22,750	22,745	5	0.0
6	23,413	23,406	7	0.0
7	23,885	23,876	9	0.0
8	23,917	23,902	15	0.1
Mathematics				
3	22,738	22,733	5	0.0
4	22,879	22,874	5	0.0
5	22,721	22,716	5	0.0
6	23,368	23,365	3	0.0
7	23,829	23,820	9	0.0
8	23,856	23,839	17	0.1
Science				
5	22,727	22,723	4	0.0

8	23,856	23,841	15	0.1
---	--------	--------	----	-----

8.3. Testing Time

Table 8.8, Table 8.9, and Table 8.10 present the number of minutes students spent taking the Spring 2022 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	47	0.2	39	0.2	27	0.1	48	0.2	52	0.2	120	0.5
10 – <20	366	1.6	225	1.0	201	0.9	195	0.8	246	1.0	377	1.6
20 – <30	930	4.1	665	2.9	521	2.3	497	2.1	688	2.9	888	3.7
30 – <40	1,853	8.1	1,617	7.1	1,368	6.0	1,235	5.3	1,542	6.5	1,917	8.0
40 – <50	2,855	12.6	2,653	11.6	2,359	10.4	2,437	10.4	2,763	11.6	3,232	13.5
50 – <60	3,229	14.2	3,242	14.2	3,113	13.7	3,522	15.0	3,680	15.4	4,269	17.9
60 – <70	3,147	13.8	3,363	14.7	3,423	15.0	3,844	16.4	4,042	16.9	3,932	16.5
70 – <80	2,764	12.1	2,951	12.9	3,016	13.3	3,399	14.5	3,452	14.5	3,082	12.9
80 – <90	2,164	9.5	2,376	10.4	2,617	11.5	2,631	11.2	2,533	10.6	2,185	9.1
90 – <100	1,609	7.1	1,768	7.7	1,896	8.3	1,882	8.0	1,708	7.2	1,433	6.0
100 – <110	1,159	5.1	1,275	5.6	1,371	6.0	1,234	5.3	1,133	4.7	874	3.7
110 – <120	778	3.4	849	3.7	894	3.9	816	3.5	716	3.0	516	2.2
120 – <130	569	2.5	596	2.6	582	2.6	573	2.4	446	1.9	387	1.6
130 – <140	327	1.4	362	1.6	423	1.9	363	1.6	283	1.2	227	0.9
140 – <150	276	1.2	301	1.3	281	1.2	192	0.8	198	0.8	134	0.6
150 – <160	165	0.7	194	0.8	184	0.8	157	0.7	121	0.5	98	0.4
160 – <170	134	0.6	115	0.5	142	0.6	110	0.5	89	0.4	70	0.3
170 – <180	74	0.3	75	0.3	91	0.4	83	0.4	48	0.2	46	0.2
>=180	303	1.3	211	0.9	236	1.0	188	0.8	136	0.6	115	0.5
Total	22,749	100.0	22,877	100.0	22,745	100.0	23,406	100.0	23,876	100.0	23,902	100.0

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	21	0.1	14	0.1	12	0.1	28	0.1	46	0.2	76	0.3
10 – <20	288	1.3	181	0.8	122	0.5	136	0.6	244	1.0	281	1.2
20 – <30	1,055	4.6	590	2.6	521	2.3	474	2.0	591	2.5	771	3.2
30 – <40	2,652	11.7	1,596	7.0	1,435	6.3	1,108	4.7	1,223	5.1	1,517	6.4
40 – <50	3,927	17.3	2,777	12.1	2,746	12.1	2,002	8.6	2,143	9.0	2,616	11.0
50 – <60	3,948	17.4	3,717	16.2	3,695	16.3	2,873	12.3	3,116	13.1	3,522	14.8
60 – <70	3,223	14.2	3,459	15.1	3,671	16.2	3,302	14.1	3,417	14.3	3,876	16.3
70 – <80	2,416	10.6	3,011	13.2	3,059	13.5	3,183	13.6	3,321	13.9	3,276	13.7
80 – <90	1,651	7.3	2,191	9.6	2,362	10.4	2,681	11.5	2,751	11.5	2,435	10.2
90 – <100	1,069	4.7	1,648	7.2	1,678	7.4	2,186	9.4	2,039	8.6	1,757	7.4
100 – <110	739	3.3	1,131	4.9	1,096	4.8	1,586	6.8	1,511	6.3	1,279	5.4
110 – <120	519	2.3	802	3.5	723	3.2	1,154	4.9	1,081	4.5	833	3.5
120 – <130	369	1.6	495	2.2	496	2.2	815	3.5	691	2.9	559	2.3
130 – <140	274	1.2	360	1.6	337	1.5	556	2.4	478	2.0	313	1.3
140 – <150	147	0.6	243	1.1	206	0.9	360	1.5	297	1.2	231	1.0
150 – <160	136	0.6	184	0.8	162	0.7	268	1.1	258	1.1	126	0.5
160 – <170	72	0.3	139	0.6	123	0.5	174	0.7	174	0.7	100	0.4
170 – <180	44	0.2	98	0.4	77	0.3	125	0.5	116	0.5	76	0.3
>=180	183	0.8	238	1.0	195	0.9	354	1.5	323	1.4	195	0.8
Total	22,733	100.0	22,874	100.0	22,716	100.0	23,365	100.0	23,820	100.0	23,839	100.0

Table 8.10. Testing Time in Minutes—Science

Time in Minutes	Grade 5		Grade 8	
	N	%	N	%
<10	103	0.5	209	0.9
10 – <20	2,079	9.1	802	3.4
20 – <30	7,492	33.0	2,864	12.0
30 – <40	6,727	29.6	6,163	25.9
40 – <50	3,546	15.6	6,006	25.2
50 – <60	1,605	7.1	3,805	16.0
60 – <70	650	2.9	1,947	8.2
70 – <80	275	1.2	904	3.8
80 – <90	124	0.5	503	2.1
90 – <100	58	0.3	258	1.1
100 – <110	33	0.1	157	0.7
110 – <120	18	0.1	77	0.3
120 – <130	4	0.0	53	0.2
130 – <140	6	0.0	23	0.1
140 – <150	2	0.0	24	0.1
150 – <160	-	0.0	15	0.1
160 – <170	-	0.0	3	0.0
170 – <180	1	0.0	7	0.0
>=180	-	0.0	21	0.1
Total	22,723	100.0	23,841	100.0

8.4. Achievement Level Distributions

Table 8.11 presents the achievement level distributions for the Spring 2022 NSCAS assessments. Appendix D provides the achievement level distributions by demographic group. For ELA, 47–57% of students are at Developing and 42–53% of students are at On Track or CCR Benchmark. For mathematics, 49–58% of students are at Developing and 42–50% of students are at On Track or CCR Benchmark. For science, 29–37% of students are at Developing and 63–71% are at On Track or Advanced.

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	%
ELA									
3	22,752	11,334	49.8	8,227	36.2	3,191	14.0	11,418	50.2
4	22,884	10,788	47.1	8,781	38.4	3,315	14.5	12,096	52.9
5	22,750	11,943	52.5	7,460	32.8	3,347	14.7	10,807	47.5
6	23,413	13,110	56.0	7,276	31.1	3,027	12.9	10,303	44.0
7	23,885	13,765	57.6	8,281	34.7	1,839	7.7	10,120	42.4
8	23,917	12,835	53.7	8,697	36.4	2,385	10.0	11,082	46.3
Mathematics									
3	22,738	11,356	49.9	8,827	38.8	2,555	11.2	11,382	50.1
4	22,879	12,259	53.6	8,302	36.3	2,318	10.1	10,620	46.4
5	22,721	11,548	50.8	9,117	40.1	2,056	9.0	11,173	49.2
6	23,368	12,633	54.1	8,744	37.4	1,991	8.5	10,735	45.9
7	23,829	13,217	55.5	8,527	35.8	2,085	8.7	10,612	44.5
8	23,856	13,870	58.1	8,232	34.5	1,754	7.4	9,986	41.9
Science									
5	22,727	6,538	28.8	12,450	54.8	3,739	16.5	16,189	71.2
8	23,856	8,751	36.7	12,997	54.5	2,108	8.8	15,105	63.3

*Achievement levels for ELA and mathematics = Level 3: Developing, Level 2: On Track, and Level 1: CCR Benchmark. Achievement levels for science = Level 3: Developing, Level 2: On Track, and Level 1: Advanced.

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 D 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	LOSS	HOSS	Min	Max	Mean	SD	Percentiles						
								P5	P10	P25	P50	P75	P90	P95
ELA														
3	22,752	2220	2840	2222	2833	2465.67	88.23	2295	2339	2412	2477	2528	2570	2593
4	22,884	2250	2850	2252	2850	2495.52	86.32	2331	2372	2446	2505	2557	2598	2621
5	22,750	2280	2860	2282	2780	2516.16	81.85	2359	2406	2467	2526	2573	2612	2636
6	23,413	2290	2870	2292	2797	2523.34	75.48	2382	2417	2478	2533	2575	2611	2633
7	23,885	2300	2880	2302	2759	2531.12	77.33	2385	2422	2483	2543	2585	2621	2642
8	23,917	2310	2890	2312	2823	2546.26	73.41	2401	2449	2509	2554	2596	2631	2651
Mathematics														
3	22,738	1000	1470	1002	1470	1187.78	81.53	1052	1081	1130	1190	1242	1291	1322
4	22,879	1010	1500	1012	1500	1215.83	78.11	1086	1113	1162	1214	1269	1318	1347
5	22,721	1020	1510	1022	1510	1231.90	75.55	1108	1132	1180	1234	1282	1326	1355
6	23,368	1030	1530	1032	1530	1238.93	73.74	1113	1147	1192	1236	1286	1334	1364
7	23,829	1040	1540	1042	1540	1241.28	73.19	1124	1149	1194	1236	1287	1337	1371
8	23,856	1050	1550	1052	1550	1250.99	76.67	1125	1152	1201	1249	1298	1349	1383
Science														
5	22,727	3000	3250	3007	3218	3115.85	31.46	3068	3074	3095	3115	3133	3150	3163
8	23,856	3000	3250	3002	3238	3107.46	30.66	3061	3068	3086	3108	3127	3146	3155

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–Table 8.18Table 8.13 provide the reporting category correlations, and Table 8.19–Table 8.24 present the disattenuated correlations.

The correlations between reporting categories within the content areas are positive and moderate in value, ranging from 0.54 (between Reading Vocabulary and Writing Skills for both Grade 7 and 8) to 0.78 (between Number and Geometry for Grade 3 and between Number and Algebra for Grade 4). The correlations between reporting categories across the content areas are positive and low to moderate in value, ranging from 0.46 (between Reading Vocabulary and Data for Grade 6) and 0.68 (between Reading Comprehension and Number for Grade 3). 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: 0.82 (between Geometry and Data for Grade 4) or higher. The disattenuated correlations between reporting categories across the content areas are positive and moderate in value, ranging from 0.72 (between Reading Comprehension 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	0.76	1.00					
Writing Skills	0.65	0.69	1.00				
Number	0.64	0.68	0.62	1.00			
Algebra	0.60	0.64	0.58	0.77	1.00		
Geometry	0.61	0.64	0.58	0.78	0.72	1.00	
Data	0.60	0.64	0.58	0.75	0.69	0.69	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	0.74	1.00					
Writing Skills	0.63	0.70	1.00				
Number	0.59	0.65	0.57	1.00			
Algebra	0.61	0.66	0.58	0.78	1.00		
Geometry	0.58	0.62	0.53	0.71	0.68	1.00	
Data	0.51	0.55	0.48	0.64	0.64	0.58	1.00

Table 8.15. Reporting Category Correlations—Grade 5

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00						
Reading Comprehension	0.70	1.00					
Writing Skills	0.61	0.69	1.00				
Number	0.58	0.64	0.60	1.00			
Algebra	0.55	0.62	0.57	0.76	1.00		
Geometry	0.55	0.60	0.55	0.71	0.66	1.00	
Data	0.57	0.64	0.60	0.70	0.66	0.63	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	0.64	1.00					
Writing Skills	0.58	0.66	1.00				
Number	0.52	0.60	0.55	1.00			
Algebra	0.56	0.65	0.59	0.76	1.00		
Geometry	0.51	0.59	0.53	0.69	0.72	1.00	
Data	0.46	0.53	0.49	0.61	0.64	0.59	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	0.68	1.00					
Writing Skills	0.54	0.64	1.00				
Number	0.50	0.58	0.50	1.00			
Algebra	0.57	0.66	0.56	0.73	1.00		
Geometry	0.49	0.56	0.48	0.63	0.68	1.00	
Data	0.55	0.62	0.52	0.68	0.74	0.65	1.00

Table 8.18. Reporting Category Correlations—Grade 8

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00						
Reading Comprehension	0.66	1.00					
Writing Skills	0.54	0.62	1.00				
Number	0.50	0.55	0.50	1.00			
Algebra	0.55	0.61	0.55	0.72	1.00		
Geometry	0.52	0.58	0.51	0.70	0.74	1.00	
Data	0.53	0.59	0.53	0.65	0.73	0.69	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	0.99	1.00					
Writing Skills	0.97	0.93	1.00				
Number	0.82	0.79	0.82	1.00			
Algebra	0.85	0.82	0.85	0.97	1.00		
Geometry	0.83	0.78	0.81	0.94	0.96	1.00	
Data	0.86	0.83	0.86	0.96	0.97	0.93	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	1.04	1.00					
Writing Skills	0.99	0.93	1.00				
Number	0.82	0.76	0.75	1.00			
Algebra	0.88	0.81	0.79	0.95	1.00		
Geometry	0.85	0.77	0.74	0.88	0.87	1.00	
Data	0.81	0.74	0.72	0.85	0.89	0.82	1.00

Table 8.21. Reporting Category Disattenuated Correlations—Grade 5

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00						
Reading Comprehension	1.01	1.00					
Writing Skills	0.97	0.93	1.00				
Number	0.82	0.77	0.79	1.00			
Algebra	0.82	0.78	0.79	0.94	1.00		
Geometry	0.82	0.75	0.76	0.88	0.86	1.00	
Data	0.91	0.86	0.89	0.93	0.92	0.88	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	0.95	1.00					
Writing Skills	0.94	0.90	1.00				
Number	0.78	0.76	0.76	1.00			
Algebra	0.82	0.80	0.79	0.95	1.00		
Geometry	0.79	0.77	0.75	0.91	0.92	1.00	
Data	0.74	0.72	0.72	0.83	0.85	0.83	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	0.95	1.00					
Writing Skills	0.89	0.90	1.00				
Number	0.75	0.74	0.75	1.00			
Algebra	0.81	0.80	0.79	0.94	1.00		
Geometry	0.76	0.74	0.74	0.89	0.91	1.00	
Data	0.81	0.78	0.77	0.91	0.94	0.91	1.00

Table 8.24. Reporting Category Disattenuated Correlations—Grade 8

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00						
Reading Comprehension	0.99	1.00					
Writing Skills	0.93	0.90	1.00				
Number	0.77	0.71	0.74	1.00			
Algebra	0.82	0.76	0.79	0.92	1.00		
Geometry	0.78	0.73	0.74	0.91	0.93	1.00	
Data	0.83	0.78	0.80	0.87	0.96	0.91	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 2022. As shown in the table, the correlation coefficients range from 0.81 to 0.86 for ELA/reading and 0.88 to 0.90 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.
ELA/Reading										
3	12,802	0.85	2459	88.34	2223	2715	197	16.63	137	244
4	12,549	0.86	2490	87.60	2253	2850	205	16.45	139	254
5	12,226	0.85	2512	82.00	2283	2780	211	15.99	140	263
6	11,862	0.84	2516	76.72	2293	2774	213	16.29	149	262
7	10,787	0.83	2524	78.51	2303	2757	216	16.80	155	267
8	10,595	0.81	2539	73.31	2313	2762	219	16.96	150	266
Mathematics										
3	12,335	0.89	1184	80.52	1003	1470	202	15.22	131	257
4	12,595	0.90	1211	79.52	1020	1500	211	17.02	133	290
5	11,929	0.90	1226	75.77	1024	1510	218	18.58	135	288
6	11,979	0.89	1230	72.29	1033	1530	222	17.33	150	282
7	11,447	0.89	1234	71.44	1052	1540	225	19.04	146	297
8	10,761	0.88	1242	76.08	1053	1550	229	20.49	150	321

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

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 2022 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
- Classification Accuracy

Combined, these data provide several ways of looking at the reliability of the NSCAS 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 all at or above 0.80 (the ELA and mathematics estimates are all 0.88 and 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 E 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 ends 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	22,752	0.91	0.69	0.85	0.65
4	22,884	0.91	0.60	0.84	0.68
5	22,750	0.90	0.59	0.82	0.67
6	23,413	0.89	0.57	0.80	0.68
7	23,885	0.89	0.61	0.83	0.61
8	23,917	0.88	0.56	0.79	0.60

Table 9.2. Marginal Reliability of Scale Scores—Mathematics

Grade	N	Total Score	Number	Algebra	Geometry	Data
3	22,738	0.95	0.87	0.72	0.78	0.70
4	22,879	0.94	0.86	0.79	0.77	0.66
5	22,721	0.94	0.85	0.77	0.77	0.67
6	23,368	0.94	0.78	0.83	0.73	0.69
7	23,829	0.93	0.74	0.82	0.68	0.75
8	23,856	0.94	0.76	0.80	0.79	0.73

Table 9.3. Marginal Reliability of Scale Scores—Science

Grade	N	Total Score
5	22,727	0.80
8	23,856	0.85

Table 9.4. Marginal Reliability: Variance

Content Area	Grade	N	Variance	MSE	Overall	Deciles										
						1	2	3	4	5	6	7	8	9	10	
ELA	3	22,752	7784.32	676.98	0.91	0.87	0.91	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.90
	4	22,884	7451.86	691.82	0.91	0.89	0.91	0.92	0.92	0.92	0.92	0.92	0.91	0.91	0.90	0.87
	5	22,750	6699.23	684.36	0.90	0.87	0.90	0.90	0.91	0.91	0.91	0.91	0.91	0.91	0.90	0.87
	6	23,413	5696.88	652.38	0.89	0.84	0.88	0.89	0.90	0.90	0.90	0.90	0.90	0.89	0.89	0.86
	7	23,885	5979.30	637.81	0.89	0.84	0.88	0.89	0.90	0.90	0.91	0.91	0.91	0.91	0.90	0.88
	8	23,917	5388.66	666.05	0.88	0.83	0.87	0.88	0.88	0.89	0.89	0.89	0.89	0.89	0.89	0.87
Mathematics	3	22,738	6647.46	364.44	0.95	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.94	0.93	
	4	22,879	6101.66	350.62	0.94	0.93	0.94	0.95	0.95	0.95	0.95	0.95	0.94	0.94	0.93	
	5	22,721	5707.58	336.10	0.94	0.93	0.94	0.94	0.94	0.95	0.95	0.95	0.94	0.94	0.92	
	6	23,368	5438.32	346.28	0.94	0.92	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.93	
	7	23,829	5356.32	360.05	0.93	0.91	0.93	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.93	
	8	23,856	5877.94	368.87	0.94	0.91	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.95	0.94	
Science	5	22,727	989.43	196.98	0.80	0.82	0.88	0.88	0.88	0.88	0.88	0.85	0.82	0.74	0.35	
	8	23,856	939.90	141.06	0.85	0.64	0.82	0.86	0.87	0.89	0.89	0.89	0.89	0.89	0.82	

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 ± 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 25.1 to 26.2) than for mathematics (from 18.2 to 19.1). 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	2,139	24.6	0.8	1,564	25.6	0.9
	4	2,405	24.6	0.8	1,557	27.1	1.0
	5	2,592	24.7	0.9	1,717	26.0	1.0
	6	3,060	23.9	0.9	1,901	25.6	0.9
	7	3,037	23.4	1.0	1,212	25.0	0.7
	8	3,288	24.7	0.9	1,564	25.0	1.0
Mathematics	3	2,417	18.1	0.4	1,089	19.6	0.7
	4	2,313	18.2	0.7	1,097	19.0	0.5
	5	2,488	17.1	0.7	973	18.8	0.7
	6	2,625	18.3	0.7	916	18.0	0.5
	7	2,571	18.1	0.9	788	18.5	0.8
	8	2,771	18.2	0.7	735	18.7	0.9
Science	5	6,197	11.0	0.0	3,451	14.9	1.0
	8	6,050	10.6	0.5	2,604	11.1	0.8

*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	25.9	31.1	26.5	25.0	24.6	24.6	24.7	24.8	25.1	25.6	27.2
	4	26.2	28.3	25.8	25.1	24.7	24.6	24.8	25.2	25.7	27.1	30.7
	5	26.1	29.2	26.1	25.4	25.2	25.0	24.6	24.5	24.8	26.0	29.9
	6	25.4	29.9	25.7	24.7	24.2	23.9	23.9	24.1	24.7	25.4	27.8
	7	25.1	30.6	26.3	25.2	24.8	24.2	23.5	23.1	23.1	23.8	26.5
	8	25.7	30.5	26.7	25.4	25.0	24.8	24.7	24.5	24.5	24.7	26.4
Mathematics	3	19.0	19.8	18.9	18.3	18.1	18.1	18.2	18.4	19.0	19.3	22.0
	4	18.7	20.5	18.7	18.1	17.7	17.9	18.2	18.3	18.4	18.7	20.3
	5	18.2	19.4	18.0	17.9	18.0	17.4	17.0	17.2	17.9	18.3	21.2
	6	18.6	20.6	18.8	18.3	18.3	18.3	18.3	18.2	18.1	17.8	19.1
	7	18.9	21.8	19.9	19.1	18.4	18.2	18.1	18.1	18.1	18.2	19.1
	8	19.1	22.3	20.1	19.4	19.1	18.4	18.2	18.1	18.0	18.1	19.5
Science	5	13.3	13.2	11.0	11.0	11.0	11.0	11.0	12.0	13.5	16.0	23.9
	8	11.6	17.9	13.0	11.5	11.0	10.0	10.0	10.0	10.0	10.2	12.7

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), 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 and achievement level. Overall classification accuracy ranges from 0.816 (ELA Grade 6) to 0.899 (mathematics Grades 7 and 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 2022 suggests that accurate level classifications are being made for Nebraska students on the NSCAS assessments. Overall classification accuracy by achievement level ranges from 0.679 (science Grade 5 Advanced) to 0.935 (mathematics Grades 4 and 5 Developing).

Table 9.7. Classification Accuracy by Achievement Level—ELA

Grade	Achievement Level	N	%	Expected Proportion*			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
3	Developing	11,334	0.50	0.46	0.04	0.00	0.920	0.846
	On Track	8,227	0.36	0.05	0.28	0.04	0.762	
	CCR Benchmark	3,191	0.14	0.00	0.03	0.11	0.800	
4	Developing	10,788	0.47	0.43	0.04	0.00	0.907	0.834
	On Track	8,781	0.38	0.05	0.29	0.04	0.758	
	CCR Benchmark	3,315	0.15	0.00	0.03	0.12	0.800	
5	Developing	11,943	0.53	0.48	0.05	0.00	0.909	0.827
	On Track	7,460	0.33	0.05	0.23	0.04	0.713	
	CCR Benchmark	3,347	0.15	0.00	0.03	0.12	0.789	
6	Developing	13,110	0.56	0.51	0.05	0.00	0.904	0.816
	On Track	7,276	0.31	0.06	0.21	0.04	0.672	
	CCR Benchmark	3,027	0.13	0.00	0.03	0.10	0.783	
7	Developing	13,765	0.58	0.52	0.05	0.00	0.910	0.843
	On Track	8,281	0.35	0.06	0.26	0.03	0.749	
	CCR Benchmark	1,839	0.08	0.00	0.02	0.06	0.766	
8	Developing	12,835	0.54	0.48	0.06	0.00	0.890	0.818
	On Track	8,697	0.36	0.06	0.26	0.04	0.725	
	CCR Benchmark	2,385	0.10	0.00	0.02	0.08	0.760	

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

Table 9.8. Classification Accuracy by Achievement Level—Mathematics

Grade	Achievement Level	N	%	Expected Proportion*			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
3	Developing	11,356	0.50	0.47	0.03	0.00	0.934	0.893
	On Track	8,827	0.39	0.04	0.33	0.02	0.851	
	CCR Benchmark	2,555	0.11	0.00	0.02	0.10	0.866	
4	Developing	12,259	0.54	0.50	0.04	0.00	0.935	0.898
	On Track	8,302	0.36	0.03	0.31	0.02	0.854	
	CCR Benchmark	2,318	0.10	0.00	0.02	0.09	0.861	
5	Developing	11,548	0.51	0.48	0.03	0.00	0.935	0.896
	On Track	9,117	0.40	0.04	0.34	0.02	0.858	
	CCR Benchmark	2,056	0.09	0.00	0.01	0.08	0.856	

Grade	Achievement Level	N	%	Expected Proportion*			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
6	Developing	12,633	0.54	0.50	0.04	0.00	0.924	0.895
	On Track	8,744	0.37	0.04	0.32	0.02	0.861	
	CCR Benchmark	1,991	0.09	0.00	0.01	0.07	0.859	
7	Developing	13,217	0.56	0.52	0.04	0.00	0.930	0.899
	On Track	8,527	0.36	0.04	0.31	0.01	0.858	
	CCR Benchmark	2,085	0.09	0.00	0.01	0.08	0.874	
8	Developing	13,870	0.58	0.54	0.04	0.00	0.931	0.899
	On Track	8,232	0.35	0.04	0.29	0.01	0.852	
	CCR Benchmark	1,754	0.07	0.00	0.01	0.06	0.865	

*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		
5	Developing	6,538	0.29	0.26	0.03	0.00	0.892	0.820
	On Track	12,450	0.55	0.07	0.45	0.03	0.823	
	Advanced	3,739	0.17	0.00	0.05	0.11	0.679	
8	Developing	8,751	0.37	0.33	0.04	0.00	0.896	0.852
	On Track	12,997	0.55	0.07	0.46	0.02	0.840	
	Advanced	2,108	0.09	0.00	0.02	0.07	0.739	

*L3: Developing, L2: On Track, and L1: Advanced.

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, σ_x^2 = the total score variance, and σ_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).

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

Grade	Demographic Group*		#Items	Reliability	SEM
5	Grade 5 Overall		21	0.84	12.58
	Gender	Female	21	0.83	12.45
		Male	21	0.85	12.62
	Ethnicity	AI/AN	21	0.81	12.17
		Asian	21	0.85	12.75
		Black or African American	21	0.81	11.87
		Hispanic	21	0.82	12.19
		NH/PI	21	0.84	11.66
		White	21	0.82	12.83
		Two or More Races	21	0.83	12.38
	FRL	Yes	21	0.83	12.13
		No	21	0.82	12.84
	LEP	Yes	21	0.81	11.97
		No	21	0.83	12.80
	SPED	Yes	21	0.84	11.98
No		21	0.82	12.79	
8	Grade 8 Overall		27	0.85	11.87
	Gender	Female	27	0.84	11.86
		Male	27	0.86	11.80
	Ethnicity	AI/AN	27	0.80	12.26
		Asian	27	0.88	11.63
		Black or African American	27	0.81	12.55
		Hispanic	27	0.82	12.00
		NH/PI	27	0.83	11.81
		White	27	0.83	11.82
		Two or More Races	27	0.85	12.08
	FRL	Yes	27	0.83	12.02
		No	27	0.84	11.65
	LEP	Yes	27	0.77	12.46
		No	27	0.84	11.98
	SPED	Yes	27	0.80	12.22
No		27	0.84	11.71	

*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.54). 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).

While NSCAS assessments offer the additional benefit of reporting category scores that indicate directions for gaining further instructional information through the interim system or classroom observation, scores based on NSCAS are equally reliable and valid as the traditional end of year assessment due to the following factors. First, NSCAS assessments go through the same rigorous psychometric analyses such as test reliability, classification accuracy, CSEMs, test information, DIF, and convergent validity check, and the analysis results we have so far strongly support the reliability and validity claim of NSCAS assessments. In addition, the test development process ensures validity of the intended test score interpretations provided through the Reporting ALDs and scale score. Last but not the least, NSCAS assessments are aligned to grade-level content and their test scores are suitable for use in accountability systems, as a result of a robust development process of table of specifications (TOS), passage and item specifications, and achievement level descriptor (ALD).

10.1. Intended Purposes and Uses of Test Scores

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

1. To measure and report Nebraska students’ depth of achievement regarding the Nebraska College and Career Ready Standards
2. To report if student achievement is sufficient academic proficiency 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

As the *Standards* notes, “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 refers 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.

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 General 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 the Nebraska College and Career Ready Standards.	✓	✓	✓	✓
2. Report if student achievement is sufficient academic proficiency to be on track for achieving college readiness.	✓	✓	✓	
3. Measure students' annual progress toward college and career readiness.	✓	✓	✓	
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 the Nebraska College and Career Ready Standards.	<ul style="list-style-type: none"> • Bias is minimized through Universal Design and accessibility resources. • Blueprint, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity. • The item pool and item selection procedures adequately support the test design. 	2, 9
2. Report if student achievement is sufficient academic proficiency to be on track for achieving college readiness.	<ul style="list-style-type: none"> • Nebraska's College and Career Ready Standards are based on skills leading to college and career readiness across grades. • Blueprint, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity. 	2
3. Measure students' annual progress toward college and career readiness.	<ul style="list-style-type: none"> • Nebraska's College and Career Ready Standards are based on skills leading to college and career readiness across grades. • Blueprint, passage specifications and item specifications are aligned to grade-level content, process skills, and associated cognitive complexity. 	2

Test Purpose	Summary of Evidence	Tech Report Sections
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"> • Blueprint, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity. • Blueprint and ALDs were developed in consultation with Nebraska educators. • Reporting categories align with the structure of the Nebraska standards to support the interpretation of the test results. 	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"> • Bias is minimized through Universal Design and accessibility resources. • Assessments are administered with appropriate accommodations. 	2, 3, 6, 9

Table 10.3. Sources of Validity Evidence based on Response Process

Test Purpose	Summary of Evidence	Tech Report Sections
1. Measure and report Nebraska students' depth of achievement regarding the Nebraska College and Career Ready Standards.	<ul style="list-style-type: none"> • Bias is minimized through Universal Design and accessibility resources. • Blueprint, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity. • Achievement levels were set consistent with best practice. 	2
2. Report if student achievement is sufficient academic proficiency to be on track for achieving college readiness.	<ul style="list-style-type: none"> • Blueprint, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity. • Achievement levels were vertically articulated. 	2
3. Measure students' annual progress toward college and career readiness.	<ul style="list-style-type: none"> • Blueprint, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity. • Achievement levels were vertically articulated. 	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"> • Blueprint, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity. • Range and Policy ALDs were developed in consultation with Nebraska educators with the goal of providing information to Nebraska educators. 	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"> • Bias is minimized through Universal Design and accessibility resources. • Assessments are administered with appropriate accommodations. 	2, 3, 6, 9

Table 10.4. Sources of Validity Evidence based on Internal Structure

Test Purpose	Summary of Evidence	Tech Report Sections
1. Measure and report Nebraska students' depth of achievement regarding the Nebraska College and Career Ready Standards.	<ul style="list-style-type: none"> • The assessment supports precise measurement and consistent classification. • Achievement levels were set consistent with best practice. 	6, 8, 9
2. Report if student achievement is sufficient academic proficiency to be on track for achieving college readiness.	<ul style="list-style-type: none"> • Scale is vertically articulated. • Achievement levels were vertically articulated. 	6, 7
3. Measure students' annual progress toward college and career readiness.	<ul style="list-style-type: none"> • The assessment supports precise measurement and consistent classification to support analysis and reporting of longitudinal data. • Scale is vertically articulated. • Achievement levels were vertically articulated. 	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"> • Range and Policy ALDs were developed in consultation with Nebraska educators with the goal of providing information to Nebraska educators. • Reporting categories align with the structure of the Nebraska standards to support the interpretation of the test results. • Items aligned with ALDs to support item writing processes. 	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"> • The assessment supports precise measurement and consistent classification for all students. • DIF analysis completed for all items across all required subgroups. 	6, 9

Table 10.5. Sources of Validity Evidence based on Other Variables

Test Purpose	Summary of Evidence	Tech Report Sections
1. Measure and report Nebraska students' depth of achievement regarding the Nebraska College and Career Ready Standards.	<ul style="list-style-type: none"> • Correlations with MAP Growth are high. 	8
2. Report if student achievement is sufficient academic proficiency to be on track for achieving college readiness.	<ul style="list-style-type: none"> • No evidence is provided. 	
3. Measure students' annual progress toward college and career readiness.	<ul style="list-style-type: none"> • No evidence is provided. 	
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"> • No evidence is provided. 	
5. Assess students' construct relevant achievement in ELA, mathematics, and science for all students and subgroups of students.	<ul style="list-style-type: none"> • No evidence is provided. 	

10.4. Interpretive Argument Claims

The test scores for the 2022 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.	6. Psychometric Analyses 9. Reliability	Simulations, analysis of test information, conditional standard errors of measurement, classification accuracy, and reliability estimates; blueprint

Arguments	Tech Report Section(s)	Evidence
		comparability across students; item analysis, calibration and linking 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
Achievement standards were rigorous and technically sound.	7. Standard Setting	Documentation of the mathematics standard setting procedures, ELA cut score review process, and the science standard setting procedures 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 2021–2022 NSCAS Growth 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 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 ELA and mathematics assessments demonstrated that the tests administered to students conform to the blueprint during the constraint-based engine simulation studies.

The item pool and item selection procedures used for the adaptive administration adequately support the test design and blueprint. 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.

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.
- 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>.
- 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.
- French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, 33, 315–332.
- Fu, J., & Monfils, L. (2016). *LDIF_ES: A SAS macro for logistic regression tests for differential item functioning of dichotomous and polytomous items*. (Research Memorandum ETS RM-16-17). Princeton, NJ: Educational Testing Service.
- 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. (2021). *Winsteps® Rasch measurement computer program (version 4.8.0.0) [Computer software]*. Portland, OR: winsteps.com.
- National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. Committee on a Conceptual Framework for New K–12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education Washington, DC: The National Academies Press.
- Nebraska Department of Education (NDE). (2019, August). *Nebraska Student-Centered Assessment System (NSCAS) summative & alternate accessibility manual*. Retrieved from <https://cdn.education.ne.gov/wp-content/uploads/2019/02/NSCAS-Summative-and-Alternate-Accessibility-Manual-2.8.19.pdf>.
- NWEA. (2020). Constraint-based engine scientific approach and methodology (confidential).

- NWEA. (2021a, October). Constraint-based engine simulation report for the winter 2021-2022 NSCAS ELA and Mathematics assessments (Tech. Rep.). Portland, OR.
- NWEA. (2022a, February). Constraint-based engine simulation report for the spring 2021-2022 NSCAS ELA and Mathematics assessments (Tech. Rep.). Portland, OR.
- NWEA. (2022b, May). Constraint-based engine evaluation report for the winter 2021-2022 NSCAS ELA and Mathematics assessments (Tech. Rep.). Portland, OR.
- NWEA. (2022c, June). Constraint-based engine evaluation report for the spring 2021-2022 NSCAS ELA and Mathematics assessments (Tech. Rep.). Portland, OR.
- 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), 59–81.
- Phillips, S., & Camara, W. J. (2006). Educational measurement. In (pp. 733–755).
- Plake, B. S., Huff, K., & Reshetar, R. (2010). Evidence-centered assessment design as a foundation for achievement-level descriptor development and for standard setting. *Applied Measurement in Education*, 23(4), 342–357.
- 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., & Johnson, R. L. (2018). *Creating and implementing student learning objectives to support student learning and teacher evaluation*. Under contract. Taylor and Francis.
- 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.
- 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>.
- Swaminathan, H., & Rogers, H. J., (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement*, 27, 361–370.
- Van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22(3), 259–270.
- Webb, N. L. (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.
- Zumbo, B. D. (1999). A Handbook on the Theory and Methods of Differential Item Functioning (DIF): *Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Directorate of Human Resources Research and Evaluation, Department of National Defense, Ottawa, ON.

Appendix A: Data Review Cheat Sheet



Measuring What Matters™

Data Review Cheat Sheet
NSCAS Data Review Meeting with NDE

Use this document as a guide when reviewing the NSCAS field test items. It includes flagging criteria for four different scenarios:

- General (both multiple-choice and non-multiple-choice items)
- Multiple-choice items
- Non-multiple-choice items (both 1- and 2-point items)
- Non-multiple-choice items (2-point items only)

References starting with “cia,” “fit,” or “dif” are how the statistics are identified in the data review file. The data review file also contains definitions above the statistics to clarify their meaning. A one-page summary of the statistical flags is located at the end of the document.

DIF			
Statistic	Flag	Meaning	Implication for Data Review
DIF of gender or ethnicity	C+ or C-	Item is flagged for potential bias toward a certain group of students.	Is there anything that could trigger the bias toward certain groups of students?

Page 1 of 5

Multiple-Choice Items			
Statistic	Flag	Meaning	Implication for Data Review
P-value Percent of students who got the item correct. (cia_Pval)	< 0.2 or > 0.9	Less than 20% of students got the item correct, or more than 90% of students got it correct.	Does it make sense that an item seems very difficult or very easy?
Option percentages (cia_Pct_Opt1-4)	Distractor % > P-value	More students chose a distractor than the key	Is the answer key accurate? Is the distractor appropriate (common error, etc.)?
Omit (cia_Pct_Omit)	> 5%	More than 5% of students are omitting this item.	Is there anything that could make this item confusing to students?
Item-total correlation aka Point Biserial (cia_ItemTotalCorr)	< 0.2	The item is not differentiating between high- and low-performing students.	Is the answer key accurate?
Item-total correlation for options (cia_ItemTotalCorr_Opt1-4)	> 0.05	An incorrect answer is pulling higher scoring students.	Is there anything that a distractor is doing for high-performing students to select it as an answer? Or is there a possibility for two correct answers? Is the distractor appropriate (common error, etc.)?
IRT Difficulty or Step parameters are extremely High	>=4.25	Probability of getting an item correct may require extremely high ability	Is the item too difficult for even high performing students to get it correct?
Do not use items if items have: <ul style="list-style-type: none"> Negative item-total correlation 			

Non-MC Items (Both 1-and 2-point items)			
Statistic	Flag	Meaning	Implication for Data Review
Low student count for each score (cia_Pct_Opt1-3)	= 0	No one got a certain score (e.g., no student got a score of 1).	Is there anything in the item that could cause students to not earn certain scores? Is the key correct?
Item-total correlation (cia_ItemTotalCorr)	< 0.2	The item is not differentiating between high- and low-performing students.	Are the keys accurate? If step parameters are flagged and item total correlation is flagged, the item may not be showing more sophisticated thinking in the content across score points. Is the item asking for the same skill more times?
Item-total correlation for score of 0 (cia_ItemTotalCorr_Opt1)	> 0.0	A score of 0 on the item is not differentiating achievement levels as expected.	Is there a reason earning 0 points is happening more often for high-performing students than low-performing?
Item-total correlation for score of 0 > Item-total correlation for score of 1	cia_ItemTotalCorr_Opt1 > cia_ItemTotalCorr_Opt2	A score of 0 on the item is better differentiating achievement levels than a score of 1.	Is there anything that could make the item perform the opposite of what is expected for high- vs. low-performing students who got a score of 0 vs. 1?
IRT Difficulty or Step parameters are extremely High	>=4.25	Probability of getting an item correct may require extremely high ability	Is the item too difficult for even high performing students to get it correct?
Step parameters [Step 1, Step2]	Step 1 > Step 2	Step parameters are not ordered in value (e.g., the difficulty of score 1 > the difficulty of score 2). There is not a good separation of students into different stages of learning.	Do students have to show more substantive knowledge to earn the second point? Is the same skill being repeated causing the difficulty to stay the same across steps 1 and 2? Is there another reason the difficulty is not increasing across points?
Do not use items if items have: <ul style="list-style-type: none"> Negative item-total correlation 			

Non-MC Items (2-point items only)			
Statistic	Flag	Meaning	Implication for Data Review
Item-total correlation for score of 1 > Item-total correlation for score of 2	$\text{cia_ItemTotalCorr_Opt2} > \text{cia_ItemTotalCorr_Opt3}$	A score of 1 on the item is better at differentiating achievement levels than a score of 2.	Is there anything that could make the item perform the opposite of what is expected for high- vs. low-performing students who got a score of 1 vs. 2?
Item-total correlation for score of 2 (cia_ItemTotalCorr_Opt3)	< 0.2	A score of 2 on the item is not differentiating achievement levels as expected.	Is there a reason earning 2 points is happening more often for low-performing students than high-performing?
IRT Difficulty or Step parameters are extremely High	≥ 4.25	Probability of getting an item correct may require extremely high ability	Is the item too difficult for even high performing students to get it correct?
Step parameters [Step 1, Step2]	Step 1 > Step 2	Step parameters are not ordered in value (e.g., the difficulty of score 1 > the difficulty of score 2). There is not a good separation of students into different stages of learning.	Do students have to show more substantive knowledge to earn the second point? Is the same skill being repeated causing the difficulty to stay the same across steps 1 and 2? Is there another reason the difficulty is not increasing across points?
Do not use 2-point items if items have: <ul style="list-style-type: none"> Negative item-total correlation No second-step parameters. 			

	Label	Statistics	Flags
MC items	Pvalue_LOW/ Pvalue_HIGH	<i>P</i> -value	< 0.2 or > 0.9
	Pvalue_Dis	Option percentages	Distractor % > <i>P</i> -value
	Pbis_LOW	Item-total correlation	< 0.20
	Pbis_Dis	Item-total correlation for distractors	> 0.05
Non-MC items (Both 1- and 2-point items)	Pvalue_LOW/ Pvalue_HIGH	<i>P</i> -value	< 0.2 or > 0.9
	N_012	Low student count for each score	= 0
	Pbis_LOW	Item-total correlation	< 0.2
	Score_0_Pbis	Item-total correlation for score of 0	> 0.0
	Score_0Vs1_Pbis	Item-total correlation for score of 0 > item-total correlation for score of 1	
Non-MC items (2-point items only)	Score_1Vs2_Pbis	Item-total correlation for score of 1 > item-total correlation for score of 2	
	Score_2_Pbis	Item-total correlation for score of 2	< 0.2
Item Parameters	itemFlag_IRT_Parameter	IRT Difficulty or Step parameters are extreme	>=4.25
	itemFlag_IRT_ReversedStep	Reversed Step parameters	Step 1 > Step 2
DIF	itemFlag_Gender_DIF/ itemFlag_Black_DIF/ itemFlag_Hispanic_DIF	DIF of gender or ethnicity	C+ or C-
Do not use items if items have: <ul style="list-style-type: none"> Negative item-total correlation No second step parameters 			

Appendix B: Summary P-values by Item Type

Table B.1. Summary P-values by Item Type—Operational Items

Operational Items																
Grade	Item Type	#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
ELA																
3	Choice Multiple	25	0.544	0.154	0.207	0.866	0	0	1	3	4	8	7	0	2	0
	Choice Single	481	0.505	0.158	0.000	1.000	4	1	19	105	134	102	68	23	14	11
	Composite	26	0.429	0.128	0.128	0.719	0	2	1	5	12	5	0	1	0	0
	Gap Match Multiple	27	0.463	0.231	0.000	1.000	1	1	4	5	4	6	3	0	2	1
	Gap Match Single	4	0.483	0.026	0.461	0.517	0	0	0	0	3	1	0	0	0	0
	Hot Text	1	0.138	–	0.138	0.138	0	1	0	0	0	0	0	0	0	0
4	Choice Multiple	33	0.582	0.175	0.218	1.000	0	0	1	4	5	7	10	4	0	2
	Choice Single	551	0.538	0.162	0.000	1.000	7	3	17	68	121	159	100	44	18	14
	Composite	28	0.497	0.136	0.000	0.757	1	0	1	0	10	11	4	1	0	0
	Gap Match Multiple	33	0.461	0.117	0.140	0.679	0	1	1	7	9	11	4	0	0	0
	Gap Match Single	1	0.580	–	0.580	0.580	0	0	0	0	0	1	0	0	0	0
	Hot Text	1	0.616	–	0.616	0.616	0	0	0	0	0	0	1	0	0	0
5	Choice Multiple	36	0.501	0.142	0.257	0.822	0	0	3	5	13	8	3	3	1	0
	Choice Single	473	0.520	0.166	0.000	1.000	8	6	19	66	119	114	82	40	11	8
	Composite	26	0.490	0.106	0.200	0.726	0	1	1	1	10	12	0	1	0	0
	Gap Match Multiple	25	0.483	0.176	0.122	0.746	0	4	0	2	4	9	4	2	0	0
	Gap Match Single	1	0.570	–	0.570	0.570	0	0	0	0	0	1	0	0	0	0
	Hot Text	1	0.616	–	0.616	0.616	0	0	0	0	0	0	1	0	0	0
6	Choice Multiple	45	0.450	0.161	0.000	0.750	3	0	2	5	18	12	2	3	0	0
	Choice Single	444	0.508	0.184	0.000	1.000	11	3	32	71	110	87	60	47	14	9
	Composite	25	0.503	0.132	0.193	0.714	0	1	1	1	10	6	5	1	0	0
	Gap Match Multiple	24	0.459	0.163	0.217	1.000	0	0	4	4	6	9	0	0	0	1
	Gap Match Single	2	0.658	0.188	0.526	0.791	0	0	0	0	0	1	0	1	0	0
	Hot Text	1	0.616	–	0.616	0.616	0	0	0	0	0	0	1	0	0	0

Operational Items																
Grade	Item Type	#Items	Mean	SD	Min.	Max.	#Items by <i>P</i> -value Range									
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	> 0.9
	Hot Text	2	0.354	0.290	0.149	0.559	0	1	0	0	0	1	0	0	0	0
7	Choice Multiple	32	0.428	0.167	0.000	1.000	1	0	5	6	16	1	2	0	0	1
	Choice Single	459	0.532	0.182	0.000	1.000	13	0	18	55	117	110	77	41	11	17
	Composite	25	0.419	0.124	0.177	0.750	0	2	1	8	8	5	0	1	0	0
	Gap Match Multiple	16	0.487	0.218	0.000	1.000	1	0	1	4	3	4	1	1	0	1
	Gap Match Single	6	0.353	0.140	0.140	0.555	0	1	0	3	1	1	0	0	0	0
	Hot Text	1	0.540	–	0.540	0.540	0	0	0	0	0	1	0	0	0	0

Operational Items																
Grade	Item Type	#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
8	Choice Multiple	43	0.409	0.135	0.056	0.750	1	2	5	11	16	5	2	1	0	0
	Choice Single	480	0.521	0.175	0.000	1.000	3	4	25	82	130	108	61	32	14	21
	Composite	16	0.386	0.110	0.140	0.554	0	1	2	7	4	2	0	0	0	0
	Gap Match Multiple	13	0.536	0.214	0.125	1.000	0	1	0	2	2	4	1	2	0	1
	Gap Match Single	1	0.518	–	0.518	0.518	0	0	0	0	0	1	0	0	0	0
Mathematics																
3	Choice Multiple	31	0.524	0.114	0.326	0.770	0	0	0	6	6	15	1	3	0	0
	Choice Single	500	0.541	0.104	0.300	0.886	0	0	1	39	128	210	83	31	8	0
	Composite	48	0.514	0.149	0.152	0.823	0	1	3	7	9	14	9	4	1	0
	Gap Match Multiple	39	0.529	0.139	0.000	0.773	1	1	0	3	7	17	9	1	0	0
	Graphic Gap Match	37	0.516	0.120	0.108	0.756	0	1	1	3	8	17	6	1	0	0
	Hot Text	9	0.574	0.087	0.428	0.691	0	0	0	0	2	4	3	0	0	0
	Schema Discrete	1	0.526	–	0.526	0.526	0	0	0	0	0	1	0	0	0	0
	Text Entry	40	0.542	0.075	0.356	0.762	0	0	0	1	8	27	2	2	0	0
4	Choice Multiple	39	0.485	0.102	0.240	0.700	0	0	1	8	13	13	4	0	0	0
	Choice Single	295	0.517	0.089	0.179	0.816	0	1	1	20	108	123	32	9	1	0
	Composite	49	0.490	0.153	0.146	1.000	0	1	3	8	15	14	4	1	2	1
	Gap Match Multiple	29	0.496	0.127	0.267	0.893	0	0	1	6	12	4	5	0	1	0
	Graphic Gap Match	32	0.534	0.067	0.372	0.657	0	0	0	1	8	18	5	0	0	0
	Hot Text	16	0.487	0.111	0.238	0.712	0	0	1	1	8	3	2	1	0	0
	Schema Discrete	2	0.534	0.031	0.512	0.555	0	0	0	0	0	2	0	0	0	0
	Text Entry	49	0.523	0.087	0.359	0.730	0	0	0	6	12	24	5	2	0	0
5	Choice Multiple	39	0.478	0.117	0.253	0.875	0	0	2	7	14	12	3	0	1	0
	Choice Single	346	0.542	0.107	0.244	0.833	0	0	6	20	98	127	68	22	5	0
	Composite	66	0.513	0.142	0.193	0.795	0	1	4	8	19	19	8	7	0	0
	Gap Match Multiple	32	0.500	0.088	0.382	0.839	0	0	0	2	14	14	1	0	1	0
	Graphic Gap Match	18	0.566	0.076	0.385	0.721	0	0	0	1	2	9	5	1	0	0
	Hot Text	11	0.572	0.067	0.476	0.689	0	0	0	0	1	7	3	0	0	0

Operational Items																
Grade	Item Type	#Items	Mean	SD	Min.	Max.	#Items by <i>P</i> -value Range									
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	> 0.9
	Schema Discrete	2	0.558	0.043	0.527	0.588	0	0	0	0	0	2	0	0	0	0
	Text Entry	42	0.517	0.101	0.300	0.768	0	0	0	5	12	19	3	3	0	0

Operational Items																
Grade	Item Type	#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
6	Choice Multiple	46	0.442	0.130	0.163	1.000	0	1	5	9	23	4	3	0	0	1
	Choice Single	510	0.516	0.112	0.000	0.857	5	0	3	43	185	179	68	22	5	0
	Composite	51	0.469	0.231	0.000	1.000	3	2	5	11	11	8	2	4	2	3
	Gap Match Multiple	31	0.496	0.107	0.203	0.730	0	0	1	2	15	9	3	1	0	0
	Graphic Gap Match	13	0.525	0.079	0.411	0.632	0	0	0	0	5	5	3	0	0	0
	Hot Text	20	0.490	0.129	0.212	0.819	0	0	1	3	9	3	3	0	1	0
	Text Entry	29	0.520	0.105	0.363	0.770	0	0	0	2	14	7	4	2	0	0
7	Choice Multiple	28	0.451	0.142	0.206	0.837	0	0	3	7	11	4	1	1	1	0
	Choice Single	432	0.482	0.101	0.208	0.857	0	0	11	79	170	120	41	10	1	0
	Composite	39	0.404	0.127	0.152	0.750	0	4	3	9	18	3	1	1	0	0
	Gap Match Multiple	24	0.416	0.058	0.331	0.517	0	0	0	9	13	2	0	0	0	0
	Graphic Gap Match	9	0.361	0.081	0.239	0.475	0	0	2	4	3	0	0	0	0	0
	Hot Text	22	0.401	0.109	0.151	0.589	0	1	3	8	6	4	0	0	0	0
	Text Entry	54	0.437	0.085	0.287	0.641	0	0	3	17	21	11	2	0	0	0
8	Choice Multiple	29	0.437	0.101	0.307	0.691	0	0	0	12	9	7	1	0	0	0
	Choice Single	313	0.488	0.095	0.262	0.826	0	0	4	43	140	94	22	9	1	0
	Composite	46	0.426	0.178	0.145	0.813	0	5	8	9	8	8	4	3	1	0
	Gap Match Multiple	40	0.415	0.087	0.225	0.600	0	0	4	10	19	6	1	0	0	0
	Graphic Gap Match	11	0.468	0.098	0.336	0.702	0	0	0	2	6	2	0	1	0	0
	Hot Text	36	0.434	0.113	0.000	0.646	1	0	1	12	12	7	3	0	0	0
	Text Entry	47	0.459	0.062	0.300	0.611	0	0	0	7	26	13	1	0	0	0
Science																
5	Choice Multiple	2	0.581	0.086	0.520	0.642	0	0	0	0	0	1	1	0	0	0
	Choice Single	10	0.639	0.114	0.482	0.844	0	0	0	0	1	3	2	3	1	0
	Composite	3	0.600	0.085	0.513	0.684	0	0	0	0	0	1	2	0	0	0
	Gap Match Multiple	1	0.746	–	0.746	0.746	0	0	0	0	0	0	0	1	0	0
	Graphic Gap Match	3	0.531	0.059	0.492	0.598	0	0	0	0	1	2	0	0	0	0
	Hot Text	2	0.637	0.229	0.475	0.798	0	0	0	0	1	0	0	1	0	0

Operational Items																
Grade	Item Type	#Items	Mean	SD	Min.	Max.	#Items by <i>P</i> -value Range									
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	> 0.9
8	Choice Multiple	3	0.392	0.101	0.279	0.473	0	0	1	0	2	0	0	0	0	0
	Choice Single	8	0.561	0.167	0.293	0.795	0	0	1	0	2	2	1	2	0	0
	Composite	9	0.259	0.102	0.139	0.457	0	3	2	3	1	0	0	0	0	0
	Gap Match Multiple	3	0.293	0.065	0.231	0.360	0	0	2	1	0	0	0	0	0	0
	Graphic Gap Match	2	0.590	0.155	0.480	0.699	0	0	0	0	1	0	1	0	0	0
	Hot Text	2	0.438	0.087	0.377	0.499	0	0	0	1	1	0	0	0	0	0

Table B.2. Summary *P*-values by Item Type—Field Test Items

Field Test Items																
Grade	Item Type	#Items	Mean	SD	Min.	Max.	#Items by <i>P</i> -value Range									
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	> 0.9
ELA																
3	Choice Multiple	37	0.519	0.105	0.308	0.755	0	0	0	4	14	11	6	2	0	0
	Choice Single	173	0.489	0.161	0.175	0.889	0	2	20	36	39	33	27	11	5	0
	Composite	31	0.354	0.131	0.141	0.654	0	3	9	7	7	4	1	0	0	0
	Gap Match Multiple	13	0.439	0.159	0.146	0.755	0	1	1	2	5	2	1	1	0	0
	Hot Text	3	0.557	0.084	0.499	0.653	0	0	0	0	1	1	1	0	0	0
4	Choice Multiple	43	0.555	0.102	0.387	0.804	0	0	0	1	12	16	11	2	1	0
	Choice-Single	141	0.542	0.169	0.169	0.899	0	2	9	21	28	29	25	16	11	0
	Composite	21	0.441	0.124	0.199	0.686	0	1	2	4	6	7	1	0	0	0
	Gap Match Multiple	12	0.525	0.159	0.180	0.808	0	1	0	0	4	5	0	1	1	0
	Hot Text	1	0.529	–	0.529	0.529	0	0	0	0	0	1	0	0	0	0
5	Choice Multiple	49	0.571	0.097	0.367	0.779	0	0	0	1	12	18	11	7	0	0
	Choice Single	137	0.501	0.153	0.187	0.854	0	1	12	27	32	25	25	12	3	0
	Composite	22	0.333	0.138	0.137	0.752	0	4	6	5	5	1	0	1	0	0
	Gap Match Multiple	20	0.559	0.207	0.216	0.913	0	0	3	1	5	2	4	2	2	1
	Gap Match Single	2	0.542	0.052	0.505	0.579	0	0	0	0	0	2	0	0	0	0
	Hot Text	10	0.506	0.138	0.292	0.793	0	0	1	1	4	2	1	1	0	0
6	Choice Multiple	34	0.490	0.118	0.215	0.686	0	0	3	3	11	10	7	0	0	0
	Choice Single	145	0.505	0.161	0.083	0.891	1	2	10	29	33	26	27	12	5	0
	Composite	36	0.363	0.167	0.124	0.782	0	8	4	11	8	2	1	2	0	0
	Gap Match Multiple	18	0.565	0.181	0.206	0.883	0	0	1	3	2	4	4	2	2	0
	Hot Text	5	0.520	0.147	0.320	0.684	0	0	0	1	1	1	2	0	0	0
7	Choice Multiple	38	0.506	0.103	0.252	0.796	0	0	2	3	12	15	5	1	0	0
	Choice Single	120	0.555	0.146	0.253	0.877	0	0	4	16	21	34	24	13	8	0
	Composite	39	0.464	0.122	0.226	0.719	0	0	2	11	12	7	6	1	0	0
	Gap Match Multiple	21	0.455	0.199	0.100	0.867	0	1	5	3	3	5	2	0	2	0
	Hot Text	12	0.593	0.121	0.387	0.815	0	0	0	1	2	4	3	1	1	0

Field Test Items																
Grade	Item Type	#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
8	Choice Multiple	36	0.496	0.125	0.255	0.734	0	0	3	5	13	7	6	2	0	0
	Choice Single	111	0.578	0.153	0.226	0.871	0	0	3	10	23	24	25	16	10	0
	Composite	41	0.463	0.131	0.167	0.750	0	1	3	10	11	10	5	1	0	0
	Gap Match Multiple	22	0.512	0.137	0.206	0.725	0	0	1	3	7	4	5	2	0	0
	Gap Match Single	2	0.339	0.143	0.238	0.440	0	0	1	0	1	0	0	0	0	0
	Hot Text	13	0.695	0.126	0.505	0.892	0	0	0	0	0	2	6	2	3	0
Mathematics																
3	Choice Multiple	26	0.417	0.182	0.120	0.780	0	4	4	4	3	8	1	2	0	0
	Choice Single	38	0.612	0.216	0.213	0.927	0	0	4	4	5	5	5	6	7	2
	Composite	9	0.429	0.199	0.208	0.812	0	0	4	0	1	3	0	0	1	0
	Gap Match Multiple	17	0.458	0.245	0.047	0.743	2	2	2	0	1	4	2	4	0	0
	Graphic Gap Match	16	0.516	0.252	0.066	0.970	1	2	1	0	2	4	2	2	1	1
	Text Entry	6	0.621	0.214	0.350	0.904	0	0	0	1	1	1	0	2	0	1
4	Choice Multiple	9	0.436	0.163	0.179	0.663	0	1	1	2	2	1	2	0	0	0
	Choice Single	14	0.529	0.193	0.217	0.789	0	0	3	1	1	2	4	3	0	0
	Composite	5	0.482	0.133	0.349	0.634	0	0	0	2	1	0	2	0	0	0
	Gap Match Multiple	7	0.324	0.139	0.230	0.615	0	0	5	1	0	0	1	0	0	0
	Graphic Gap Match	11	0.435	0.257	0.110	0.803	0	3	2	0	1	2	1	1	1	0
	Hot Text	1	0.689	–	0.689	0.689	0	0	0	0	0	0	1	0	0	0
5	Text Entry	7	0.617	0.206	0.352	0.916	0	0	0	1	1	2	1	0	1	1
	Choice Multiple	16	0.395	0.169	0.089	0.650	1	1	3	3	4	1	3	0	0	0
	Choice Single	19	0.557	0.151	0.232	0.768	0	0	1	2	4	4	5	3	0	0
	Composite	11	0.474	0.111	0.249	0.612	0	0	1	2	3	3	2	0	0	0
	Gap Match Multiple	11	0.424	0.281	0.091	0.880	1	1	3	2	0	0	2	0	2	0
	Graphic Gap Match	11	0.321	0.177	0.019	0.526	1	3	1	0	4	2	0	0	0	0
Hot Text	1	0.225	–	0.225	0.225	0	0	1	0	0	0	0	0	0	0	
Text Entry	9	0.511	0.196	0.220	0.882	0	0	1	1	4	1	0	1	1	0	

Field Test Items																
Grade	Item Type	#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
6	Choice Multiple	36	0.237	0.150	0.026	0.761	5	8	16	5	0	0	1	1	0	0
	Choice Single	107	0.539	0.197	0.078	0.908	1	4	3	23	19	16	10	20	9	2
	Composite	25	0.389	0.121	0.219	0.637	0	0	5	12	3	2	3	0	0	0
	Gap Match Multiple	16	0.371	0.231	0.019	0.945	1	2	2	8	0	0	2	0	0	1
	Graphic Gap Match	21	0.282	0.138	0.003	0.560	3	2	5	9	0	2	0	0	0	0
	Hot Text	6	0.298	0.111	0.147	0.459	0	1	1	3	1	0	0	0	0	0
	Schema Discrete	1	0.699	–	0.699	0.699	0	0	0	0	0	0	1	0	0	0
	Text Entry	41	0.341	0.189	0.016	0.775	2	8	13	3	5	6	2	2	0	0
7	Choice Multiple	26	0.233	0.155	0.043	0.607	9	3	5	7	0	1	1	0	0	0
	Choice Single	29	0.466	0.140	0.209	0.758	0	0	4	4	10	5	5	1	0	0
	Composite	21	0.343	0.169	0.047	0.705	1	4	3	5	4	3	0	1	0	0
	Gap Match Multiple	16	0.293	0.203	0.048	0.760	4	3	0	4	4	0	0	1	0	0
	Graphic Gap Match	1	0.419	–	0.419	0.419	0	0	0	0	1	0	0	0	0	0
	Hot Text	1	0.191	–	0.191	0.191	0	1	0	0	0	0	0	0	0	0
	Text Entry	27	0.257	0.201	0.034	0.764	7	7	5	2	3	0	2	1	0	0
8	Choice Multiple	26	0.200	0.157	0.039	0.490	9	8	2	1	6	0	0	0	0	0
	Choice Single	17	0.460	0.146	0.237	0.678	0	0	4	3	1	5	4	0	0	0
	Composite	15	0.305	0.138	0.028	0.549	1	3	3	4	3	1	0	0	0	0
	Gap Match Multiple	12	0.267	0.208	0.015	0.650	4	1	1	4	0	1	1	0	0	0
	Graphic Gap Match	2	0.176	0.159	0.064	0.289	1	0	1	0	0	0	0	0	0	0
	Hot Text	1	0.226	–	0.226	0.226	0	0	1	0	0	0	0	0	0	0
	Schema Discrete	1	0.323	–	0.323	0.323	0	0	0	1	0	0	0	0	0	0
	Text Entry	15	0.211	0.138	0.028	0.422	5	2	4	1	3	0	0	0	0	0
Science																
5	Choice Multiple	10	0.551	0.249	0.154	0.868	0	1	1	1	1	1	2	0	3	0
	Choice Single	47	0.575	0.168	0.184	0.861	0	1	1	5	10	8	10	6	6	0
	Composite	7	0.453	0.249	0.094	0.769	1	0	1	1	1	0	2	1	0	0
	Gap Match Multiple	16	0.551	0.237	0.256	0.898	0	0	3	3	1	2	2	2	3	0

Field Test Items																
Grade	Item Type	#Items	Mean	SD	Min.	Max.	#Items by <i>P</i> -value Range									
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	> 0.9
	Graphic Gap Match	9	0.683	0.253	0.115	0.911	0	1	0	0	0	2	0	2	2	2
	Hot Text	1	0.190	–	0.190	0.190	0	1	0	0	0	0	0	0	0	0
8	Choice Multiple	15	0.228	0.122	0.051	0.458	4	2	5	3	1	0	0	0	0	0
	Choice Single	44	0.534	0.149	0.173	0.789	0	1	2	5	8	14	7	7	0	0
	Composite	10	0.482	0.246	0.093	0.799	1	1	1	0	2	2	0	3	0	0
	Gap Match Multiple	18	0.391	0.245	0.079	0.858	2	3	2	2	4	1	1	2	1	0
	Graphic Gap Match	9	0.466	0.278	0.033	0.797	1	1	1	0	2	0	2	2	0	0

Appendix C: Summary Item-Total Correlations by Item Type

Table C.1. Summary Item-Total Correlations by Item Type—Operational Items

Operational Items													
Grade	Item Type	#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
ELA													
3	Choice Multiple	25	0.330	0.184	-0.275	0.595	3	1	5	7	5	4	0
	Choice Single	481	0.293	0.159	-0.559	1.000	36	61	146	144	63	21	10
	Composite	26	0.358	0.159	0.060	0.619	1	5	2	7	6	4	1
	Gap Match Multiple	27	0.285	0.150	-0.079	0.518	4	2	6	10	4	1	0
	Gap Match Single	4	0.279	0.052	0.228	0.349	0	0	3	1	0	0	0
	Hot Text	1	0.258	–	0.258	0.258	0	0	1	0	0	0	0
4	Choice Multiple	33	0.308	0.186	-0.282	0.560	5	4	2	12	6	4	0
	Choice Single	551	0.306	0.162	-1.000	1.000	63	30	107	218	104	25	4
	Composite	28	0.295	0.180	-0.041	0.525	5	4	4	2	10	3	0
	Gap Match Multiple	33	0.262	0.205	-0.251	0.965	8	2	6	13	3	0	1
	Gap Match Single	1	0.364	–	0.364	0.364	0	0	0	1	0	0	0
	Hot Text	1	0.425	–	0.425	0.425	0	0	0	0	1	0	0
5	Choice Multiple	36	0.285	0.264	-0.452	0.967	8	3	5	5	10	4	1
	Choice Single	473	0.289	0.178	-0.974	0.920	53	40	118	154	88	16	4
	Composite	26	0.239	0.280	-0.547	0.511	6	3	4	3	8	2	0
	Gap Match Multiple	25	0.220	0.240	-0.413	0.539	5	2	8	5	4	1	0
	Gap Match Single	1	0.356	–	0.356	0.356	0	0	0	1	0	0	0
6	Choice Multiple	45	0.288	0.290	-0.486	1.000	11	8	7	9	5	0	5
	Choice Single	444	0.267	0.202	-1.000	0.964	76	40	105	132	67	14	10
	Composite	25	0.354	0.235	-0.157	1.000	3	2	4	7	3	4	2
	Gap Match Multiple	24	0.322	0.214	-0.082	1.000	3	3	4	8	3	2	1
	Gap Match Single	2	0.382	0.114	0.301	0.463	0	0	0	1	1	0	0
	Hot Text	2	0.339	0.152	0.231	0.446	0	0	1	0	1	0	0
7	Choice Multiple	32	0.233	0.284	-0.627	1.000	9	4	1	11	6	0	1

Appendix C: Summary Item-Total Correlations by Item Type

Operational Items													
Grade	Item Type	#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
	Choice Single	459	0.280	0.207	-1.000	1.000	70	43	98	132	92	17	7
	Composite	25	0.306	0.257	-0.220	1.000	4	3	4	6	4	3	1
	Gap Match Multiple	16	0.271	0.144	0.000	0.476	2	3	4	5	2	0	0
	Gap Match Single	6	0.176	0.083	0.081	0.326	1	4	0	1	0	0	0
	Hot Text	1	0.173	–	0.173	0.173	0	1	0	0	0	0	0

Appendix C: Summary Item-Total Correlations by Item Type

Operational Items													
Grade	Item Type	#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
8	Choice Multiple	43	0.224	0.322	-1.000	1.000	10	4	7	12	9	0	1
	Choice Single	480	0.274	0.193	-1.000	1.000	72	42	106	160	87	9	4
	Composite	16	0.306	0.152	0.035	0.518	3	1	2	5	4	1	0
	Gap Match Multiple	13	0.260	0.156	0.000	0.489	2	2	3	3	3	0	0
	Gap Match Single	1	0.261	–	0.261	0.261	0	0	1	0	0	0	0
Mathematics													
3	Choice Multiple	31	0.386	0.095	0.089	0.542	1	0	4	10	12	4	0
	Choice Single	500	0.363	0.083	0.026	0.754	2	13	91	236	136	21	1
	Composite	48	0.428	0.064	0.198	0.526	0	1	1	10	31	5	0
	Gap Match Multiple	39	0.394	0.096	0.000	0.531	1	0	4	11	18	5	0
	Graphic Gap Match	37	0.369	0.067	0.246	0.517	0	0	7	19	10	1	0
	Hot Text	9	0.405	0.084	0.272	0.509	0	0	1	4	3	1	0
	Schema Discrete	1	0.525	–	0.525	0.525	0	0	0	0	0	1	0
	Text Entry	40	0.385	0.075	0.211	0.522	0	0	5	17	15	3	0
4	Choice Multiple	39	0.348	0.094	0.120	0.498	0	3	7	18	11	0	0
	Choice Single	295	0.349	0.084	0.125	0.587	0	17	54	144	73	7	0
	Composite	49	0.400	0.119	0.000	0.591	2	0	6	12	24	5	0
	Gap Match Multiple	29	0.377	0.122	0.000	0.804	1	0	3	14	10	0	1
	Graphic Gap Match	32	0.398	0.094	0.192	0.570	0	1	4	10	13	4	0
	Hot Text	16	0.390	0.071	0.253	0.493	0	0	2	6	8	0	0
	Schema Discrete	2	0.469	0.063	0.424	0.514	0	0	0	0	1	1	0
	Text Entry	49	0.389	0.080	0.174	0.548	0	1	6	17	22	3	0
5	Choice Multiple	39	0.364	0.128	0.049	0.979	1	1	4	21	11	0	1
	Choice Single	346	0.348	0.087	0.094	0.578	1	11	89	142	89	14	0
	Composite	66	0.386	0.096	0.000	0.592	2	1	6	22	31	4	0
	Gap Match Multiple	32	0.388	0.081	0.237	0.520	0	0	6	9	16	1	0
	Graphic Gap Match	18	0.412	0.084	0.236	0.572	0	0	1	7	8	2	0
	Hot Text	11	0.367	0.076	0.174	0.441	0	1	0	7	3	0	0

Appendix C: Summary Item-Total Correlations by Item Type

Operational Items													
Grade	Item Type	#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
	Schema Discrete	2	0.384	0.078	0.328	0.439	0	0	0	1	1	0	0
	Text Entry	42	0.397	0.077	0.170	0.552	0	1	3	20	14	4	0

Appendix C: Summary Item-Total Correlations by Item Type

Operational Items													
Grade	Item Type	#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
6	Choice Multiple	46	0.381	0.106	0.000	0.525	2	0	5	17	20	2	0
	Choice Single	510	0.348	0.087	-0.073	0.561	8	11	109	255	109	18	0
	Composite	51	0.390	0.178	-0.059	0.852	6	1	2	9	25	6	2
	Gap Match Multiple	31	0.356	0.073	0.259	0.543	0	0	8	14	7	2	0
	Graphic Gap Match	13	0.381	0.062	0.258	0.487	0	0	2	7	4	0	0
	Hot Text	20	0.368	0.101	0.183	0.488	0	1	4	6	9	0	0
	Text Entry	29	0.378	0.070	0.248	0.532	0	0	5	11	12	1	0
7	Choice Multiple	28	0.422	0.076	0.300	0.636	0	0	0	13	11	3	1
	Choice Single	432	0.344	0.084	0.063	0.567	1	23	100	192	104	12	0
	Composite	39	0.402	0.194	-0.468	0.774	3	0	1	11	16	6	2
	Gap Match Multiple	24	0.413	0.098	0.185	0.527	0	1	3	6	7	7	0
	Graphic Gap Match	9	0.374	0.096	0.280	0.510	0	0	2	4	2	1	0
	Hot Text	22	0.365	0.087	0.159	0.513	0	2	2	12	5	1	0
	Text Entry	54	0.415	0.066	0.286	0.579	0	0	2	20	27	5	0
8	Choice Multiple	29	0.374	0.073	0.194	0.503	0	1	4	11	12	1	0
	Choice Single	313	0.341	0.073	0.096	0.553	1	7	83	163	54	5	0
	Composite	46	0.411	0.144	0.111	1.000	0	3	7	10	19	5	2
	Gap Match Multiple	40	0.368	0.091	0.095	0.512	1	2	4	19	13	1	0
	Graphic Gap Match	11	0.417	0.049	0.335	0.489	0	0	0	4	7	0	0
	Hot Text	36	0.345	0.099	0.000	0.481	1	1	6	13	15	0	0
	Text Entry	47	0.422	0.058	0.318	0.566	0	0	0	20	22	5	0
Science													
5	Choice Multiple	2	0.459	0.132	0.366	0.552	0	0	0	1	0	1	0
	Choice Single	10	0.434	0.088	0.254	0.538	0	0	1	1	6	2	0
	Composite	3	0.572	0.046	0.522	0.614	0	0	0	0	0	2	1
	Gap Match Multiple	1	0.457	–	0.457	0.457	0	0	0	0	1	0	0
	Graphic Gap Match	3	0.536	0.044	0.496	0.582	0	0	0	0	1	2	0
	Hot Text	2	0.436	0.106	0.361	0.512	0	0	0	1	0	1	0

Appendix C: Summary Item-Total Correlations by Item Type

Operational Items													
Grade	Item Type	#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
8	Choice Multiple	3	0.391	0.043	0.356	0.439	0	0	0	2	1	0	0
	Choice Single	8	0.385	0.127	0.211	0.544	0	0	3	0	3	2	0
	Composite	9	0.420	0.071	0.336	0.529	0	0	0	4	3	2	0
	Gap Match Multiple	3	0.539	0.093	0.443	0.628	0	0	0	0	1	1	1
	Graphic Gap Match	2	0.611	0.058	0.570	0.652	0	0	0	0	0	1	1
	Hot Text	2	0.488	0.038	0.461	0.516	0	0	0	0	1	1	0

Table C.2. Summary P-Values by Item Type—Field Test Items

Field Test Items													
Grade	Item Type	#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
ELA													
3	Choice Multiple	37	0.404	0.129	0.004	0.590	1	2	3	11	10	10	0
	Choice Single	173	0.305	0.145	-0.180	0.559	17	22	42	41	39	12	0
	Composite	31	0.339	0.157	0.031	0.599	4	2	7	5	8	5	0
	Gap Match Multiple	13	0.462	0.092	0.305	0.610	0	0	0	3	5	4	1
	Hot Text	3	0.349	0.063	0.282	0.407	0	0	1	1	1	0	0
4	Choice Multiple	43	0.400	0.103	0.178	0.572	0	3	4	12	16	8	0
	Choice Single	141	0.313	0.135	-0.083	0.553	12	15	35	36	35	8	0
	Composite	21	0.372	0.140	0.021	0.524	1	2	1	6	7	4	0
	Gap Match Multiple	12	0.406	0.132	0.051	0.552	1	0	0	3	7	1	0
	Hot Text	1	0.357	–	0.357	0.357	0	0	0	1	0	0	0
5	Choice Multiple	49	0.385	0.091	0.093	0.552	1	1	5	22	16	4	0
	Choice Single	137	0.279	0.126	-0.196	0.511	13	23	36	39	24	2	0
	Composite	22	0.246	0.145	-0.055	0.523	3	5	6	5	2	1	0
	Gap Match Multiple	20	0.421	0.130	0.202	0.565	0	0	5	2	5	8	0
	Gap Match Single	2	0.368	0.088	0.306	0.430	0	0	0	1	1	0	0
	Hot Text	10	0.261	0.118	0.129	0.457	0	4	3	1	2	0	0
6	Choice Multiple	34	0.393	0.130	0.064	0.596	1	2	4	10	9	8	0
	Choice Single	145	0.291	0.133	-0.234	0.530	14	16	40	40	31	4	0
	Composite	36	0.329	0.139	-0.065	0.620	2	4	8	9	11	1	1
	Gap Match Multiple	18	0.448	0.117	0.170	0.690	0	1	0	4	8	4	1
	Hot Text	5	0.344	0.122	0.152	0.489	0	1	0	3	1	0	0
7	Choice Multiple	38	0.377	0.123	0.024	0.563	1	3	6	9	15	4	0
	Choice Single	120	0.336	0.120	0.023	0.553	4	11	33	31	34	7	0
	Composite	39	0.375	0.124	0.029	0.577	1	2	8	11	10	7	0
	Gap Match Multiple	21	0.427	0.077	0.309	0.572	0	0	0	9	8	4	0
	Hot Text	12	0.330	0.152	-0.097	0.505	1	0	1	8	1	1	0

Appendix C: Summary Item-Total Correlations by Item Type

Field Test Items													
Grade	Item Type	#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
8	Choice Multiple	36	0.344	0.116	0.102	0.617	0	3	11	11	8	2	1
	Choice Single	111	0.321	0.124	-0.001	0.572	6	17	22	31	30	5	0
	Composite	41	0.360	0.110	0.056	0.552	1	3	6	13	15	3	0
	Gap Match Multiple	22	0.415	0.094	0.191	0.596	0	1	1	8	8	4	0
	Gap Match Single	2	0.253	0.065	0.207	0.300	0	0	2	0	0	0	0
	Hot Text	13	0.334	0.089	0.197	0.496	0	1	4	6	2	0	0
Mathematics													
3	Choice Multiple	26	0.442	0.120	0.092	0.617	1	1	1	5	10	7	1
	Choice Single	38	0.310	0.095	0.062	0.455	1	6	7	18	6	0	0
	Composite	9	0.529	0.062	0.480	0.662	0	0	0	0	4	4	1
	Gap Match Multiple	17	0.405	0.105	0.155	0.548	0	1	2	4	6	4	0
	Graphic Gap Match	16	0.420	0.101	0.226	0.546	0	0	4	1	8	3	0
	Text Entry	6	0.451	0.102	0.341	0.620	0	0	0	2	3	0	1
4	Choice Multiple	9	0.453	0.091	0.299	0.565	0	0	1	2	3	3	0
	Choice Single	14	0.318	0.132	0.075	0.471	1	2	3	2	6	0	0
	Composite	5	0.557	0.132	0.332	0.671	0	0	0	1	0	2	2
	Gap Match Multiple	7	0.480	0.073	0.383	0.579	0	0	0	1	3	3	0
	Graphic Gap Match	11	0.478	0.068	0.373	0.585	0	0	0	2	5	4	0
	Hot Text	1	0.571	–	0.571	0.571	0	0	0	0	0	1	0
	Text Entry	7	0.463	0.069	0.347	0.547	0	0	0	1	3	3	0
5	Choice Multiple	16	0.392	0.131	0.138	0.560	0	2	2	2	7	3	0
	Choice Single	19	0.402	0.122	0.178	0.568	0	2	2	5	5	5	0
	Composite	11	0.572	0.084	0.426	0.672	0	0	0	0	3	2	6
	Gap Match Multiple	11	0.407	0.134	0.162	0.663	0	1	0	4	4	1	1
	Graphic Gap Match	11	0.490	0.129	0.238	0.630	0	0	1	2	3	2	3
	Hot Text	1	0.604	–	0.604	0.604	0	0	0	0	0	0	1
	Text Entry	9	0.464	0.092	0.305	0.547	0	0	0	3	2	4	0

Appendix C: Summary Item-Total Correlations by Item Type

Field Test Items													
Grade	Item Type	#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
6	Choice Multiple	36	0.377	0.112	0.150	0.590	0	3	6	9	13	5	0
	Choice Single	107	0.342	0.130	-0.110	0.548	7	5	19	39	30	7	0
	Composite	25	0.512	0.082	0.340	0.643	0	0	0	2	8	13	2
	Gap Match Multiple	16	0.427	0.126	0.175	0.640	0	1	1	3	5	5	1
	Graphic Gap Match	21	0.468	0.118	0.117	0.577	0	1	1	3	5	11	0
	Hot Text	6	0.374	0.180	0.140	0.630	0	1	1	2	0	1	1
	Schema Discrete	1	0.369	–	0.369	0.369	0	0	0	1	0	0	0
	Text Entry	41	0.435	0.086	0.182	0.591	0	1	3	7	21	9	0
7	Choice Multiple	26	0.361	0.135	0.131	0.551	0	4	7	3	7	5	0
	Choice Single	29	0.340	0.109	0.172	0.527	0	3	9	6	10	1	0
	Composite	21	0.424	0.130	0.060	0.615	1	0	4	2	8	5	1
	Gap Match Multiple	16	0.382	0.132	0.199	0.594	0	1	5	4	1	5	0
	Graphic Gap Match	1	0.572	–	0.572	0.572	0	0	0	0	0	1	0
	Hot Text	1	0.449	–	0.449	0.449	0	0	0	0	1	0	0
	Text Entry	27	0.448	0.107	0.274	0.609	0	0	4	4	10	7	2
8	Choice Multiple	26	0.321	0.118	0.106	0.573	0	4	5	12	3	2	0
	Choice Single	17	0.304	0.137	-0.073	0.466	1	3	2	6	5	0	0
	Composite	15	0.390	0.128	0.173	0.620	0	1	3	4	5	1	1
	Gap Match Multiple	12	0.340	0.159	0.056	0.492	2	1	0	1	8	0	0
	Graphic Gap Match	2	0.380	0.121	0.294	0.466	0	0	1	0	1	0	0
	Hot Text	1	0.134	–	0.134	0.134	0	1	0	0	0	0	0
	Schema Discrete	1	0.555	–	0.555	0.555	0	0	0	0	0	1	0
	Text Entry	15	0.429	0.127	0.141	0.551	0	2	0	3	3	7	0
Science													
5	Choice Multiple	10	0.427	0.095	0.239	0.559	0	0	1	3	4	2	0
	Choice Single	47	0.312	0.111	0.032	0.516	4	1	16	16	9	1	0
	Composite	7	0.431	0.100	0.266	0.523	0	0	1	1	3	2	0
	Gap Match Multiple	16	0.354	0.114	0.117	0.550	0	2	1	6	6	1	0

Appendix C: Summary Item-Total Correlations by Item Type

Field Test Items													
Grade	Item Type	#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
	Graphic Gap Match	9	0.368	0.073	0.220	0.474	0	0	1	5	3	0	0
	Hot Text	1	0.191	–	0.191	0.191	0	1	0	0	0	0	0
8	Choice Multiple	15	0.305	0.116	0.152	0.465	0	3	5	2	5	0	0
	Choice Single	44	0.340	0.118	0.037	0.525	3	2	10	14	14	1	0
	Composite	10	0.417	0.101	0.260	0.608	0	0	1	4	3	1	1
	Gap Match Multiple	18	0.325	0.111	0.011	0.486	1	0	5	8	4	0	0
	Graphic Gap Match	9	0.315	0.122	0.114	0.511	0	1	4	2	1	1	0

Appendix D: Achievement Level Distributions & Scale Score Descriptive Statistics by Demographics

Table D.1. Achievement Level Distributions & Scale Score Descriptive Statistics by Demographics—ELA

		ELA							
Grade	Demographic Sub-Group*		N	SS Descriptive Statistics		Percent of Students in Each Achievement Level**			
				Mean	SD	Level 3	Level 2	Level 1	L2 + L1
3	Grade 3 Overall		22,752	2465.67	88.23	49.8	36.2	14.0	50.2
	Gender	Female	11,041	2471.09	85.69	47.1	38.2	14.8	52.9
		Male	11,711	2460.56	90.26	52.4	34.3	13.3	47.6
	Ethnicity	AI/AN	289	2418.39	82.48	74.7	21.8	3.5	25.3
		Asian	759	2473.94	96.60	46.5	32.5	20.9	53.5
		Black	1,480	2411.97	91.97	73.0	22.4	4.6	27.0
		Hispanic	4,682	2427.28	87.93	67.8	26.7	5.4	32.2
		NH/PI	35	2440.71	94.46	62.9	25.7	11.4	37.1
		White	14,399	2484.80	80.40	41.0	41.2	17.8	59.0
		2 or more Races	1,108	2458.47	89.08	52.5	35.6	11.9	47.5
	FRL	Yes	9,983	2431.56	87.60	66.7	27.2	6.1	33.3
		No	12,769	2492.33	79.06	36.6	43.2	20.2	63.4
	LEP	Yes	3,964	2419.90	88.74	71.8	22.9	5.2	28.2
No		18,788	2475.33	85.03	45.2	39.0	15.9	54.8	
SPED	Yes	3,861	2405.51	91.01	76.8	18.5	4.7	23.2	
	No	18,891	2477.97	82.41	44.3	39.8	15.9	55.7	
4	Grade 4 Overall		22,884	2495.52	86.32	47.1	38.4	14.5	52.9
	Gender	Female	11,138	2501.57	84.40	44.5	39.7	15.8	55.5
		Male	11,746	2489.79	87.73	49.6	37.1	13.2	50.4
	Ethnicity	AI/AN	305	2430.15	83.87	78.7	19.3	2.0	21.3
		Asian	745	2500.31	94.77	45.9	36.1	18.0	54.1
		Black	1,473	2440.15	90.98	71.4	23.8	4.8	28.6
		Hispanic	4,578	2456.00	86.83	67.0	27.2	5.7	33.0
		NH/PI	43	2461.93	93.02	65.1	25.6	9.3	34.9
		White	14,635	2515.28	77.29	37.3	44.2	18.4	62.7
		2 or more Races	1,105	2487.50	87.29	53.6	33.6	12.9	46.4
	FRL	Yes	9,753	2459.57	85.64	65.3	28.7	6.1	34.7
		No	13,131	2522.23	76.59	33.7	45.6	20.7	66.3
	LEP	Yes	3,710	2445.36	87.66	71.8	24.0	4.3	28.2
No		19,174	2505.23	82.62	42.4	41.2	16.5	57.6	
SPED	Yes	3,897	2428.92	88.25	77.6	18.4	4.0	22.4	
	No	18,987	2509.19	79.28	40.9	42.5	16.6	59.1	

Appendix D: Achievement Level Distributions & Scale Score Descriptive Statistics by Demographics

ELA									
Grade	Demographic Sub-Group*		N	SS Descriptive Statistics		Percent of Students in Each Achievement Level**			
				Mean	SD	Level 3	Level 2	Level 1	L2 + L1
5	Grade 5 Overall		22,750	2516.16	81.85	52.5	32.8	14.7	47.5
	Gender	Female	11,060	2522.49	78.35	49.6	34.5	15.9	50.4
		Male	11,690	2510.16	84.59	55.2	31.2	13.6	44.8
	Ethnicity	AI/AN	268	2460.07	78.93	79.5	17.9	2.6	20.5
		Asian	697	2526.81	91.78	46.2	31.3	22.5	53.8
		Black	1,410	2465.71	83.93	75.8	19.7	4.5	24.2
		Hispanic	4,575	2482.17	81.12	70.1	24.1	5.9	29.9
		NH/PI	42	2475.81	88.33	66.7	26.2	7.1	33.3
		White	14,695	2532.80	75.14	44.3	37.2	18.5	55.7
		2 or more Races	1,063	2508.01	84.36	55.8	31.4	12.8	44.2
	FRL	Yes	9,698	2484.28	80.74	69.3	24.6	6.1	30.7
		No	13,052	2539.84	74.28	40.0	38.9	21.1	60.0
	LEP	Yes	3,446	2469.97	82.06	75.0	20.9	4.1	25.0
No		19,304	2524.40	79.02	48.5	34.9	16.6	51.5	
SPED	Yes	3,746	2448.00	84.02	82.5	14.0	3.6	17.5	
	No	19,004	2529.59	74.38	46.6	36.5	16.9	53.4	
6	Grade 6 Overall		23,413	2523.34	75.48	56.0	31.1	12.9	44.0
	Gender	Female	11,410	2527.68	71.84	54.5	32.7	12.8	45.5
		Male	12,003	2519.22	78.56	57.5	29.5	13.0	42.5
	Ethnicity	AI/AN	295	2477.43	71.70	82.4	13.6	4.1	17.6
		Asian	690	2530.18	81.48	51.2	30.6	18.3	48.8
		Black	1,570	2470.06	77.58	81.7	14.8	3.4	18.3
		Hispanic	4,772	2492.05	75.20	73.3	21.2	5.5	26.7
		NH/PI	46	2494.61	71.93	71.7	23.9	4.3	28.3
		White	14,961	2540.14	68.55	47.1	36.5	16.4	52.9
		2 or more Races	1,079	2515.78	76.02	61.3	28.2	10.6	38.7
	FRL	Yes	10,110	2494.04	75.05	72.5	21.8	5.7	27.5
		No	13,303	2545.61	67.80	43.5	38.1	18.4	56.5
	LEP	Yes	3,273	2474.57	72.65	82.5	14.6	2.9	17.5
No		20,140	2531.27	72.91	51.7	33.8	14.6	48.3	
SPED	Yes	3,498	2455.73	76.39	86.2	10.9	2.9	13.8	
	No	19,915	2535.22	68.77	50.7	34.6	14.7	49.3	

Appendix D: Achievement Level Distributions & Scale Score Descriptive Statistics by Demographics

ELA									
Grade	Demographic Sub-Group*		N	SS Descriptive Statistics		Percent of Students in Each Achievement Level**			
				Mean	SD	Level 3	Level 2	Level 1	L2 + L1
7	Grade 7 Overall		23,885	2531.12	77.33	57.6	34.7	7.7	42.4
	Gender	Female	11,637	2537.54	74.31	55.1	36.5	8.4	44.9
		Male	12,248	2525.03	79.62	60.0	32.9	7.1	40.0
	Ethnicity	AI/AN	303	2484.98	76.66	82.5	15.5	2.0	17.5
		Asian	640	2541.57	81.15	51.7	35.9	12.3	48.3
		Black	1,563	2481.53	80.24	79.5	18.6	1.9	20.5
		Hispanic	5,008	2500.91	76.37	73.8	24.0	2.2	26.2
		NH/PI	35	2518.26	90.71	60.0	34.3	5.7	40.0
		White	15,288	2547.24	71.16	49.6	40.3	10.2	50.4
		2 or more Races	1,048	2521.63	79.81	61.7	32.9	5.3	38.3
	FRL	Yes	10,157	2500.92	78.01	73.1	23.9	3.0	26.9
		No	13,728	2553.47	68.75	46.2	42.6	11.2	53.8
	LEP	Yes	2,689	2471.62	74.16	87.7	11.8	0.6	12.3
		No	21,196	2538.67	74.39	53.8	37.6	8.6	46.2
SPED	Yes	3,502	2460.17	77.05	87.9	10.5	1.7	12.1	
	No	20,383	2543.31	70.52	52.4	38.8	8.7	47.6	
8	Grade 8 Overall		23,917	2546.26	73.41	53.7	36.4	10.0	46.3
	Gender	Female	11,585	2554.05	69.39	49.6	39.4	11.0	50.4
		Male	12,332	2538.95	76.27	57.5	33.5	9.0	42.5
	Ethnicity	AI/AN	289	2515.74	69.67	73.4	24.2	2.4	26.6
		Asian	683	2556.11	80.28	48.9	34.4	16.7	51.1
		Black	1,581	2497.98	80.19	78.2	18.8	3.0	21.8
		Hispanic	4,864	2516.13	77.03	69.9	26.3	3.8	30.1
		NH/PI	34	2533.41	81.06	58.8	32.4	8.8	41.2
		White	15,417	2561.37	65.34	45.7	41.8	12.6	54.3
		2 or more Races	1,049	2539.06	75.82	56.6	34.9	8.5	43.4
	FRL	Yes	9,753	2518.89	75.08	69.3	26.5	4.2	30.7
		No	14,164	2565.11	65.93	42.9	43.2	13.9	57.1
	LEP	Yes	2,188	2480.00	77.20	87.7	11.7	0.6	12.3
		No	21,729	2552.94	69.60	50.2	38.8	10.9	49.8
SPED	Yes	3,290	2480.35	76.97	86.2	12.0	1.8	13.8	
	No	20,627	2556.78	67.08	48.5	40.2	11.3	51.5	

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

**Level 3 = Developing. Level 2 = On Track. Level 1 = CCR Benchmark.

Appendix D: Achievement Level Distributions & Scale Score Descriptive Statistics by Demographics

Table D.2. Achievement Level Distributions & Scale Score Descriptive Statistics by Demographics—Mathematics

		Mathematics							
Grade	Demographic Sub-Group*		N	SS Descriptive Statistics		Percent of Students in Each Achievement Level**			
				Mean	SD	Level 3	Level 2	Level 1	L2 + L1
3	Grade 3 Overall		22,738	1187.78	81.53	49.9	38.8	11.2	50.1
	Gender	Female	11,023	1182.87	76.63	52.4	38.7	8.9	47.6
		Male	11,715	1192.40	85.64	47.6	38.9	13.5	52.4
	Ethnicity	AI/AN	288	1138.99	72.26	75.3	22.6	2.1	24.7
		Asian	760	1207.60	93.83	43.0	36.6	20.4	57.0
		Black	1,481	1130.20	73.75	78.3	18.9	2.8	21.7
		Hispanic	4,667	1150.64	71.20	69.8	27.1	3.0	30.2
		NH/PI	35	1171.57	87.65	60.0	31.4	8.6	40.0
		White	14,399	1206.89	77.41	39.7	45.6	14.7	60.3
		2 or more Races	1,108	1172.57	78.80	59.6	31.9	8.5	40.4
	FRL	Yes	9,977	1153.81	74.98	67.7	28.2	4.1	32.3
		No	12,761	1214.35	76.43	36.0	47.2	16.8	64.0
	LEP	Yes	3,947	1147.70	74.25	71.8	24.4	3.9	28.2
No		18,791	1196.20	80.48	45.4	41.9	12.8	54.6	
SPED	Yes	3,864	1133.38	80.50	76.4	19.3	4.3	23.6	
	No	18,874	1198.92	77.15	44.5	42.8	12.7	55.5	
4	Grade 4 Overall		22,879	1215.83	78.11	53.6	36.3	10.1	46.4
	Gender	Female	11,137	1211.49	74.61	55.4	36.7	7.9	44.6
		Male	11,742	1219.95	81.09	51.8	35.9	12.2	48.2
	Ethnicity	AI/AN	304	1148.61	70.42	85.9	12.2	2.0	14.1
		Asian	745	1229.54	91.98	49.3	33.7	17.0	50.7
		Black	1,470	1157.51	70.54	82.0	15.6	2.4	18.0
		Hispanic	4,546	1180.78	70.11	73.0	23.9	3.1	27.0
		NH/PI	43	1204.84	77.71	69.8	23.3	7.0	30.2
		White	14,666	1234.50	72.90	43.5	43.4	13.2	56.5
		2 or more Races	1,105	1199.53	76.41	63.3	29.9	6.8	36.7
	FRL	Yes	9,745	1181.04	71.36	72.6	24.0	3.4	27.4
		No	13,134	1241.64	72.71	39.5	45.4	15.1	60.5
	LEP	Yes	3,677	1174.85	71.12	75.5	21.6	2.9	24.5
No		19,202	1223.67	76.93	49.4	39.1	11.5	50.6	
SPED	Yes	3,898	1163.41	72.89	80.3	16.7	3.0	19.7	
	No	18,981	1226.59	74.72	48.1	40.3	11.6	51.9	

Appendix D: Achievement Level Distributions & Scale Score Descriptive Statistics by Demographics

Mathematics									
Grade	Demographic Sub-Group*		N	SS Descriptive Statistics		Percent of Students in Each Achievement Level**			
				Mean	SD	Level 3	Level 2	Level 1	L2 + L1
5	Grade 5 Overall		22,721	1231.90	75.55	50.8	40.1	9.0	49.2
	Gender	Female	11,039	1229.19	71.51	52.2	40.4	7.4	47.8
		Male	11,682	1234.46	79.09	49.5	39.8	10.6	50.5
	Ethnicity	AI/AN	268	1176.11	66.10	82.5	16.4	1.1	17.5
		Asian	696	1253.32	94.86	42.8	37.2	20.0	57.2
		Black	1,408	1175.58	69.62	79.9	18.0	2.1	20.1
		Hispanic	4,548	1200.48	67.99	69.6	27.3	3.1	30.4
		NH/PI	42	1198.71	84.93	64.3	31.0	4.8	35.7
		White	14,696	1248.20	71.07	41.4	47.2	11.4	58.6
		2 or more Races	1,063	1216.98	72.62	59.5	34.3	6.1	40.5
	FRL	Yes	9,683	1200.43	69.15	68.8	28.1	3.1	31.2
		No	13,038	1255.27	71.51	37.5	49.0	13.4	62.5
	LEP	Yes	3,418	1192.31	69.08	73.5	23.9	2.6	26.5
		No	19,303	1238.91	74.48	46.8	43.0	10.2	53.2
SPED	Yes	3,741	1174.65	70.73	81.5	16.0	2.5	18.5	
	No	18,980	1243.18	71.23	44.8	44.9	10.3	55.2	
6	Grade 6 Overall		23,368	1238.93	73.74	54.1	37.4	8.5	45.9
	Gender	Female	11,394	1237.84	70.43	54.7	38.1	7.2	45.3
		Male	11,974	1239.96	76.76	53.4	36.8	9.8	46.6
	Ethnicity	AI/AN	294	1185.53	67.13	82.0	15.6	2.4	18.0
		Asian	691	1259.36	90.85	45.0	34.3	20.7	55.0
		Black	1,560	1177.46	69.24	83.6	15.1	1.3	16.4
		Hispanic	4,742	1210.21	67.67	71.5	25.1	3.4	28.5
		NH/PI	45	1222.96	76.67	55.6	40.0	4.4	44.4
		White	14,955	1255.78	68.19	44.6	44.6	10.7	55.4
		2 or more Races	1,081	1222.67	72.00	63.4	31.6	5.0	36.6
	FRL	Yes	10,076	1207.62	67.80	72.2	24.9	2.9	27.8
		No	13,292	1262.66	69.07	40.3	46.9	12.8	59.7
	LEP	Yes	3,247	1196.33	65.88	78.1	19.8	2.1	21.9
		No	20,121	1245.80	72.63	50.2	40.3	9.6	49.8
SPED	Yes	3,495	1179.95	66.61	85.2	13.1	1.7	14.8	
	No	19,873	1249.30	69.97	48.6	41.7	9.7	51.4	

Appendix D: Achievement Level Distributions & Scale Score Descriptive Statistics by Demographics

Mathematics									
Grade	Demographic Sub-Group*		N	SS Descriptive Statistics		Percent of Students in Each Achievement Level**			
				Mean	SD	Level 3	Level 2	Level 1	L2 + L1
7	Grade 7 Overall		23,829	1241.28	73.19	55.5	35.8	8.7	44.5
	Gender	Female	11,601	1240.01	69.47	56.6	35.8	7.6	43.4
		Male	12,228	1242.49	76.53	54.4	35.7	9.8	45.6
	Ethnicity	AI/AN	304	1193.30	66.50	81.6	15.8	2.6	18.4
		Asian	638	1265.17	93.54	44.4	35.9	19.7	55.6
		Black	1,559	1188.42	63.10	84.2	13.7	2.1	15.8
		Hispanic	4,975	1212.74	64.35	72.9	24.1	3.0	27.1
		NH/PI	35	1222.77	92.04	65.7	28.6	5.7	34.3
		White	15,275	1256.93	70.09	46.2	42.7	11.1	53.8
		2 or more Races	1,043	1227.27	70.39	63.9	29.5	6.6	36.1
	FRL	Yes	10,121	1210.80	64.97	73.3	23.6	3.0	26.7
		No	13,708	1263.79	70.73	42.3	44.8	13.0	57.7
	LEP	Yes	2,660	1192.12	60.32	83.2	15.3	1.5	16.8
No		21,169	1247.46	72.32	52.0	38.4	9.7	48.0	
SPED	Yes	3,494	1182.35	60.64	87.4	11.0	1.6	12.6	
	No	20,335	1251.41	70.33	50.0	40.0	10.0	50.0	
8	Grade 8 Overall		23,856	1250.99	76.67	58.1	34.5	7.4	41.9
	Gender	Female	11,554	1252.57	72.31	57.5	36.1	6.5	42.5
		Male	12,302	1249.50	80.52	58.8	33.0	8.2	41.2
	Ethnicity	AI/AN	287	1206.75	62.97	82.2	16.4	1.4	17.8
		Asian	682	1282.32	97.63	45.7	32.4	21.8	54.3
		Black	1,568	1195.99	68.65	85.5	12.9	1.6	14.5
		Hispanic	4,826	1218.89	68.36	76.0	21.8	2.2	24.0
		NH/PI	34	1248.44	69.19	61.8	35.3	2.9	38.2
		White	15,409	1266.95	72.38	49.4	41.5	9.1	50.6
		2 or more Races	1,050	1238.32	80.16	65.3	28.6	6.1	34.7
	FRL	Yes	9,707	1218.78	69.16	75.8	21.8	2.5	24.2
		No	14,149	1273.09	73.68	46.0	43.3	10.7	54.0
	LEP	Yes	2,156	1192.45	61.69	88.2	11.2	0.6	11.8
No		21,700	1256.81	75.56	55.2	36.8	8.0	44.8	
SPED	Yes	3,273	1186.55	64.53	89.1	9.9	1.0	10.9	
	No	20,583	1261.24	73.39	53.2	38.4	8.4	46.8	

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

**Level 3 = Developing. Level 2 = On Track. Level 1 = CCR Benchmark.

Appendix D: Achievement Level Distributions & Scale Score Descriptive Statistics by Demographics

Table D.3. Achievement Level Distributions & Scale Score Descriptive Statistics by Demographics—Science

		Science							
Grade	Demographic Sub-Group*		N	SS Descriptive Statistics		Percent of Students in Each Achievement Level*			
				Mean	SD	Level 3	Level 2	Level 1	L2 + L1
5	Grade 5 Overall		22,727	3115.85	31.46	28.8	54.8	16.5	71.2
	Gender	Female	11,047	3114.84	30.20	29.2	56.0	14.8	70.8
		Male	11,680	3116.80	32.57	28.4	53.6	18.0	71.6
	Ethnicity	AI/AN	266	3095.39	27.92	57.9	38.7	3.4	42.1
		Asian	698	3116.05	32.91	29.5	52.7	17.8	70.5
		Black	1,411	3094.20	27.22	59.3	36.4	4.3	40.7
		Hispanic	4,552	3102.72	28.74	44.7	48.3	7.0	55.3
		NH/PI	42	3096.88	29.15	54.8	42.9	2.4	45.2
		White	14,694	3122.76	30.23	20.0	58.9	21.1	80.0
		2 or more Races	1,064	3110.95	30.03	32.5	55.5	12.0	67.5
	FRL	Yes	9,398	3104.27	29.42	43.1	48.8	8.2	56.9
		No	13,329	3124.01	30.26	18.7	59.0	22.3	81.3
LEP	Yes	3,434	3097.88	27.46	51.1	44.2	4.7	48.9	
	No	19,293	3119.04	31.04	24.8	56.7	18.5	75.2	
SPED	Yes	3,834	3095.69	29.94	57.5	36.4	6.1	42.5	
	No	18,893	3119.94	30.15	22.9	58.5	18.6	77.1	
8	Grade 8 Overall		23,856	3107.46	30.66	36.7	54.5	8.8	63.3
	Gender	Female	11,555	3108.05	29.65	35.1	56.6	8.3	64.9
		Male	12,301	3106.91	31.56	38.2	52.5	9.4	61.8
	Ethnicity	AI/AN	286	3090.40	27.41	62.2	36.0	1.7	37.8
		Asian	682	3112.88	33.57	31.4	52.6	16.0	68.6
		Black	1,576	3083.65	28.83	71.0	27.3	1.6	29.0
		Hispanic	4,824	3094.53	28.27	54.3	42.5	3.2	45.7
		NH/PI	34	3100.91	28.64	50.0	47.1	2.9	50.0
		White	15,405	3114.37	28.67	26.9	61.8	11.3	73.1
		2 or more Races	1,049	3102.58	31.19	43.2	49.8	7.1	56.8
	FRL	Yes	9,157	3095.46	29.17	53.3	43.2	3.5	46.7
		No	14,699	3114.94	29.15	26.3	61.5	12.1	73.7
LEP	Yes	2,154	3082.56	26.00	73.2	26.1	0.7	26.8	
	No	21,702	3109.93	29.98	33.1	57.3	9.6	66.9	
SPED	Yes	3,325	3082.86	27.35	73.2	25.3	1.5	26.8	
	No	20,531	3111.45	29.28	30.8	59.2	10.0	69.2	

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

**Level 3 = Below the Standards. Level 2 = Meets the Standards. Level 1 = Exceeds the Standards.

Appendix E: Marginal Reliability by Demographics

Table E.1. Marginal Reliability by Demographics—ELA

ELA						
Grade	Demographic Sub-Group*	N	Variance	MSE	Marginal Reliability	
3	Grade 3 Overall		22,752	7784.3	677.0	0.91
	Gender	Female	11,041	7343.5	671.1	0.91
		Male	11,711	8146.8	682.5	0.92
	Ethnicity	AI/AN	289	6802.8	704.2	0.90
		Asian	759	9331.9	688.3	0.93
		Black	1,480	8459.2	725.3	0.91
		Hispanic	4,682	7731.3	701.1	0.91
		NH/PI	35	8921.8	686.8	0.92
		White	14,399	6463.6	662.7	0.90
		2 or more Races	1,108	7935.3	680.9	0.91
	FRL	Yes	9,983	7673.2	695.9	0.91
		No	12,769	6251.3	662.2	0.89
	LEP	Yes	3,964	7875.3	710.3	0.91
		No	18,788	7230.3	669.9	0.91
SPED	Yes	3,861	8282.8	735.2	0.91	
	No	18,891	6791.9	665.1	0.90	
4	Grade 4 Overall		22,884	7451.9	691.8	0.91
	Gender	Female	11,138	7123.8	693.7	0.90
		Male	11,746	7696.1	690.0	0.91
	Ethnicity	AI/AN	305	7034.4	689.3	0.90
		Asian	745	8980.7	721.6	0.92
		Black	1,473	8277.8	694.7	0.92
		Hispanic	4,578	7539.9	680.0	0.91
		NH/PI	43	8653.4	686.8	0.92
		White	14,635	5973.0	693.7	0.88
		2 or more Races	1,105	7620.2	693.1	0.91
	FRL	Yes	9,753	7334.0	678.8	0.91
		No	13,131	5866.5	701.5	0.88
	LEP	Yes	3,710	7684.0	685.9	0.91
		No	19,174	6826.1	693.0	0.90
SPED	Yes	3,897	7788.5	697.5	0.91	
	No	18,987	6285.8	690.7	0.89	

Appendix F: Scatterplots for Scale Score CSEM

ELA						
Grade	Demographic Sub-Group*		N	Variance	MSE	Marginal Reliability
5	Grade 5 Overall		22,750	6699.2	684.4	0.90
	Gender	Female	11,060	6139.3	680.9	0.89
		Male	11,690	7155.7	687.7	0.90
	Ethnicity	AI/AN	268	6229.4	698.7	0.89
		Asian	697	8423.6	712.8	0.92
		Black	1,410	7044.0	699.8	0.90
		Hispanic	4,575	6580.2	683.9	0.90
		NH/PI	42	7802.9	690.5	0.91
		White	14,695	5646.7	681.4	0.88
		2 or more Races	1,063	7116.6	684.8	0.90
	FRL	Yes	9,698	6518.3	682.9	0.90
		No	13,052	5518.1	685.4	0.88
	LEP	Yes	3,446	6733.1	693.7	0.90
		No	19,304	6244.7	682.7	0.89
SPED	Yes	3,746	7060.0	721.0	0.90	
	No	19,004	5532.2	677.1	0.88	
6	Grade 6 Overall		23,413	5696.9	652.4	0.89
	Gender	Female	11,410	5161.4	644.8	0.88
		Male	12,003	6171.5	659.6	0.89
	Ethnicity	AI/AN	295	5140.2	675.8	0.87
		Asian	690	6639.2	672.4	0.90
		Black	1,570	6019.3	699.2	0.88
		Hispanic	4,772	5655.2	664.0	0.88
		NH/PI	46	5173.8	648.6	0.87
		White	14,961	4698.9	641.8	0.86
		2 or more Races	1,079	5778.8	659.8	0.89
	FRL	Yes	10,110	5633.1	663.8	0.88
		No	13,303	4597.3	643.7	0.86
	LEP	Yes	3,273	5278.1	680.4	0.87
		No	20,140	5315.7	647.8	0.88
SPED	Yes	3,498	5834.8	718.3	0.88	
	No	19,915	4728.7	640.8	0.86	

Appendix F: Scatterplots for Scale Score CSEM

ELA						
Grade	Demographic Sub-Group*		N	Variance	MSE	Marginal Reliability
7	Grade 7 Overall		23,885	5979.3	637.8	0.89
	Gender	Female	11,637	5521.3	629.2	0.89
		Male	12,248	6338.7	646.0	0.90
	Ethnicity	AI/AN	303	5877.1	695.9	0.88
		Asian	640	6585.3	641.3	0.90
		Black	1,563	6438.8	702.9	0.89
		Hispanic	5,008	5832.7	662.6	0.89
		NH/PI	35	8227.9	671.5	0.92
		White	15,288	5064.0	620.6	0.88
		2 or more Races	1,048	6368.9	653.1	0.90
	FRL	Yes	10,157	6085.0	667.8	0.89
		No	13,728	4727.0	615.6	0.87
	LEP	Yes	2,689	5499.6	711.1	0.87
		No	21,196	5534.3	628.5	0.89
SPED	Yes	3,502	5937.1	739.5	0.88	
	No	20,383	4973.2	620.3	0.88	
8	Grade 8 Overall		23,917	5388.7	666.0	0.88
	Gender	Female	11,585	4814.7	657.2	0.86
		Male	12,332	5817.8	674.4	0.88
	Ethnicity	AI/AN	289	4854.5	687.0	0.86
		Asian	683	6444.2	676.0	0.90
		Black	1,581	6430.5	730.0	0.89
		Hispanic	4,864	5934.0	695.2	0.88
		NH/PI	34	6570.9	677.4	0.90
		White	15,417	4268.7	648.6	0.85
		2 or more Races	1,049	5749.2	677.8	0.88
	FRL	Yes	9,753	5637.4	690.7	0.88
		No	14,164	4346.2	649.1	0.85
	LEP	Yes	2,188	5959.3	748.4	0.87
		No	21,729	4844.7	657.8	0.86
SPED	Yes	3,290	5923.6	749.2	0.87	
	No	20,627	4500.1	652.8	0.85	

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

Table E.2. Marginal Reliability by Demographics—Mathematics

		Mathematics				
Grade	Demographic Sub-Group*	N	Variance	MSE	Marginal Reliability	
3	Grade 3 Overall		22,738	6647.5	364.4	0.95
	Gender	Female	11,023	5872.2	358.2	0.94
		Male	11,715	7333.5	370.3	0.95
	Ethnicity	AI/AN	288	5222.1	360.5	0.93
		Asian	760	8803.4	392.0	0.96
		Black	1,481	5439.0	361.9	0.93
		Hispanic	4,667	5069.8	352.9	0.93
		NH/PI	35	7682.0	386.1	0.95
		White	14,399	5991.6	367.4	0.94
		2 or more Races	1,108	6209.0	359.3	0.94
	FRL	Yes	9,977	5621.4	356.7	0.94
		No	12,761	5842.0	370.5	0.94
	LEP	Yes	3,947	5512.9	356.3	0.94
		No	18,791	6477.8	366.1	0.94
SPED	Yes	3,864	6480.6	366.1	0.94	
	No	18,874	5952.0	364.1	0.94	
4	Grade 4 Overall		22,879	6101.7	350.6	0.94
	Gender	Female	11,137	5566.2	347.6	0.94
		Male	11,742	6575.2	353.5	0.95
	Ethnicity	AI/AN	304	4958.4	369.5	0.93
		Asian	745	8460.9	375.4	0.96
		Black	1,470	4975.6	362.7	0.93
		Hispanic	4,546	4915.5	349.2	0.93
		NH/PI	43	6039.5	370.7	0.94
		White	14,666	5314.6	348.1	0.93
		2 or more Races	1,105	5839.1	351.3	0.94
	FRL	Yes	9,745	5091.9	351.3	0.93
		No	13,134	5286.7	350.1	0.93
	LEP	Yes	3,677	5058.2	353.8	0.93
		No	19,202	5918.6	350.0	0.94
SPED	Yes	3,898	5313.3	361.7	0.93	
	No	18,981	5583.7	348.3	0.94	

Appendix F: Scatterplots for Scale Score CSEM

Mathematics						
Grade	Demographic Sub-Group*		N	Variance	MSE	Marginal Reliability
5	Grade 5 Overall		22,721	5707.6	336.1	0.94
	Gender	Female	11,039	5113.4	331.2	0.94
		Male	11,682	6256.0	340.8	0.95
	Ethnicity	AI/AN	268	4368.8	336.6	0.92
		Asian	696	8998.9	392.4	0.96
		Black	1,408	4847.1	339.9	0.93
		Hispanic	4,548	4622.0	329.0	0.93
		NH/PI	42	7212.3	345.2	0.95
		White	14,696	5051.1	335.6	0.93
		2 or more Races	1,063	5274.1	331.3	0.94
	FRL	Yes	9,683	4782.1	330.2	0.93
		No	13,038	5113.9	340.4	0.93
	LEP	Yes	3,418	4772.3	333.3	0.93
		No	19,303	5546.7	336.6	0.94
SPED	Yes	3,741	5003.4	342.4	0.93	
	No	18,980	5073.3	334.8	0.93	
6	Grade 6 Overall		23,368	5438.3	346.3	0.94
	Gender	Female	11,394	4959.8	344.4	0.93
		Male	11,974	5891.9	348.1	0.94
	Ethnicity	AI/AN	294	4506.6	367.3	0.92
		Asian	691	8253.5	359.7	0.96
		Black	1,560	4794.2	375.8	0.92
		Hispanic	4,742	4579.7	350.9	0.92
		NH/PI	45	5877.8	350.7	0.94
		White	14,955	4650.4	340.3	0.93
		2 or more Races	1,081	5184.1	351.1	0.93
	FRL	Yes	10,076	4596.3	353.8	0.92
		No	13,292	4770.5	340.6	0.93
	LEP	Yes	3,247	4340.7	358.3	0.92
		No	20,121	5275.5	344.3	0.93
SPED	Yes	3,495	4436.6	370.3	0.92	
	No	19,873	4895.4	342.1	0.93	

Appendix F: Scatterplots for Scale Score CSEM

Mathematics						
Grade	Demographic Sub-Group*		N	Variance	MSE	Marginal Reliability
7	Grade 7 Overall		23,829	5356.3	360.1	0.93
	Gender	Female	11,601	4826.0	357.3	0.93
		Male	12,228	5856.9	362.6	0.94
	Ethnicity	AI/AN	304	4422.4	393.4	0.91
		Asian	638	8750.5	374.1	0.96
		Black	1,559	3982.2	396.5	0.90
		Hispanic	4,975	4141.3	371.1	0.91
		NH/PI	35	8470.8	391.3	0.95
		White	15,275	4912.6	350.9	0.93
		2 or more Races	1,043	4954.9	367.3	0.93
	FRL	Yes	10,121	4220.5	374.7	0.91
		No	13,708	5002.6	349.2	0.93
	LEP	Yes	2,660	3638.9	389.3	0.89
		No	21,169	5230.5	356.4	0.93
SPED	Yes	3,494	3677.1	400.1	0.89	
	No	20,335	4945.8	353.2	0.93	
8	Grade 8 Overall		23,856	5877.9	368.9	0.94
	Gender	Female	11,554	5228.9	363.6	0.93
		Male	12,302	6483.4	373.8	0.94
	Ethnicity	AI/AN	287	3964.8	391.6	0.90
		Asian	682	9531.5	383.2	0.96
		Black	1,568	4712.6	409.3	0.91
		Hispanic	4,826	4672.7	383.9	0.92
		NH/PI	34	4786.7	361.4	0.92
		White	15,409	5239.3	358.2	0.93
		2 or more Races	1,050	6424.9	380.4	0.94
	FRL	Yes	9,707	4782.8	385.4	0.92
		No	14,149	5429.4	357.5	0.93
	LEP	Yes	2,156	3806.0	406.5	0.89
		No	21,700	5709.6	365.1	0.94
SPED	Yes	3,273	4163.9	417.4	0.90	
	No	20,583	5385.4	361.2	0.93	

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

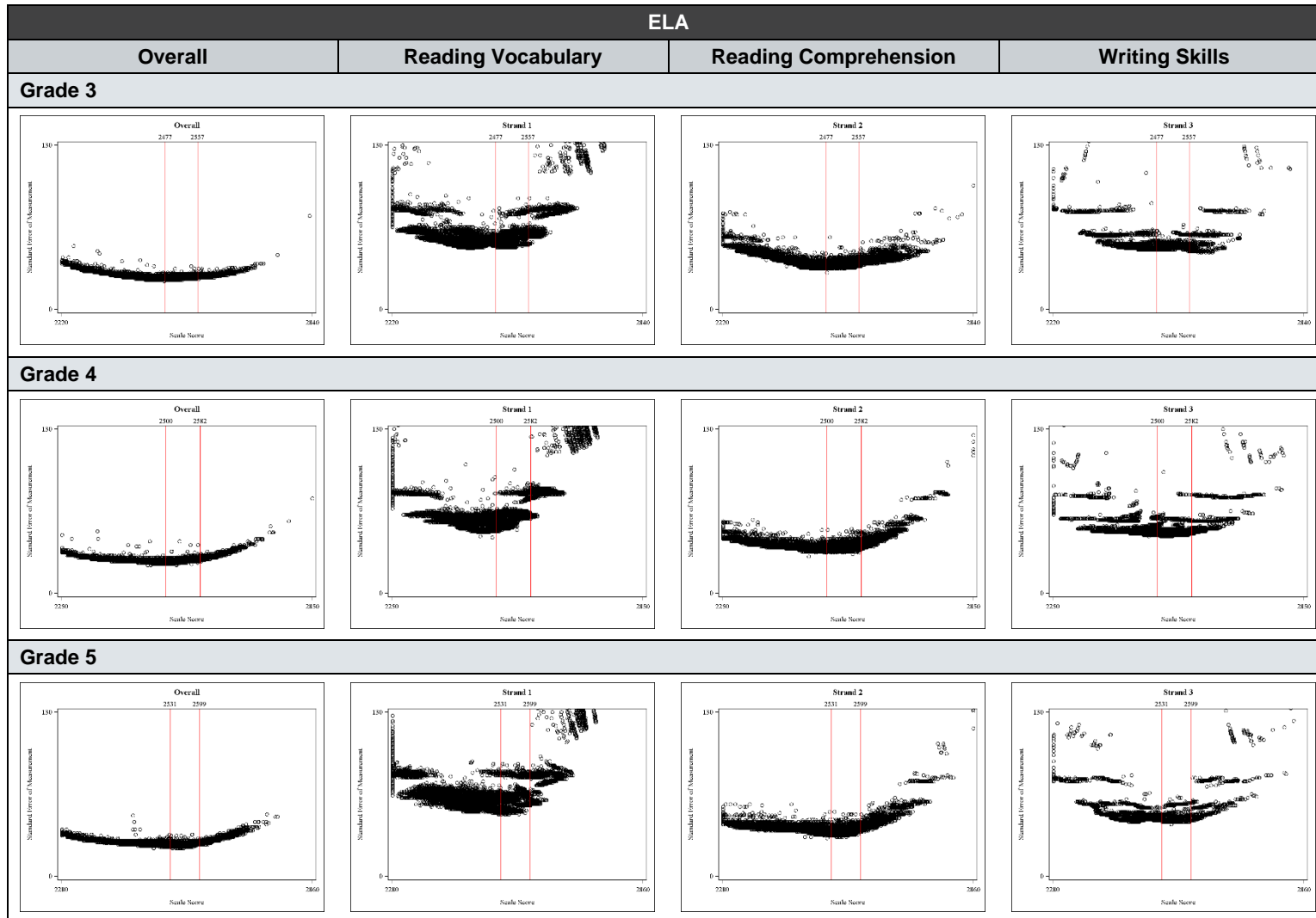
Table E.3. Marginal Reliability by Demographics—Science

		Science				
Grade	Demographic Sub-Group*	N	Variance	MSE	Marginal Reliability	
5	Grade 5 Overall		22,727	989.4	197.0	0.80
	Gender	Female	11,047	911.8	185.9	0.80
		Male	11,680	1061.0	207.4	0.80
	Ethnicity	AI/AN	266	779.5	158.7	0.80
		Asian	698	1083.4	207.8	0.81
		Black	1,411	741.1	154.0	0.79
		Hispanic	4,552	825.8	162.0	0.80
		NH/PI	42	849.9	157.7	0.81
		White	14,694	913.9	213.4	0.77
		2 or more Races	1,064	902.0	180.3	0.80
	FRL	Yes	9,398	865.8	166.2	0.81
		No	13,329	915.7	218.7	0.76
	LEP	Yes	3,434	754.3	154.5	0.80
		No	19,293	963.7	204.5	0.79
SPED	Yes	3,834	896.6	165.0	0.82	
	No	18,893	909.1	203.5	0.78	
8	Grade 8 Overall		23,856	939.9	141.1	0.85
	Gender	Female	11,555	879.4	137.8	0.84
		Male	12,301	996.2	144.1	0.86
	Ethnicity	AI/AN	286	751.3	165.7	0.78
		Asian	682	1127.2	144.0	0.87
		Black	1,576	831.4	194.3	0.77
		Hispanic	4,824	799.3	157.0	0.80
		NH/PI	34	820.1	141.0	0.83
		White	15,405	822.1	129.5	0.84
		2 or more Races	1,049	972.9	148.4	0.85
	FRL	Yes	9,157	850.8	157.9	0.81
		No	14,699	849.9	130.6	0.85
	LEP	Yes	2,154	675.8	187.1	0.72
		No	21,702	898.5	136.5	0.85
SPED	Yes	3,325	747.9	188.2	0.75	
	No	20,531	857.1	133.4	0.84	

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

Appendix F: Scatterplots for Scale Score CSEM

Figure F.1. Scatterplots for Scale Score CSEM—ELA



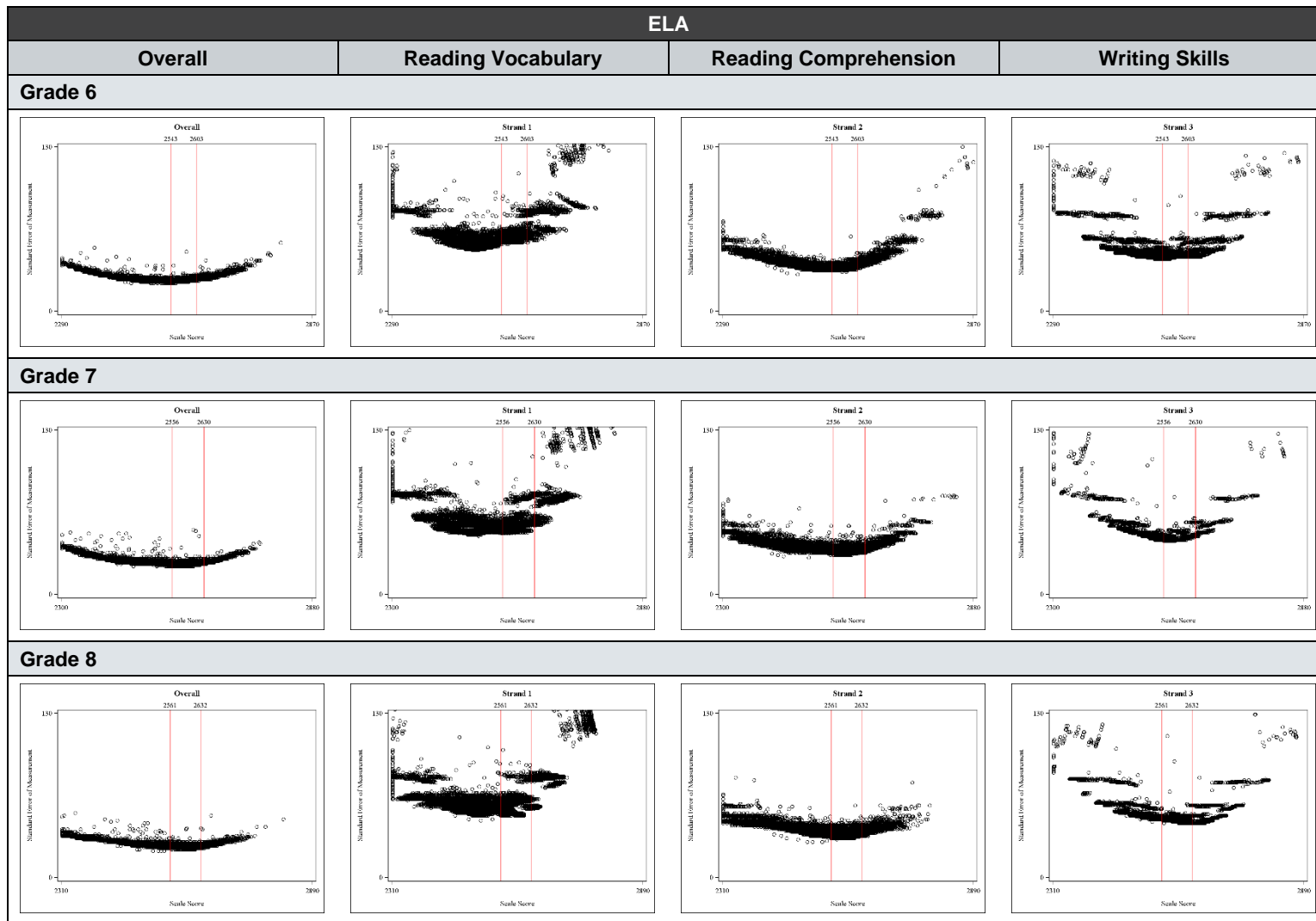
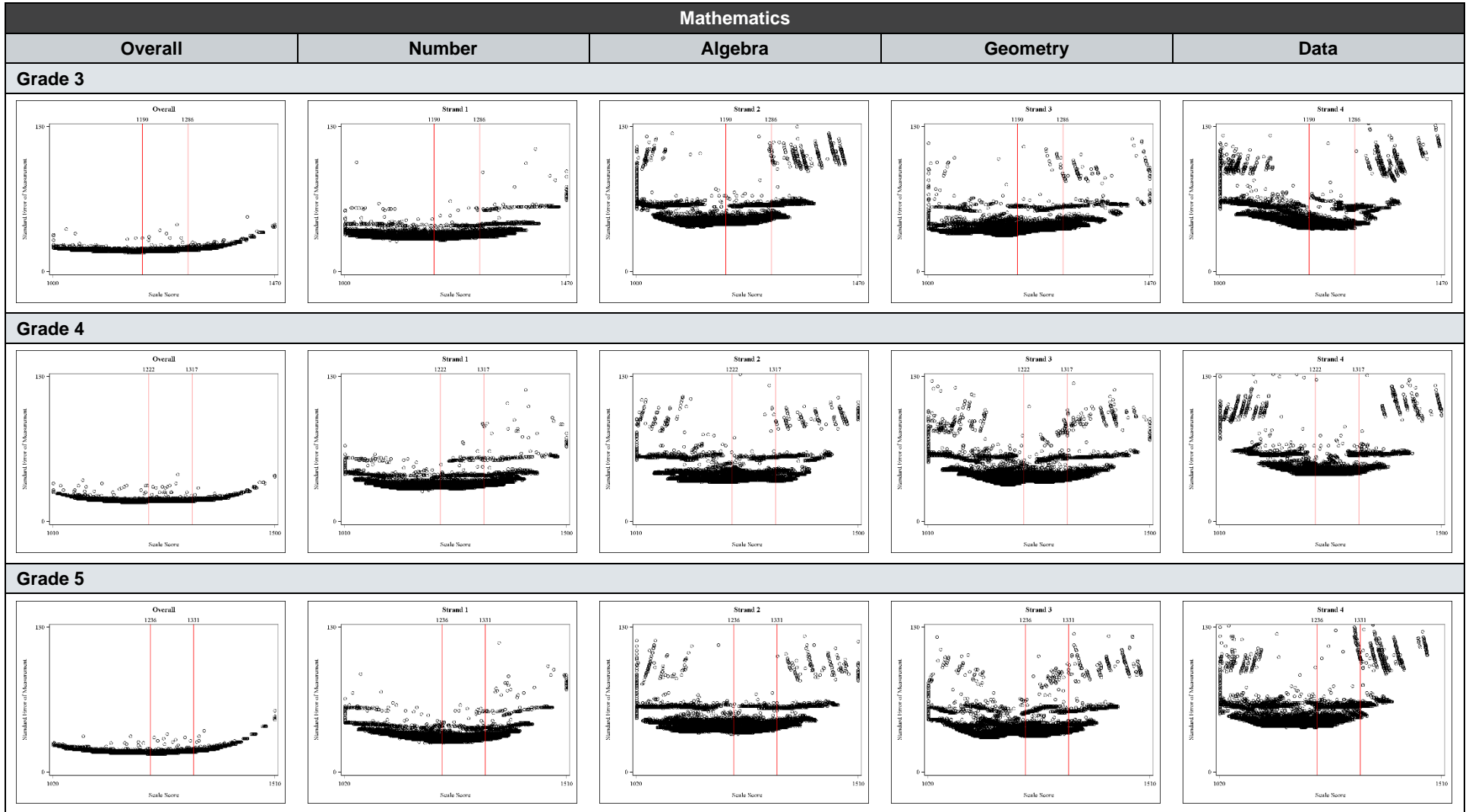


Figure F.2. Scatterplots for Scale Score CSEM—Mathematics



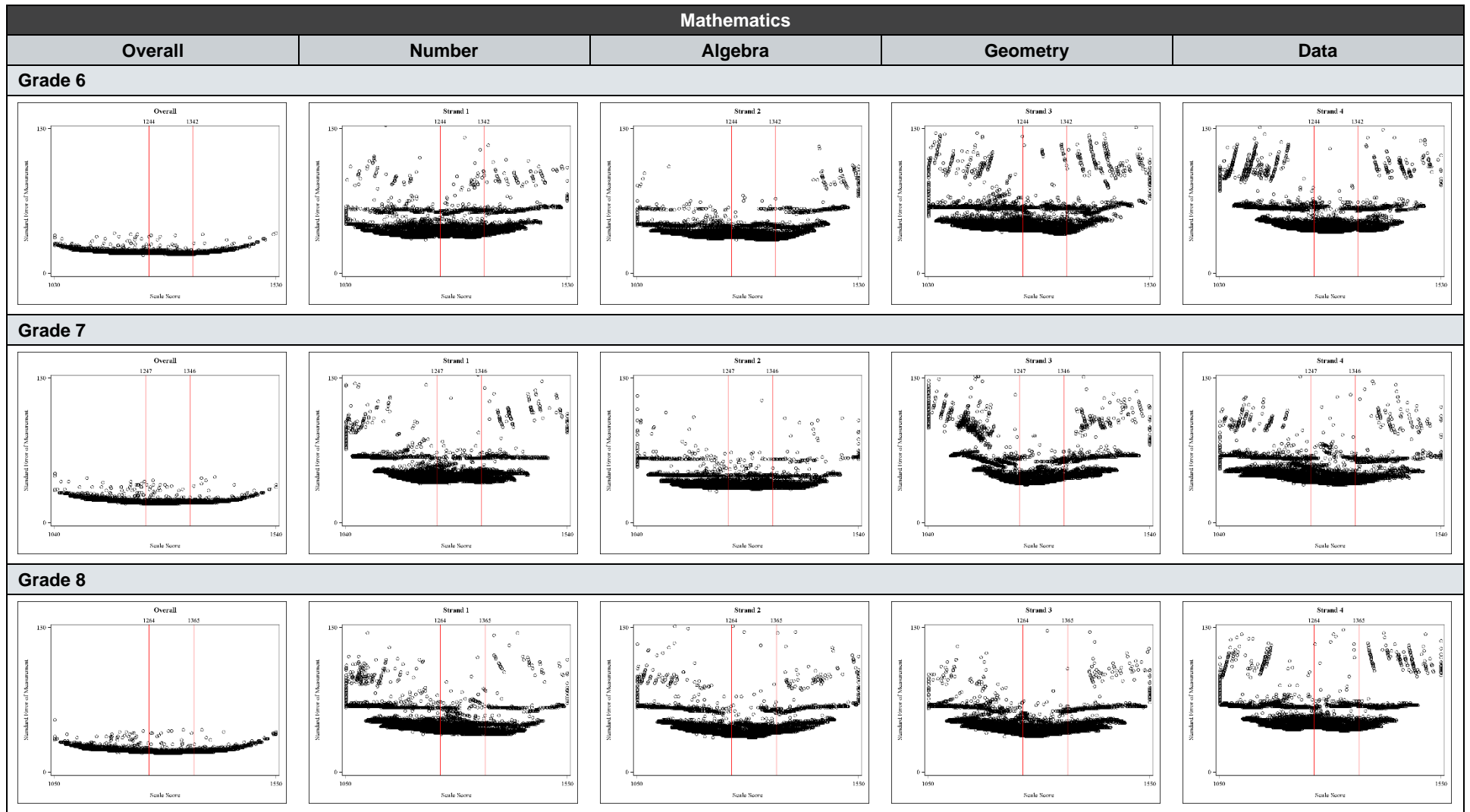


Figure F.3. Scatterplots for Scale Score CSEM—Science

