Spring 2021 NSCAS Phase I Pilot ELA, Mathematics, and Science Technical Report

October 25, 2021

NWEA Operational Content and Psychometrics

Table of Contents

Executive Summary **7**

1. Introduction **10**
   1.1 NSCAS Overview 10
   1.2 Background 11
   1.3 Schedule of Major Events 11
   1.4 Building a Validity Argument 12
      1.4.1 Intended Purposes and Uses of Test Results 13
      1.4.2 Theory of Action 14

2. Test Design and Development **16**
   2.1 Test Design 16
   2.2 Academic Content Standards 17
   2.3 Blueprints 17
   2.4 Item Types 18
   2.5 Depth of Knowledge (DOK) 18
   2.6 ALD Development 19
      2.6.1 Policy ALDs 19
      2.6.2 Range ALDs 20
         2.6.2.1 ELA and Mathematics 20
         2.6.2.2 Science 22
      2.6.3 Reporting ALDs 23
   2.7 ELA Passage Development 23
   2.8 Item Development 23
      2.8.1 Item Specifications 24
      2.8.2 Item Retirement 24
   2.9 Content Alignment 24
      2.9.1 Alignment and Adaptive Testing 25
      2.9.2 2019 Mathematics Alignment Study 26
   2.10 Universal Design 26
   2.11 Sensitivity and Fairness 26
   2.12 Test Construction 27
   2.13 Data Review of Field Tested Items 27

3. Test Administration and Security **30**
   3.1 User Roles and Responsibilities 30
   3.2 Administration Training 31
   3.3 Item Type Samplers 32
   3.4 Accommodations and Accessibility Features 32
   3.5 User Acceptance Testing (UAT) 35
   3.6 Student Participation 36
      3.6.1 Paper-Pencil Participation Criteria 37
      3.6.2 Participation of English Language Learners (ELLs) 37
      3.6.3 Participation of Recently Arrived Limited English Proficient Students 37
   3.7 Test Security 38
      3.7.1 Test Security 38

3.7.1.1 Physical Warehouse Security. 38
3.7.1.2 Secure Destruction of Test Materials. 38
3.7.1.3 Shipping Security. 38
3.7.1.4 Electronic Security of Test Materials and Data. 39
3.7.2 Caveon Test Security 39
3.7.2.1 Monitoring for Disclosure of Test Content. 39
3.7.2.2 Monitoring for Potential Test Security Violations. 39
3.8 Partner Support 41

4. Scoring and Reporting **43**
4.1 Scoring Rules 43
4.2 Paper-Pencil Scoring 43
4.3 Score Reporting Methods 43
4.4 Report Summary 45
4.5 Report Process 46
4.5.1 Online Reports 46
4.5.2 Printed ISRs 46
4.5.3 Report Verification 46
4.6 NSCAS Matrix 48

5. Constraint-Based Engine **50**
5.1 Overview 50
5.2 Engine Simulations and Evaluation 51
5.2.1 Evaluation Criteria 51
5.2.2 Blueprint Constraint Accuracy 53
5.2.3 Item Exposure Rates 56
5.2.4 Score Precision and Reliability 57

6. Psychometric Analyses **63**
6.1 Number of Student Included in the Analyses 63
6.2 Classical Item Analyses 64
6.2.1 Item Difficulty (P-Value) 64
6.2.2 Item Discrimination (Item-Total Correlation) 65
6.2.3 Item Suppression 66
6.3 Differential Item Functioning (DIF) 67
6.3.1 Logistic Regression (LR) DIF Method 68
6.3.2 Mantel-Haenszel (MH) DIF Methods 69
6.3.3 DIF Results 71
6.4 IRT Calibration 80
6.4.1 Calibration Methods 80
6.4.2 Calibration Results 81
6.5 Science Field Test 83
6.5.1 Design 84
6.5.2 Constraint-Based Engine 85
6.5.3 Analyses and Calibration 85
6.6 Common Item Linking Between NSCAS and MAP Growth (ELA and Mathematics) . 85
6.6.1 Embedded MAP Growth Items 86
6.6.2 Data 86

6.6.3 Linking Procedure . . . 87
6.6.4 Linking Results . . . 88
6.6.5 Further Considerations . . . 90
6.7 Scaling . . . 91
6.7.1 NSCAS Scale Score . . . 91
6.7.2 Linked RIT Score . . . 93

7. Standard Setting . . . **95**
7.1 Overview . . . 95
7.2 ID Matching Method . . . 96
7.3 Meeting Process . . . 96
7.4 RALD Revision . . . 97
7.5 Final Results . . . 97

8. Test Results . . . **99**
8.1 Demographics and Accommodations . . . 99
8.2 Administration Mode (Online vs. Paper-Pencil) . . . 105
8.3 Testing Time . . . 106
8.4 Achievement Level Distributions . . . 108
8.5 Descriptive Statistics of Scale Scores . . . 109
8.6 Reporting Category Correlations . . . 110
8.7 Correlations with MAP Growth . . . 110
8.8 Score Differences . . . 111

9. Reliability . . . **112**
9.1 Marginal Reliability . . . 112
9.2 Conditional Standard Error of Measurement (CSEM) . . . 113
9.3 Classification Accuracy . . . 115
9.4 Reliability for Fixed Forms (Science) . . . 119

10. Validity . . . **126**
10.1 Intended Purposes and Uses of Test Scores . . . 126
10.2 Sources of Validity Evidence . . . 127
10.3 Evidentiary Validity Framework . . . 128
10.4 Interpretive Argument Claims . . . 131
10.5 NSCAS Validity Argument . . . 132

Appendix A. Data Review Cheat Sheet . . . **138**

Appendix B: Summary P-Values by Item Type . . . **143**

Appendix C: Summary Item-Total Correlations by Item Type . . . **147**

Appendix D. Achievement Level Distributions and Scale Score Descriptive Statistics by Demographics . . . **151**

Appendix E: Marginal Reliability by Demographics . . . **160**

Appendix F. Scatterplots for Scale Score CSEM . . . **165**

List of Tables

Table 1.1 Schedule of Major Events for the Spring 2021 Administration 12
Table 2.1 NSCAS Phase I Pilot Assessments in 2021 16
Table 2.2 Number of Items and Points Per Test 17
Table 2.3 Online Item Types 18
Table 2.4 Data Review Flagging Criteria: Multiple-Choice Items 28
Table 2.5 Data Review Flagging Criteria: Non-Multiple-Choice Items 28
Table 2.6 Data Review Results for 2021 Field Test Items 29
Table 3.1 User Roles and Responsibilities 31
Table 3.2 Test Administration Workshop Dates and Participation 32
Table 3.3 Accommodations and Universal Features 33
Table 3.4 Statistical Analysis and Potential Incidents 40
Table 3.5 Partner Support Communication Options 41
Table 3.6 Test Administration Workshop Dates and Participation 41
Table 4.1 Attemptedness Rules for Scoring 43
Table 4.2 Score Range (LOSS and HOSS) for NSCAS scale score and linked RIT score . . 44
Table 4.3 Achievement Level Descriptions 44
Table 4.4 Non-Tested Code 45
Table 5.1 Blueprint Constraint by Reporting Category - Simulation 54
Table 5.2 Blueprint Constraint by Reporting Category - Engine Evaluation 55
Table 5.3 Item Exposure Rates - Simulation 56
Table 5.4 Item Exposure Rates - Engine Evaluation 57
Table 5.5 Mean Bias of the Ability Estimation (True - Estimated) 58
Table 5.6 Score Precision and Reliability – Simulation 59
Table 5.7 Score Precision and Reliability - Engine Evaluation 60
Table 5.8 SEM by Deciles -Simulation 62
Table 5.9 SEM by Deciles -Engine Evaluation 62
Table 6.1 Number of Students Included in the Psychometric Analyses 63
Table 6.2 Summary P-Values: Operational Items 64
Table 6.3 Summary P-Values: Field Test Items 65
Table 6.4 Summary Item-Total Correlations: Operational Items 66
Table 6.5 Summary Item-Total Correlations: Field Test Items 66
Table 6.6 Flagging Criteria for MC Items 67
Table 6.7 Flagging Criteria for non-MC Items 67
Table 6.8 Focal and Reference Groups for Gender- and Ethnicity-Based DIF 67
Table 6.9 LR DIF Categories 69
Table 6.10 MH DIF Categories for Dichotomous Items 71
Table 6.11 MH DIF Categories for Polytomous Items 71
Table 6.12 LR DIF Results: Operational Items 71
Table 6.13 LR UIDIF Results: Operational Items 73
Table 6.14 LR NUIDIF Results: Operational Items 74
Table 6.15 LR DIF Results: Field Test Items 76
Table 6.16 LR UIDIF Results: Field Test Items 77
Table 6.17 LR NUIDIF Results: Field Test Items 79
Table 6.18 MH DIF Results: Science Field Test Items 80
Table 6.19 Number of NSCAS Anchor Items used for MAP Growth Calibration 82
Table 6.20 Summary IRT Item Statistics: Operational Items 83

Table 6.21 Summary IRT Item Statistics: Field Test Items 83
Table 6.22 Spring 2021 NSCAS Science Field Test Form Design 84
Table 6.23 Number of Embedded MAP Growth Items in the Spring 2021 Phase 1 Pilot 86
Table 6.24 Number of Embedded MAP Growth Items used for Transformation 89
Table 6.25 Descriptive Statistics of RIT and Linked RIT Scores 89
Table 6.26 Mean SEM 90
Table 6.27 NSCAS vs. Linked RIT Achievement Level Distributions 90
Table 6.28 Score Range (LOSS and HOSS) and Assigned Score 93
Table 6.29 Cut Scores and Conversion of Theta to Scale Scores 93
Table 6.30 Score Range (LOSS and HOSS) for NSCAS scale score and linked RIT score . . 94
Table 7.1 Final Approved Cut Scores and Impact Data -ELA and Mathematics 98
Table 8.1 Number of Students Tested by Demographics - Grade 3 100
Table 8.2 Number of Students Tested by Demographics - Grade 4 101
Table 8.3 Number of Students Tested by Demographics - Grade 5 102
Table 8.4 Number of Students Tested by Demographics - Grade 6 103
Table 8.5 Number of Students Tested by Demographics - Grade 7 104
Table 8.6 Number of Students Tested by Demographics - Grade 8 105
Table 8.7 Number of Students Tested by Administration Mode 106
Table 8.8 Testing Time in Minutes - ELA 106
Table 8.9 Testing Time in Minutes - Mathematics 107
Table 8.10 Testing Time in Minutes - Science 108
Table 8.11 Achievement Level Distributions 109
Table 8.12 Scale Score Descriptive Statistics 109
Table 8.13 Raw Score Descriptive Statistics for Science Fixed Forms 110
Table 8.14 Correlation and Descriptive Statistics of NSCAS and MAP Growth Scores 111
Table 9.1 Marginal Reliability of Scale Scores–ELA 113
Table 9.2 Marginal Reliability of Scale Scores–Mathematics 113
Table 9.3 Marginal Reliability: Variance 113
Table 9.4 CSEMs at the Proficient Cut Scores 114
Table 9.5 Mean CSEMs by Deciles 114
Table 9.6 Classification Accuracy by Achievement Level and Reporting Category - ELA . . 115
Table 9.7 Classification Accuracy by Achievement Level and Reporting Category - Math . . 117
Table 9.8 Cronbach's Alpha for Science Fixed Forms 120
Table 9.9 Cronbach's Alpha by Demographics for Science Fixed Forms 121
Table 10.1 Sources of Validity Evidence for Each NSCAS Test Purpose 128
Table 10.2 Sources of Validity Evidence based on Test Content 128
Table 10.3 Sources of Validity Evidence based on Response Process 129
Table 10.4 Sources of Validity Evidence based on Internal Structure 130
Table 10.5 Sources of Validity Evidence based on Other Variables 130
Table 10.6 Interpretive Argument Claims, Evidence to Support the Essential Validity Elements 131
Table B.1 Summary P-Values by Item Type: Operational Items 143
Table B.2 Summary P-Values by Item Type: Field Test Items 144
Table C.1 Summary Item-Total Correlations by Item Type: Operational Items 147
Table C.2 Summary Item-Total Correlations by Item Type: Field Test Items 148
Table D.1 Achievement Level Distributions and Scale Score Descriptive Statistics by Demographics–ELA 151

Table D.2 Achievement Level Distributions and Scale Score Descriptive Statistics by Demographics–Mathematics 153
Table D.3 Raw Score Descriptive Statistics by Demographics–Science 155
Table E.1 Marginal Reliability by Demographics - ELA 160
Table E.2 Marginal Reliability by Demographics - Mathematics 162

List of Figures

Figure 1.1 Principled Test Design Process to Support Test Score Interpretations and Uses 15
Figure 2.1 Test Development Process 16
Figure 2.2 Range ALD Example: ELA Grade 3 21
Figure 3.1 CAP Student Login Screen 30
Figure 4.1 Reports Landing Page Example - District Assessment Coordinator 47
Figure 4.2 Matrix Example 49
Figure 5.1 Adaptive Engine Overview 50
Figure 5.2 Student-Specific Plan (SSP) Approach 51
Figure 6.1 Example Plot of ICC and Student Responses - Dichotomous Item 82
Figure 6.2 Example Plot of ICC and Student Responses - Polytomous Item 82
Figure A.1 Data Review Cheat Sheet 138
Figure F.1 Scatterplots for Scale Score CSEM 165

List of Abbreviations

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

ALD	achievement level descriptor
CAP	Comprehensive Assessment Platform
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
HOSS	highest obtainable scale score
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 Spring 2021 Nebraska Student-Centered Assessment System (NSCAS) Phase I Pilot 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

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 science assessment is being transitioned to the Nebraska College and Career Ready Standards for Science (NCCRS-S) and was administered as a full-scale field test. 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. Students taking the NSCAS tests are placed into one of the following achievement levels: Developing, On Track, or College and Career Readiness (CCR) Benchmark.

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 2021 NSCAS testing window was scheduled from March 22 to April 30, 2021. The tests were administered online via NWEA's Comprehensive Assessment Platform (CAP) test management system with paper-pencil versions available as an accommodation. Appropriate accommodations and universal features were provided, and test security was adhered to throughout the entire test administration process for both online and paper-pencil testing. User acceptance testing (UAT) was conducted prior to the operational administration to make sure the technology and item functionality were working properly.

Section 4: Scoring and Reporting

The online ELA and Mathematics assessments were administered adaptively via NWEA's constraint-based engine. 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. A visual interface, referred to as the NSCAS Matrix, allows users to select specific filters for schools and compare the data across schools in the state. No new data was added to the NSCAS Matrix for 2021.

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 were conducted prior to the Spring 2021 testing window to evaluate the constraint-based engine's item selection algorithm and estimation of student ability based on the blueprint. After the Spring 2021 testing window closed, a post-administration evaluation study was then conducted. Overall, the constraint-based engine performed as expected.

Section 6: Psychometric Analyses

The following post-administration analyses were conducted for the ELA, Mathematics, and Science assessments: classical item analyses, including item difficulty (p-value), item discrimination, and item suppression; differential item functioning (DIF) based on gender and ethnicity; item response theory (IRT) calibration; Science field testing and the common item linking between NSCAS and MAP Growth for ELA and Mathematics. The item-total correlation results appear out of bounds from traditional metrics, but this is because ELA and Mathematics were adaptive. Most items were categorized as DIF Category A (negligible DIF). Operational item parameter means increased by grade for ELA and Mathematics, as can be expected for vertical scales. Field test items were calibrated onto the NSCAS scale. The item characteristic curves (ICCs) created by the existing item parameters and the distribution of student responses were examined to determine which operational items would be used as anchor items. Based on the results from the common item linking between NSCAS and MAP Growth, NWEA recommended that IRT linked RIT with the Mean/Sigma transformation be used for the Nebraska through-year assessments, using items from the two reading reporting categories only for ELA (i.e., Reading Vocabulary and Reading Comprehension) and all items for mathematics.

Section 7: Standard Setting

No standard setting was held in 2020–2021. 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 will take place for the new NSCAS Science assessment following the first operational administration.

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. Among the students across grades, about 46% to 49% are eligible for free and reduced lunch (FRL), 7–16% have limited English proficiency (LEP) status, and 13–16% belong to at least one special education (SPED) category. The 2021 NSCAS assessments were administered online.

Most students completed the ELA test in 20–120 minutes, the Mathematics test in 20–100 minutes and the Science test in 10–60 minutes. For ELA, 46–55% of students are at Developing and 44–53% of students are at On Track or CCR Benchmark. For Mathematics, 52–54% of students are at Developing and 45–47% of students are at On Track or CCR Benchmark. The mean scale score increases with the grade for ELA and Mathematics, as expected. 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 2021 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 (84 or higher), which is typically considered the minimally acceptable level of reliability. The CSEM represents the degree of measurement error in scale score units and are conditioned on the ability of the student. When applied to an adaptive assessment, the CSEM will vary for the same scale score. It is therefore necessary to report averages. The overall CSEM is slightly higher for ELA than for Mathematics. Results also suggest that item pools have more items in the middle than at both ends and that more difficult items are needed for both ELA and Mathematics, which is consistent with reliability results. The classification accuracy results suggest that accurate classifications are being made for Nebraska students on the NSCAS assessments.

Section 10: Validity

Validating a test score interpretation is not a quantifiable property but an ongoing process, beginning at initial conceptualization of the construct and continuing throughout the entire assessment process. As the technical report progresses, it covers the different phases of the testing cycle and the procedures and processes applied in the NSCAS 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* (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).

1. Introduction

The purpose of this technical report is to summarize the design, development, administration, technical processes, and results of the Spring 2021 Nebraska Student-Centered Assessment System (NSCAS) Phase I Pilot assessments in English Language Arts (ELA) and Mathematics for Grades 3–8 and field test in Science for Grades 5 and 8 to support test users in evaluating the intended purposes, uses, and interpretations of the test scores. NSCAS was designed by the state of Nebraska with support from its vendor NWEA® to meet the requirements of the *Standards for Educational and Psychological Testing*(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 Spring 2021 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 not assigned achievement levels as this was a field test designed to calibrate the new items.

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 2019. The 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 for use on the new science assessment that will be aligned to the Nebraska College and Career Ready Standards for Science (NCCRS-S; NDE, 2017).

1.2 Background

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

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

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

NSCAS replaced the NeSA assessments beginning in 2017–2018. Spring 2021 was the third 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 it being scheduled in 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 must be provided between the NSCAS and MAP Growth scales. Whereas equipercentile linking was used to produce the Rasch Unit (RIT) scores for the Spring 2021 Phase 1 Pilot administration, NWEA conducted a common item linking study and recommended that IRT linked RIT with the Mean/Sigma transformation be used for the Nebraska through-year assessments (see Section 6.6).

1.3 Schedule of Major Events

Table 1.1 presents the major events that occurred for the 2021 NSCAS assessments, including the new science assessment. NDE involves educators throughout the development process

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

to produce customized items and provide an invaluable professional development opportunity, including item/task writing and review meetings and achievement level descriptor (ALD) reviews.

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

Event	Date(s)
Science Formative Task Development	June 14-17 & July 13-14, 2021
Fall 2020 Technical Advisory Committee (TAC) meeting	September 2, 2020 & November 18, 2020
Fall 2020 regional workshop	October 7, 2020
Test administration training	February 16-19, 2021
Technical Advisory Committee (TAC) meeting	April 19, 2021
Follow up Technical Advisory Committee (TAC) meeting	May 27,2021
Operational Testing window	March 22 – April 30, 2021
Make-up testing window	May 3 – May 7, 2021
District review preliminary data and submit updates	July 8-13, 2021
Data file available online	August 12, 2021
Delivery of online Individual Student Reports (ISRs)	September 8, 2021
Data Review with NDE (ELA, Mathematics)	September 2021
Data Review with NDE (Science)	October 2021

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 (Egan, Schneider, & Ferrara, 2012; Huff, Warner, & Schweid, 2016; Schneider & Johnson, 2018). This approach builds validity evidence into the design from the very beginning of the process, which is especially important when the assessments are intended to support interpretations regarding how student learning grows more sophisticated over time (Pellegrino, DiBello, & Goldman, 2016). The purposes of a test design centered in 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, Huff, Egan, Gaines, & Ferrara, 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 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 run simulations of the CAT engine in preparation for the Student Test Event (CAT) 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 version 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, Huff, & Reshetar, 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.

1.4.1 Intended Purposes and Uses of Test Results

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

1. To measure and report Nebraska students' depth of achievement regarding Nebraska's academic content standards
2. To report if student achievement is sufficient academic proficiency in ELA and Mathematics to be on track for achieving college readiness

3. To measure students' annual progress toward college and career readiness
4. To inform teachers how student thinking differs along different areas of the scale as represented by the RALDs 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:

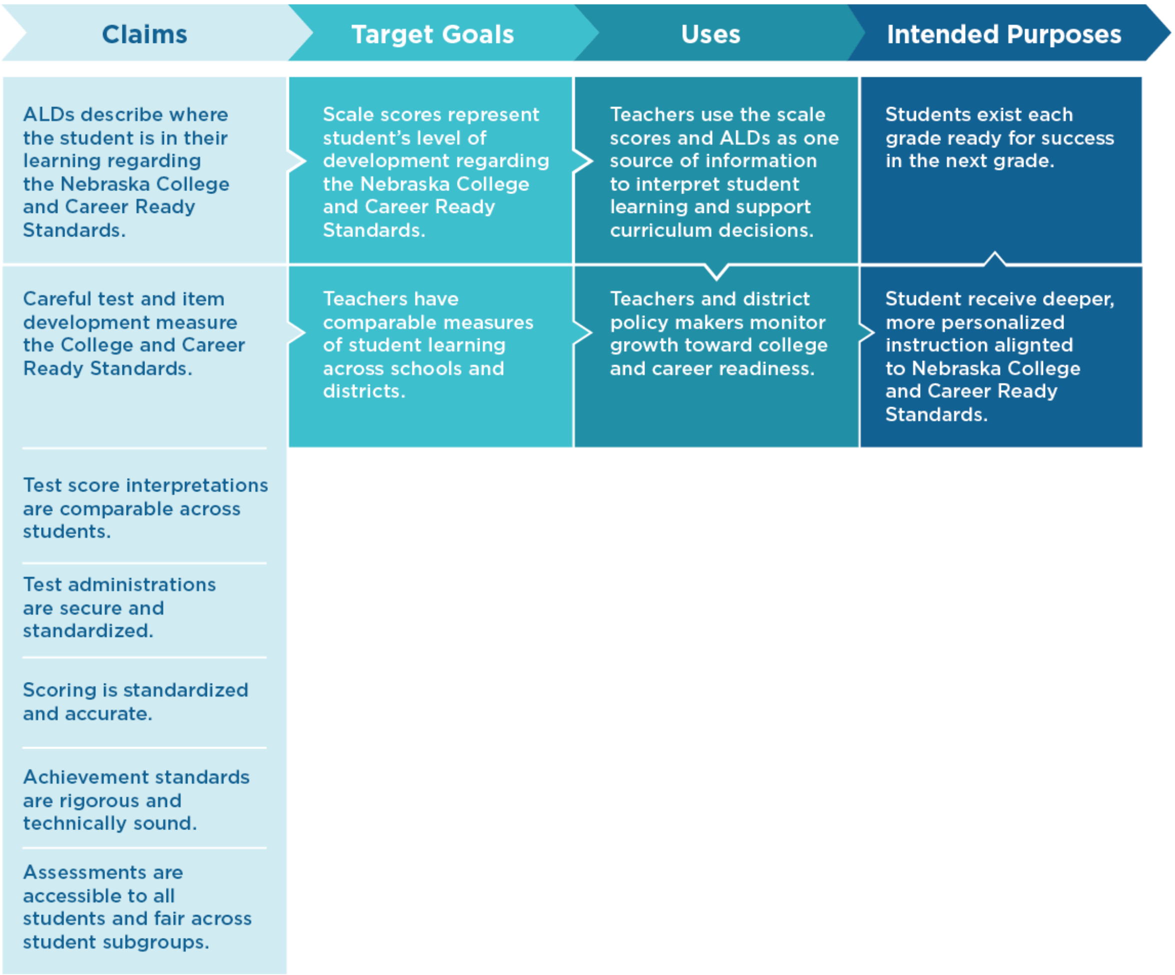
- To supplement teachers' observations and classroom assessment data and to improve the decisions teachers make about sequencing instructional goals, designing instructional materials, and selecting instructional approaches for groups and individuals
- To identify individuals for summer school and other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

1.4.2 Theory of Action

A theory of action is a tool that connects test users and their needs to decisions made during test design and development. In other words, it connects the design of the assessment, such as decisions about what evidence to collect and how to provide that evidence, to the claims that test score interpretation and use contribute to a positive solution to the broader problem for the test user. Figure 1.1 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

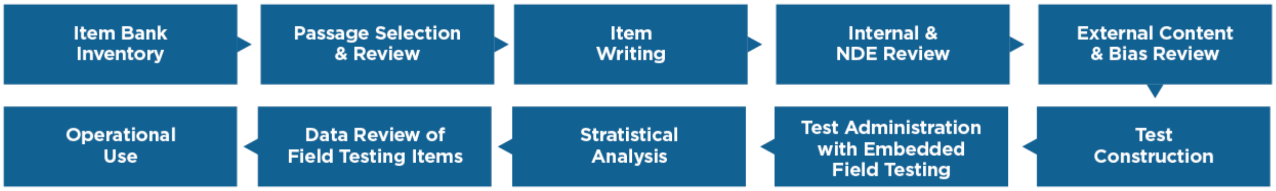


Claims	Target Goals	Uses	Intended Purposes
ALDs describe where the student is in their learning regarding the Nebraska College and Career Ready Standards.	Scale scores represent student's level of development regarding the Nebraska College and Career Ready Standards.	Teachers use the scale scores and ALDs as one source of information to interpret student learning and support curriculum decisions.	Students exist each grade ready for success in the next grade.
Careful test and item development measure the College and Career Ready Standards.	Teachers have comparable measures of student learning across schools and districts.	Teachers and district policy makers monitor growth toward college and career readiness.	Student receive deeper, more personalized instruction alignted to Nebraska College and Career Ready Standards.
Test score interpretations are comparable across students.			
Test administrations are secure and standardized.			
Scoring is standardized and accurate.			
Achievement standards are rigorous and technically sound.			
Assessments are accessible to all students and fair across student subgroups.			

2. Test Design and Development

This section describes the test design and development processes for the 2021 NSCAS Phase I Pilot 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 2020-2021 to have enough content available to populate field test slots in the Spring 2021 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.

Figure 2.1: Test Development Process


Item Bank Inventory
Passage Selection & Review
Item Writing
Internal & NDE Review
External Content & Bias Review
Test Construction
Test Administration with Embedded Field Testing
Stratistical Analysis
Data Review of Field Testing Items
Operational Use

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.

2.1 Test Design

Table 2.1 summarizes the versions of the NSCAS Phase I Pilot assessments available for 2021. For the Spring 2021 administration, students who required a paper form were exempt from the assessments. Table 2.2 presents the number of items and points possible. Science was administered as a full-scale field test in Spring 2021 (see Section 6.6).

Table 2.1: NSCAS Phase I Pilot Assessments in 2021

Content Area	Grade(s)	Online
ELA	3–8	Adaptive (35 total per grade, 23 OP + 7 FT +5 MAP)
Mathematics	3–8	Adaptive (35 total per grade, 23 OP + 7 FT +5 MAP)
Science	5	Fixed (58 FT for 20 Prompts)
	8	Fixed (51 FT for 17 Prompts)

[*] OP = operational. FT = field test. MAP = MAP Growth items embedded for linking.

Table 2.2: Number of Items and Points Per Test

		Online					
		Operational		FT/MAP*		Total	
Content Area	Grade(s)	#Items	#Points	#Items	#Points	#Items	#Points
ELA	3–8	23	27-28	12	12-15	35	39-43
Mathematics	3–8	23	27	12	12-13	35	39-41
Science	5	-	-	58	59	58	59
	8	-	-	51	59	51	59

* FT/MAP = field test/MAP Growth. Items in this slot are either FT or MAP items.

2.2 Academic Content Standards

As stated in Nebraska Revised Statute 79-760.01[2] 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 2021 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 2020-2021 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 currently in draft form and is available online at `https://cdn.education.ne.gov/`

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

[3] https://www.education.ne.gov/contentareastandards/

wp-content/uploads/2019/12/NE-Science-Draft-Public-Blueprint-V15.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 were 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.

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 (Egan et al., 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

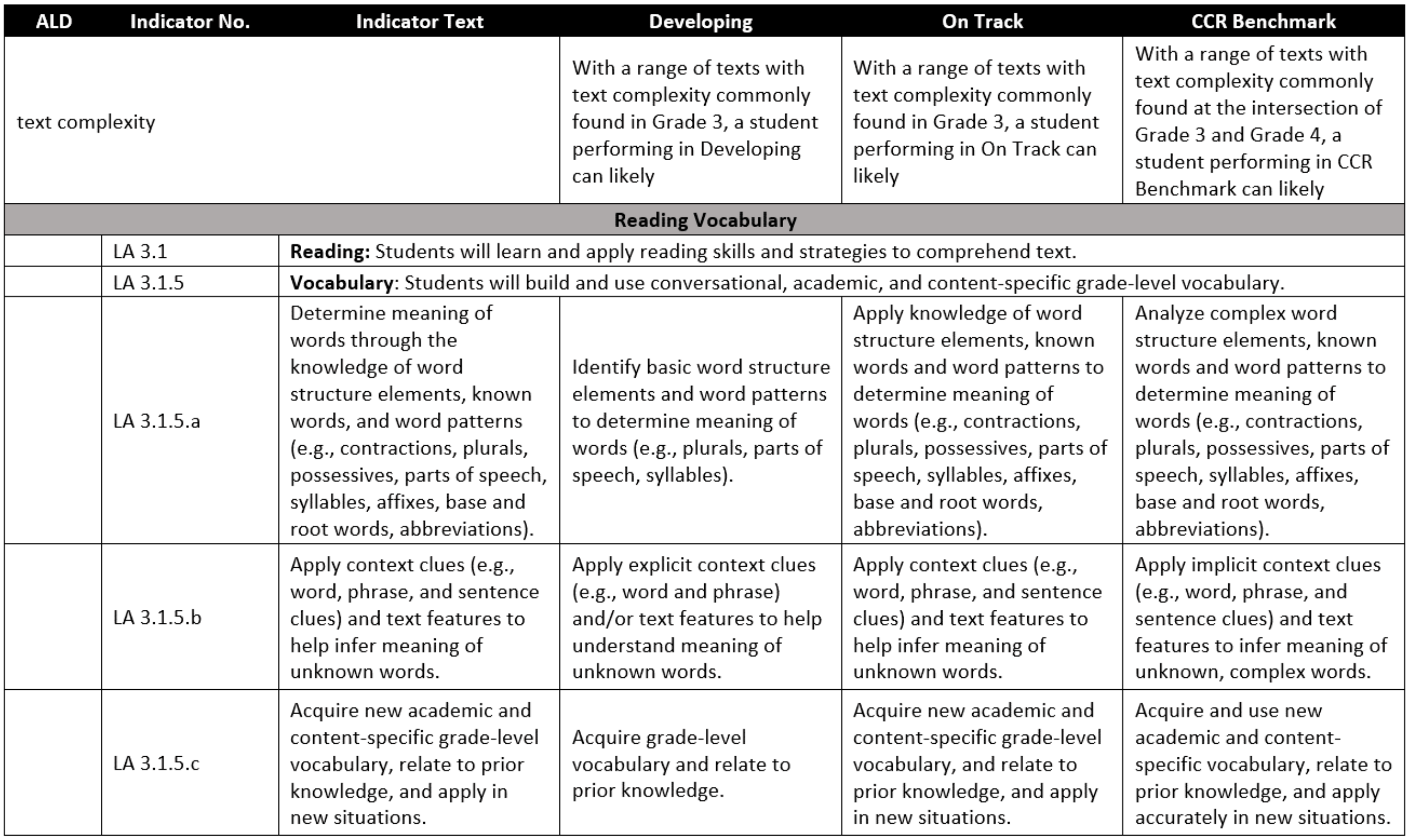
Range ALDs provide the intended content-based interpretations of what test scores within an achievement level represent and explicate observable evidence of achievement, demonstrating how the skill changes and becomes more sophisticated across achievement levels for each standard and achievement level 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. These science ALDs are still in draft form.

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 current version of the ELA ALDs available online at https://www.education.ne.gov/assessment/nscas-general-summative-assessment/nscas-english-language-arts-ela/.

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

Figure 2.2 presents example Range ALDs for ELA Grade 3. The progression descriptor (i.e., Developing, On Track, and CCR Benchmark) describes where a student is in their learning regarding the standard. Within a single expectation (e.g., LA 3.1.5.a) can be ranges of content- and thinking-skill difficulty that describe different stages of reasoning.

Figure 2.2: Range ALD Example: ELA Grade 3



ALD	Indicator No.	Indicator Text	Developing	On Track	CCR Benchmark
text complexity			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 draft science ALDs are available online at `https://cdn.education.ne.gov/wp-content/uploads/2020/02/NSCAS-Science-Summative-Achievement-Level-Descriptors-ALDs.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 (Council et al., 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 2020–2021 assessment administration was not required for math and ELA due to the shortened pilot. Items field tested in 2020-2021 had already been developed in prior years. Science development was put on hold at NDE's request to allow NDE to focus on formative task development.

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. Science also created pre-assessments for Grades 5 and 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 workshops to observe the development process. Information learned will be implemented in development for the 2020–2021 assessment administration.

2.8.1 Item Specifications

While there was no new item development for 2020-2021, 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 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 to August 8, 2019, based on Webb's DOK framework (Webb, 1997, 2002, 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

The adaptive tests were produced by selecting the item pools, building the test models that configured the engine and provided the constraints, running simulations, approving the results, and conducting user acceptance testing (UAT). The fixed forms were not created for Spring 2021 assessments.

2.13 Data Review of Field Tested Items

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 2021 field tested items occurred virtually in August/September 2021 between NDE and NWEA. Table 2.4 and Table 2.5 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 2.6 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 2021 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 2.4: 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	$\geq$4.25	Probability of getting an item correct may require extremely high ability
p-value	$<$ 0.2 or $>$ 0.9	very difficult item
p-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

Table 2.5: 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	$\geq$4.25	Probability of getting an item correct may require extremely high ability
step parameters	Step 1 $>$ Step 2	not a good separation of students into different stages of learning
Item-total correlation	$<$ 0.2	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	$<$ 0.2	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)

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

Table 2.6: Data Review Results for 2021 Field Test Items

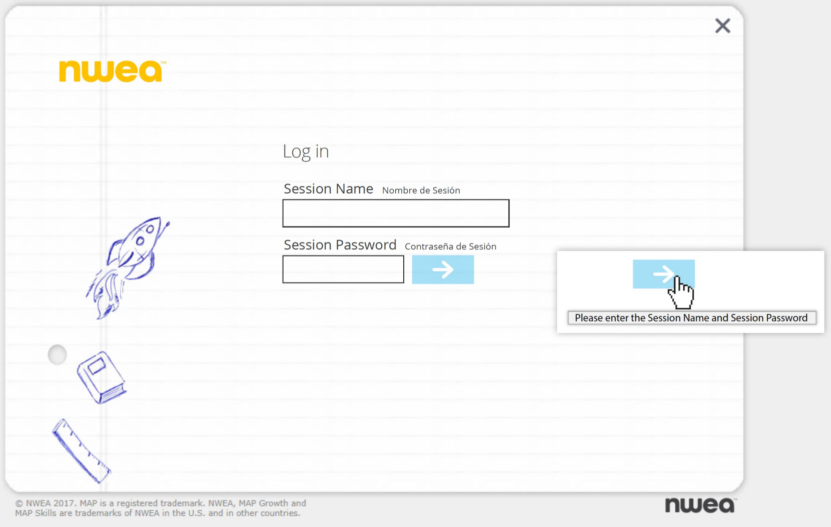
	#FT Items		Data Review				#Total
Grade	in the	#Administered	#Included	#Rejected	#Revise	#Accepted	Accepted
	Pool			/DNU	/ReFT		Items
ELA							
3	184	184	59	6	10	43	168
4	185	185	64	9	20	35	156
5	186	186	74	13	24	37	149
6	173	173	75	16	16	43	141
7	180	180	63	5	21	37	154
8	227	227	80	9	17	54	201
Mathematics							
3	231	231	52	7	11	34	213
4	150	150	24	3	5	16	142
5	182	182	34	5	4	25	173
6	231	231	47	14	7	26	210
7	226	226	76	16	18	42	192
8	157	157	50	9	14	27	134
Science							
5	58	58	7	1	5	1	52
8	51	51	16	0	12	4	39

3. Test Administration and Security

The Spring 2021 NSCAS testing window was from March 22 to April 30, 2021, and the make-up testing window was from May 3 to May 7, 2021. The tests were to be untimed and administered online via the NWEA Comprehensive Assessment Platform (CAP). 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 NWEA Comprehensive Assessment Platform (CAP) test management system, a roles-based platform that allowed users to roster students, set up test sessions, and administer the assessment. Figure 3.1 presents the student CAP login screen. CAP works with the NWEA secure lockdown testing browser to administer the assessments, which is required for NSCAS testing.

Figure 3.1: CAP Student Login Screen


nwea
Log in
Session Name Nombre de Sesión
Session Password Contraseña de Sesión
nwea

Please enter the Session Name and Session Password

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.

- Windows 7, 8.1, or 10
- Mac OS X® v.10.12 to 10.15
- iOS 11 to 12 and iPadOS 13.1.2 or higher recommended
- Google Chrome™ OS 65 or higher

3.1 User Roles and Responsibilities

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

Table 3.1: User Roles and Responsibilities

User	Roles and Responsibilities
District Assessment 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 Coordinators	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.
Proctors	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:

- 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 was a half-day, virtual workshop held across multiple regions of the state from October 8, 2020. Information on the spring administration including test sessions, accessibility, and student rostering was presented. The three test administration workshops in February 2021 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 availble online[5].

[5] https://www.youtube.com/watch?v=POh_P9Tcptshttps://cdn.education.ne.gov/wp-content/uploads/2020/10/Regional-Workshop-2020-2021-Publishing.pptx
https://vimeo.com/user84717829/review/515870657/f69712e944

Table 3.2: Test Administration Workshop Dates and Participation

Date	# Participants
Feb. 16, 2021	198
Feb. 17, 2021	112
Feb. 19, 2021	72

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.

3.4 Accommodations and Accessibility Features

Table 3.3 presents the accessibility supports available for the Spring 2021 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 (Nebraska Department of Education, 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	A student can use this feature to hear audio of the item content.
Non-Embedded Accommodations	
Paper-pencil Classroom Assessment*	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.

Table 3.3: Accommodations and Universal Features, cont.

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.

Table 3.3: Accommodations and Universal Features, cont.

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.
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.

*For the Spring 2021 administration, students who required a paper form were exempt from the assessments. However, for districts that wanted to gain information on the mastery of college and career-ready standards for students who need paper accommodation (English, Spanish translation, large print, or braille), NWEA provided electronic copies of an English and Spanish paper form in ELA and Math for districts to download and print. Additionally, districts could contact NWEA to have a large print or braille form shipped to the district by NWEA. Paper forms can be scored by the district but will not be returned to the vendor for scoring.

3.5 User Acceptance Testing (UAT)

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

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

- Item-specific features (e.g., ruler, protractor)
- Accessibility features (e.g., TTS)
- New features/enhancements

From February 4-10, 2021, 29 testers participated in UAT in 2021. Each were assigned 1-9 tests. Each were assigned 1–9 tests. 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. Daily 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 State Solution and Data Operations and Operational Content and Psychometrics staff review the data produced from the UAT to ensure it conforms to expectations for completed tests and tests assigned NTCs.

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. Accommodations may have required paper-pencil testing, those students were exempt from Spring 2021 testing.
- Students could be tested with the NSCAS Alternate assessment if they qualified for these assessments. Only students with the most significant cognitive disabilities (typically less than 1% of students) could take these tests. The NSCAS Alternate test was distributed and administered by DRC.

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

3.6.1 Paper-Pencil Participation Criteria

Students participating in the paper-pencil administration, those exempt from testing in Spring 2021, 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 (Nebraska Department of Education, 2018).
- Districts must be aware of the difference between linguistic supports (accommodations for ELLs) and modifications.
- For students learning the English language, linguistic supports are changes to testing procedures, testing materials, or the testing situation that allow the students meaningful participation in the assessment. Effective linguistic supports for ELL students address their unique linguistic and socio-cultural needs. Linguistic supports for ELL students may be determined appropriate without prior use during instruction throughout the year.
- Modifications are adjustments or changes in the test or testing process that change the test expectation, grade level, construct, or content being measured. Modifications are not acceptable in the NSCAS assessments.

3.6.3 Participation of Recently Arrived Limited English Proficient 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 12, 2020. 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 2021 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

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

3.7.2.2 Monitoring for Potential Test Security Violations. Caveon data forensics analyses were performed to discover anomalous results that may be indicative of potential test security violations. These analyses provided information regarding where and when test security incidents may have occurred, by whom, and their effects on the testing program. Table 3.4. summarizes the statistical analyses performed. The data forensics analyses were conducted to identify potential test security violations relating to individual students, schools, and items on the exams.

Table 3.4: Statistical Analysis and Potential Incidents

Statistical Analysis	Potential Incidents
Response Times	Responding to items inconsistently regarding time or supplying answers in unusually short lengths of time can indicate pre-knowledge of test content or unsanctioned aid given to students while taking the test (i.e., test coaching).
Person-fit (Aberrance) Statistics	When students respond in a manner that is inconsistent with the student population, supportive evidence of pre-knowledge or test coaching may be present.
Item Performance Changes	Performance shifts, indicating the items have become easier during the test administration window, provide evidence that the item might have been disclosed to the students.
Exposed Differences	Item exposure (i.e., administrations to individual students) levels vary in CAT pools (i.e., ELA and Mathematics). When student performance is higher on frequently exposed items than on the other items, there is a possibility that some or a few students had access to some of the test content prior to the exam.
M4 Similarity	Exams that use fixed forms (i.e., Science) were analyzed for excessive agreement between students. These statistics can identify where answer copying by students, sharing of test responses between students, coaching, pre-knowledge, or large-scale collusion may have occurred.
Identical Test	When students receive the same items (i.e., because they were administered the same form as in the Science exam), it is possible they may have identical responses to all of the items. This is more likely if they use a disclosed answer key. When this happens, students often will receive very high scores on the exam.
Perfect Test	A concentration of perfect scores at a school, which are very unusual, may indicate the presence of a test security incident.
Synchronicity	For fixed-form tests (i.e., Science), when students answered items at or near the same time of day, there is a possibility that they were guided or paced through the exam.

As provided in the data forensics report from Caveon (Drane, Torton, & Scott, 2021), data for 302,446 test instances administered at 812 schools in 245 districts were analyzed. The most significant findings are as follows:

- For ELA and Mathematics, three schools had score gains associated with detections by the Score Aberrance Statistic, which is designed to detect discrepancies between performance estimates as measured by the score and the ability level. However, the rates were not anomalous, and these findings are not strong enough to infer that the schools were involved with a security violation. All other anomalies detected for ELA and Mathematics tests were associated with decreased performance.
- For Science, high detection rates by the Synchronicity Statistic, which was designed to detect when test items are taken at the same time by multiple test-takers[6], accompanied by increased performance for detected test instances, may be evidence of a security violation for a few schools. All other anomalies detected for Science tests were associated with decreased performance.
- Mathematics Grade 6 continues to exhibit the most item performance changes of any ELA and Mathematics subject-grade group. However, the detected performance changes do not appear to be the result of a security violation.

[6] Synchronicity analysis was conducted for only Science 5 and 8 because those were the only exams with fixed forms; ELA and Mathematics forms were administered using CAT.

With the possible exception of the anomalies described above, the exams appear to have been administered securely.

3.8 Partner Support

The NWEA Partner Support Services team provided implementation and technical support throughout the 2020-2021 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 3.5 Table 3.6 presents the number of cases presented to the Partner Support team by case type for the entire 2020-2021 school year from July 2020 to June 2021 for the NSCAS tests. More than half of the cases were related to testing (i.e., administration questions).

Table 3.5: 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 3.6: Test Administration Workshop Dates and Participation

Case Type	# Cases	% of Total Cases
Student Mobility	1	0.2
Reports	8	1.6
Navigation	61	12.22
Setup and Management	122	24.44
Other	62	12.42
Testing	2454	49.09
Total	499	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.

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. Due to science being a full-scale field test, reporting is only available for ELA and Mathematics.

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 after 10 operational items from the simulation data and the finding of a small number of 2017 students who attempted less than 10 items.

Students who took the adaptive assessment (i.e., ELA and Mathematics online adaptive forms) received straight MLE scoring (i.e., regular MLE scoring with no penalty) regardless of the test completion status.

For the Spring 2021 administration, no scores were produced for fixed forms. Science was a field test. Students who would test with paper-pencil or in Spanish were exempt from the assessments. However, for districts that wanted to gain information on the mastery of college and career-ready standards for students who need paper accommodation (English, Spanish translation, large print, or braille), NWEA provided electronic copies of an English and Spanish paper form in ELA and Math for districts to download and print. Additionally, districts could contact NWEA to have a large print or braille form shipped to the district by NWEA. Paper forms can be scored by the district but will not be returned to the vendor for scoring.

4.2 Paper-Pencil Scoring

Students requiring a paper assessment were exempt from taking the Spring 2021 NSCAS assessments, therefore there were no answer sheets to scan.

4.3 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 Phase I Pilot reports also provide linked RIT scores, which were converted from the NSCAS scale scores.

Table 4.2 presents score range for both scores. Science was a field test and no score data were produced.

Table 4.2: Score Range (LOSS and HOSS) for NSCAS scale score and linked RIT score

	NSCAS Scale Score			Linked RIT Score		
Grade	**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

* 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 4.3.

Table 4.3: Achievement Level Descriptions

Achievement Level	Description
Developing	Developing learners do not yet demonstrate proficiency in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards. These results provide evidence that the student may need additional support for academic success at the next grade level.
On Track	On Track learners demonstrate proficiency in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards. These results provide evidence that the student will likely be ready for academic success at the next grade level.
CCR Benchmark	CCR Benchmark learners demonstrate advanced proficiency in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards. These results provide evidence that the student will likely be ready for academic success at the next grade level.

4.4 Report Summary

The following reports were prepared for the 2021 NSCAS test administration. Examples of the reports can be found in the Interpretive Guide[7].

- Preliminary Student Data Files
- Final Student Data Files
- Individual Student Report (ISR) English
- Individual Student Report (ISR) Spanish
- School Roster

ISRs show a student's performance on the NSCAS Phase I Pilot tests. Content areas are combined across schools and districts to produce a single ISR report. Where more than one test exists for a single student within a content area the ISR reports scored tests over non-scored tests events and valid tests over any with NTCs applied. Some ISRs will be routed to their new fall enrollment school while others will be routed to the latest District of Accountability. 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 4.4. If a student has an NTC of INV, PAR, SAE, 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 4.4: Non-Tested Code

Code	Translation	Description	Score Reporting
ALT	Alternate Assessment	Student took the NSCAS Alternate assessment and is not included in results from this testing vendor	• No Scale Score provided for a test with this code • Score Suppressed • NTC only
COV	COVID-19 Waiver	Student did not test because of an ongoing and continued concern about exposure to COVID-19	• No Scale Score provided for a test with this code • Score Suppressed • NTC only
EMW	Emergency Medical Waiver	Student was not tested because of an approved emergency medical waiver	• No Scale Score provided for a test with this code • Score Suppressed • NTC only
EXP	Exception	Due to testing irregularities, the assessment was not scored	• Score not included in any reports or calculations
INV	Invalid	Student's assessment was invalidated; such as security breach	• Score as LOSS • NTC only
NLE	No Longer Enrolled	Student was not enrolled in the district/school during testing window(s)	• No Scale Score provided for a test with this code • NTC only
OTH	Other	Student's score was removed from performance for reasons not covered by other descriptions	• Score Suppressed • NTC only
PAR	Parental Refusal	Student was not tested because of a written request from parent or guardian	• Score as LOSS • NTC only
RMV	Removed	Student was removed from the file for reasons not otherwise covered	• Score Suppressed • Suppress from all reports or calculations

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

Table 4.4: Non-Tested Code, cont.

SAE	Student Absent for Entire Testing Window	Student was absent from School for the entire testing window(s)	• Score as LOSS • NTC only
STR	Student Refusal	Student was not tested due to student refusal to participate	• Score as LOSS • NTC only
TXP	Tested at External Program	Student is attending an external program and test scores should be transferred to district/school of accountability	• Score not included in any reports or calculations
UTT	District Unable to Test Student	District unable to test student during the testing window and none of the other NTCs are applicable	• Score as LOSS • NTC only

The School Roster report lists students required to take the NSCAS tests and presented a report of their performance. The size of this document depends on the class size.

4.5 Report Process

4.5.1 Online Reports

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

4.5.2 Printed ISRs

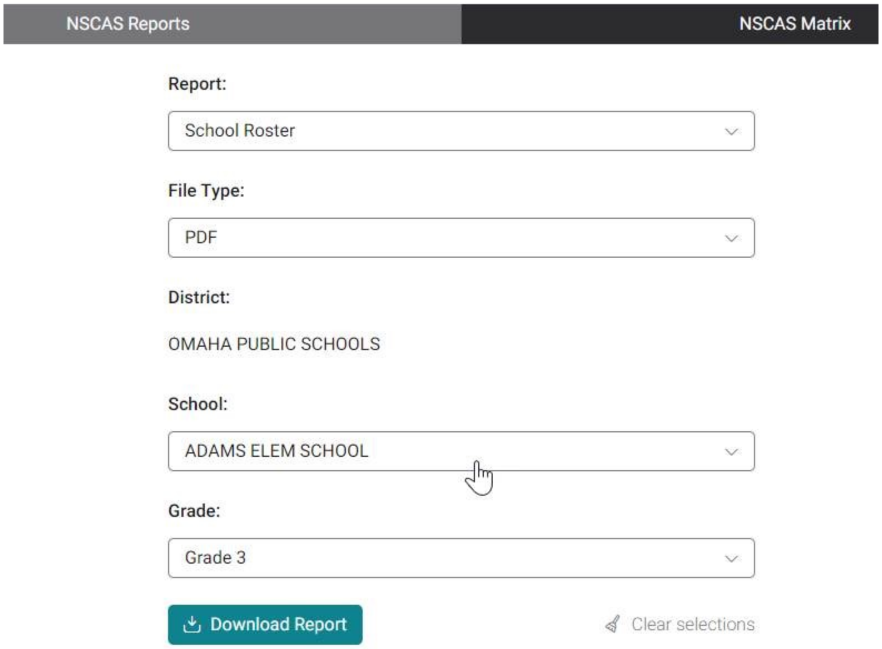
ISRs were only available in electronic format for Spring 2021.

4.5.3 Report Verification

The NSCAS report quality assurance (QA) process consisted of validating the data and reports using the scoring and reporting specifications, mockups, layouts, and scale score and cut information. The first step was to validate that the data were accurate and the appropriate rules were applied. PDF reports were then generated and validated. Specific schools were identified to validate the scoring and reporting rules. After the reports passed quality control, they were loaded to a staging environment to verify the Reports Page, user interface functionality, and user access.
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.

Figure 4.1: Reports Landing Page Example - District Assessment Coordinator



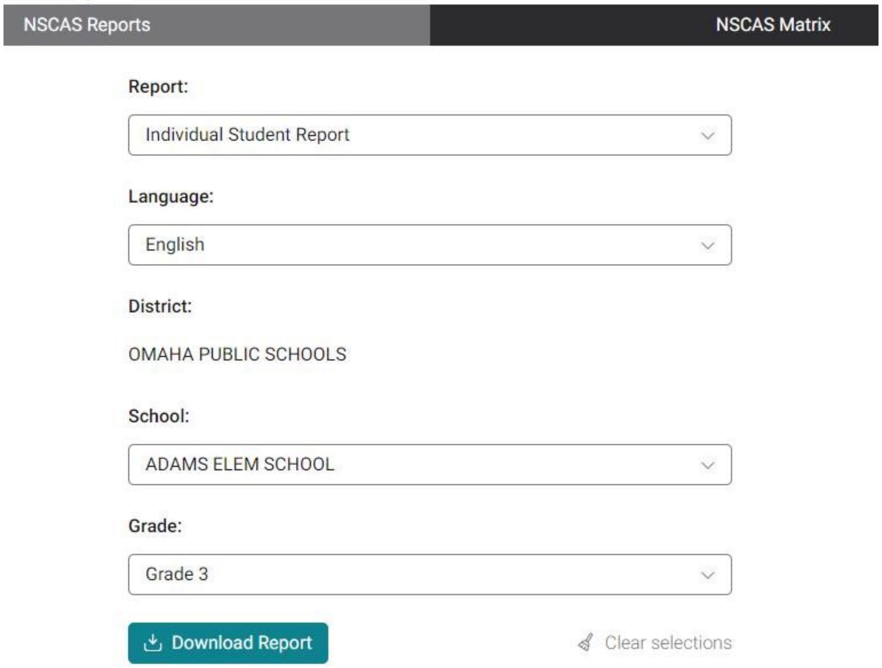
Report: School Roster

File Type: PDF

District: OMAHA PUBLIC SCHOOLS

School: ADAMS ELEM SCHOOL

Grade: Grade 3



Report: Individual Student Report

Language: English

District: OMAHA PUBLIC SCHOOLS

School: ADAMS ELEM SCHOOL

Grade: Grade 3

The following report sections were checked during the QA process:

- Formatting
- Static text (text that does not change)
- Dynamic text (text that changes)
- Student data (demographic information)
- Score-related data (scale scores, achievement levels)

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

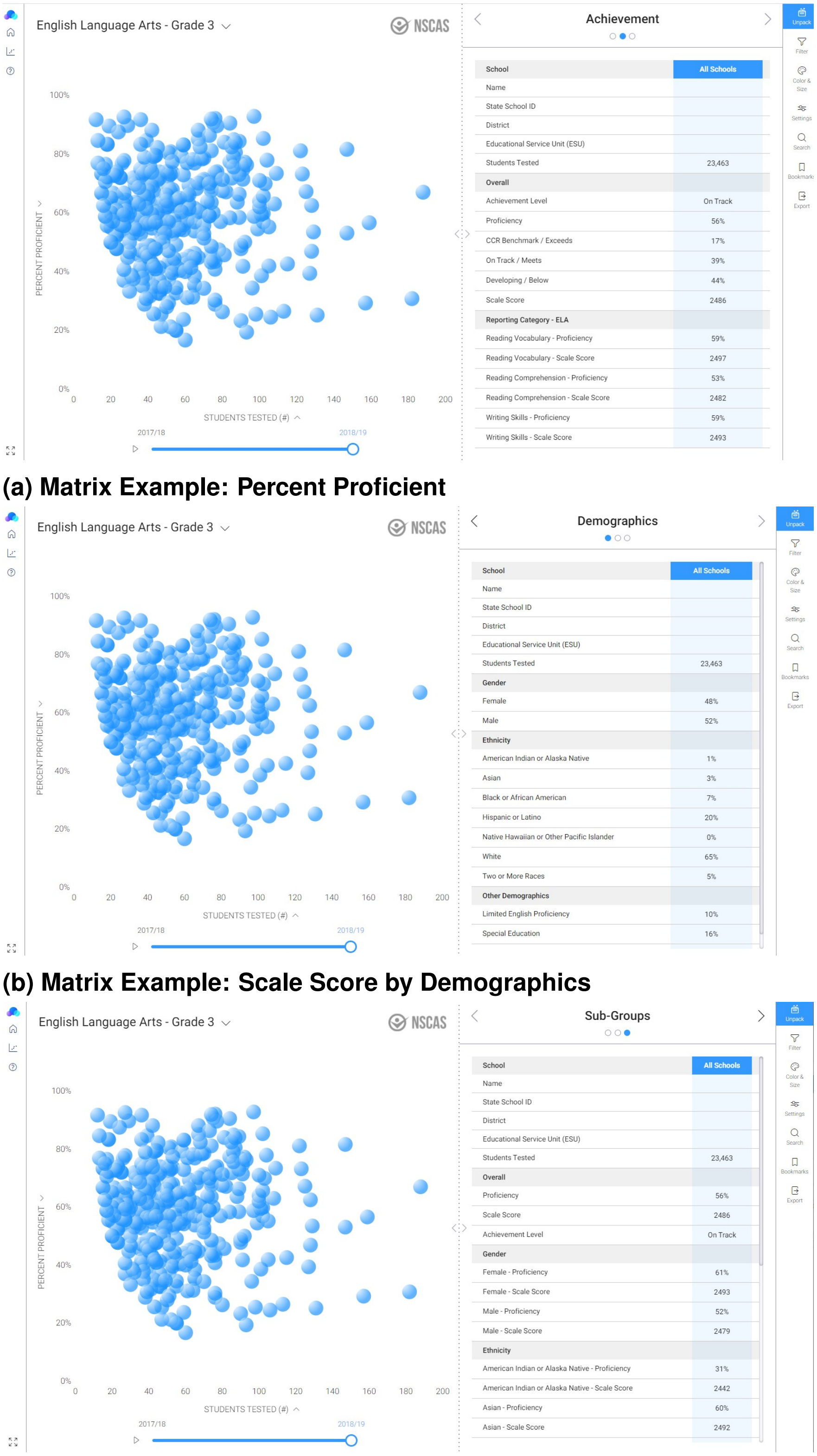
4.6 NSCAS Matrix

Education Strategy Consulting (ESC) is maintaining the NSCAS Matrix with historical info for reference. Users still have access to this tool; however, there was no new data added to the NSCAS Matrix for 2021.

NWEA used ESC's tools to view web-based visualizations for the NSCAS assessments, including combinations of aggregate and disaggregate information of results by demographics and other filtering options. This visual interface, referred to as the NSCAS Matrix, allows users to select specific filters for schools and compare the data across schools in the state. Users can interact with and explore many different levels of information to answer targeted questions about their district, school, or state. The main feature of this tool is an interactive scatterplot designed to display longitudinal data, as shown in Figure 4.2. The X and Y axes are modifiable. Users can export data in excel or csv from available variables within the export function. This feature allows for easy access to high-quality data that has gone through rigorous auditing. Users can then explore and sort data to meet their individual needs. Suppression rules are applied to the data for all users. For example, all data is suppressed for a school if the number of tested students was less than 10.

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

Figure 4.2: Matrix Example



English Language Arts - Grade 3

Achievement

School	All Schools
Name	
State School ID	
District	
Educational Service Unit (ESU)	
Students Tested	23,463
Overall	
Achievement Level	On Track
Proficiency	56%
CCR Benchmark / Exceeds	17%
On Track / Meets	39%
Developing / Below	44%
Scale Score	2486
Reporting Category - ELA	
Reading Vocabulary - Proficiency	59%
Reading Vocabulary - Scale Score	2497
Reading Comprehension - Proficiency	53%
Reading Comprehension - Scale Score	2482
Writing Skills - Proficiency	59%
Writing Skills - Scale Score	2493

(a) Matrix Example: Percent Proficient

English Language Arts - Grade 3

Demographics

School	All Schools
Name	
State School ID	
District	
Educational Service Unit (ESU)	
Students Tested	23,463
Gender	
Female	48%
Male	52%
Ethnicity	
American Indian or Alaska Native	1%
Asian	3%
Black or African American	7%
Hispanic or Latino	20%
Native Hawaiian or Other Pacific Islander	0%
White	65%
Two or More Races	5%
Other Demographics	
Limited English Proficiency	10%
Special Education	16%

(b) Matrix Example: Scale Score by Demographics

English Language Arts - Grade 3

Sub-Groups

School	All Schools
Name	
State School ID	
District	
Educational Service Unit (ESU)	
Students Tested	23,463
Overall	
Proficiency	56%
Scale Score	2486
Achievement Level	On Track
Gender	
Female - Proficiency	61%
Female - Scale Score	2493
Male - Proficiency	52%
Male - Scale Score	2479
Ethnicity	
American Indian or Alaska Native - Proficiency	31%
American Indian or Alaska Native - Scale Score	2442
Asian - Proficiency	60%
Asian - Scale Score	2492

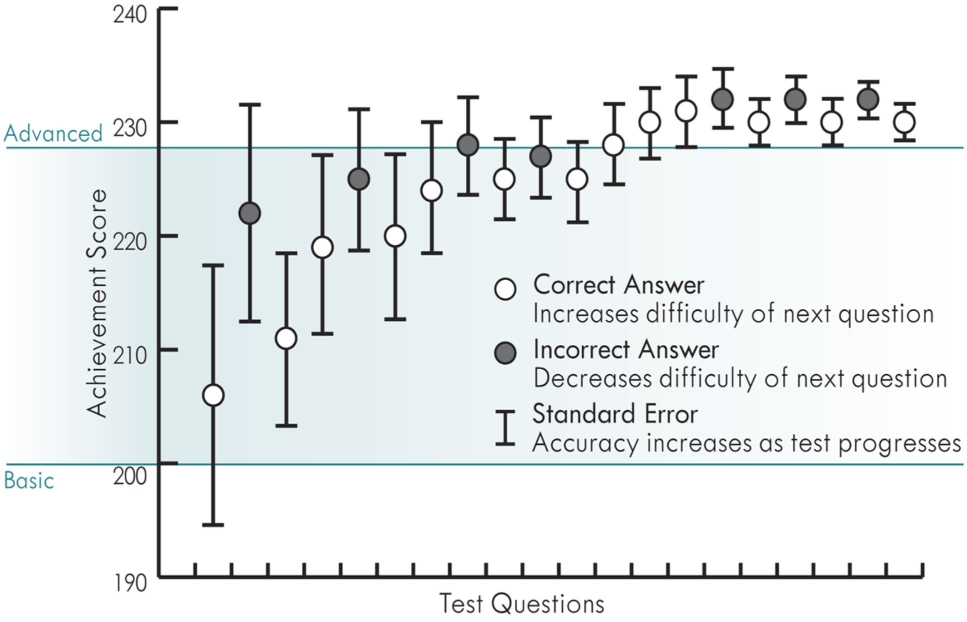
(c) Matrix Example: Scale Score by Sub-Groups

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



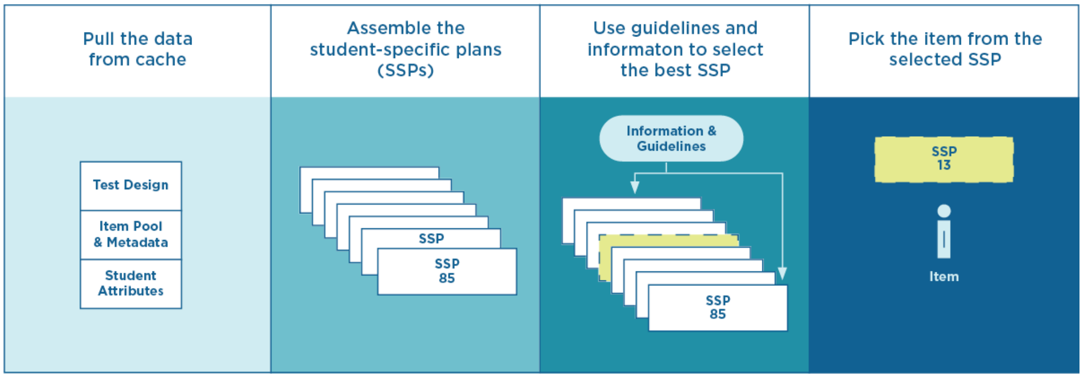
Items were selected based on item difficulty. The goal of the constraint-based engine's item selection was to provide a test that meets ”must-have” constraints and ”nice-to-have” guidelines.

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 5.2. 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, 2020a).

As compared to the previous simulation reports provided by NWEA, this simulation study was based on the following test design updates:

1. The operational test is shorter in Spring 2021 (i.e., 23 operational items vs. 41 operational items previously).
2. Indicator level guidelines were removed for this shorter version because they could not be met within 23 items. Strand level guidelines were maintained.
3. Item exposure was controlled by assigning a weight to an item based on the number of times the item is seen by students. This feature is an update to the test model that improves operational item pool utilization.
4. Pseudo-random assignment of field test items was implemented. In previous administrations, items delivered in vertical linking and field test sections were alternated between students. Because this test does not have a vertical linking set, field test items were pseudo-randomly assigned to students.

Figure 5.2: Student-Specific Plan (SSP) Approach


Pull the data from cache
Assemble the student-specific plans (SSPs)
Use guidelines and informaton to select the best SSP
Pick the item from the selected SSP
Test Design
Item Pool & Metadata
Student Attributes
SSP
SSP 85
Information & Guidelines
SSP 85
SSP 13
Item

5.2 Engine Simulations and Evaluation

Pre-administration engine simulations and a post-administration engine evaluation studies are important evidence, along with post-administration analyses, for confirming interpretation and test score use arguments regarding student proficiency with the state standards. Pre-administration simulations were conducted prior to the 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, 2021b).

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, 2021c).

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 each statistic are as follows (CRESST, 2015):

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

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

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 $g'(\hat{\theta}_i)^2$ of the above equation is used as follows:

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

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

$$Z = \frac{bias}{\sqrt{var(bias)}} \tag{5.4}$$

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:

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

where $se(\hat{\theta}_i)^2$ is the standard error of the estimated θ for individual i. The CBE provided the estimated θ and the standard error. The standard error is calculated by summing the item information at the current estimate for all items answered and taking the inverse of the square root of that total. This is applied for each scale individually, as shown below (NWEA, 2020a, p. 42).

$$se(\hat{\theta}_i) = \left(\sum_{i=1} I_i(\theta) \right)^{-1/2} \tag{5.6}$$

To determine the number of students falling outside the 95$\%$ and 99$\%$ confidence interval coverage, a t-test was performed as follows:

$$t = \frac{\theta_1 - \hat{\theta}_i}{se(\hat{\theta}_i)} \tag{5.7}$$

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 confidence interval 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 students 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:

$$MarginalReliability = \frac{var(\hat{\theta}) - \sigma_{\varepsilon}^{2}}{var(\hat{\theta})} \tag{5.8}$$

where σ_{ε} is defined as the expectation (E) of the item response information function:

$$\sigma_{\varepsilon} = E[I(\theta)]^{-1} = \int_{-\infty}^{\infty} [I(\theta)]^{-1} f(\theta) d\theta \tag{5.9}$$

5.2.2 Blueprint Constraint Accuracy

Table 5.1 and Table 5.2 present the blueprint constraint results at the reporting category level for the pre-administration simulation study and the post-administration evaluation, respectively. The findings from the engine evaluation study appeared similar to those in the simulation study, as expected. For both studies and content areas, the number of items and points at the reporting category level resulted in a 100$\%$ match for all grades based on the blueprint.

Table 5.1: Blueprint Constraint by Reporting Category - Simulation

		#Items			#Points		
Grade	**Reporting Category**	**Min.**	**Max.**	**%Match**	**Min.**	**Max.**	**%Match**
ELA							
3	Reading Vocabulary	4	4	100.0	4	5	100.0
	Reading Comprehension	13	13	100.0	15	15	100.0
	Writing Skills	6	6	100.0	8	8	100.0
4	Reading Vocabulary	4	4	100.0	4	5	100.0
	Reading Comprehension	13	13	100.0	15	15	100.0
	Writing Skills	6	6	100.0	8	8	100.0
5	Reading Vocabulary	4	4	100.0	4	5	100.0
	Reading Comprehension	13	13	100.0	15	15	100.0
	Writing Skills	6	6	100.0	8	8	100.0
6	Reading Vocabulary	4	4	100.0	4	5	100.0
	Reading Comprehension	13	13	100.0	15	15	100.0
	Writing Skills	6	6	100.0	8	8	100.0
7	Reading Vocabulary	4	4	100.0	4	5	100.0
	Reading Comprehension	12	12	100.0	14	14	100.0
	Writing Skills	7	7	100.0	9	9	100.0
8	Reading Vocabulary	4	4	100.0	4	5	100.0
	Reading Comprehension	13	13	100.0	15	15	100.0
	Writing Skills	6	6	100.0	8	8	100.0
Mathematics							
3	Number	9	9	100.0	10	10	100.0
	Algebra	4	4	100.0	5	5	100.0
	Geometry	6	6	100.0	7	7	100.0
	Data	4	4	100.0	5	5	100.0
4	Number	9	9	100.0	10	10	100.0
	Algebra	5	5	100.0	6	6	100.0
	Geometry	5	5	100.0	6	6	100.0
	Data	4	4	100.0	5	5	100.0
5	Number	9	9	100.0	10	10	100.0
	Algebra	5	5	100.0	6	6	100.0
	Geometry	5	5	100.0	6	6	100.0
	Data	4	4	100.0	5	5	100.0
6	Number	6	6	100.0	7	7	100.0
	Algebra	9	9	100.0	10	10	100.0
	Geometry	4	4	100.0	5	5	100.0
	Data	4	4	100.0	5	5	100.0
7	Number	5	5	100.0	6	6	100.0
	Algebra	9	9	100.0	10	10	100.0
	Geometry	5	5	100.0	6	6	100.0
	Data	4	4	100.0	5	5	100.0
8	Number	6	6	100.0	7	7	100.0
	Algebra	6	6	100.0	7	7	100.0
	Geometry	7	7	100.0	8	8	100.0
	Data	4	4	100.0	5	5	100.0

Table 5.2: Blueprint Constraint by Reporting Category - Engine Evaluation

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

5.2.3 Item Exposure Rates

Table 5.3 and Table 5.4 present the item exposure rates from the pre-administration engine simulation study and post-administration engine evaluation study, respectively. Because students receive 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 Table 5.3 and Table 5.4, ”Total” is the total number of items in the operational item pool. ”Unused” shows the number and percentage of items that were never administered to students.

The patterns of exposure rate for the engine evaluation study are very similar between two studies. For both studies, , most items across grades and content areas had a 0 - 20% exposure rate. Compared to the previous years’ results, the unused percentage of adaptive items decreased a lot, improving the item pool usage.

Table 5.3: Item Exposure Rates - Simulation

					Exposure Rate											
		#Items			0-20%		21-40%		41-60%		61-80%		81-99%		100%	
Grade	Total	Used	Unused	Unused %	N	%	N	%	N	%	N	%	N	%	N	%
ELA																
3	590	589	1	0.17	583	98.98	6	1.02	0	0.00	0	0.00	0	0.00	0	0.00
4	579	578	1	0.17	578	100.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
5	513	508	5	0.97	500	98.43	7	1.38	0	0.00	1	0.20	0	0.00	0	0.00
6	520	519	1	0.19	513	98.84	6	1.16	0	0.00	0	0.00	0	0.00	0	0.00
7	486	482	4	0.82	472	97.93	8	1.66	2	0.41	0	0.00	0	0.00	0	0.00
8	557	553	4	0.72	547	98.92	4	0.72	2	0.36	0	0.00	0	0.00	0	0.00
Mathematics																
3	541	540	1	0.18	537	99.44	3	0.56	0	0.00	0	0.00	0	0.00	0	0.00
4	418	417	1	0.24	417	100.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
5	432	431	1	0.23	430	99.77	1	0.23	0	0.00	0	0.00	0	0.00	0	0.00
6	538	537	1	0.19	537	100.0	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
7	465	457	8	1.72	452	98.91	5	1.09	0	0.00	0	0.00	0	0.00	0	0.00
8	435	435	0	0.00	431	99.08	4	0.92	0	0.00	0	0.00	0	0.00	0	0.00

Table 5.4: Item Exposure Rates - Engine Evaluation

					Exposure Rate											
		#Items			0-20%		21-40%		41-60%		61-80%		81-99%		100%	
Grade	Total	Used	Unused	Unused %	N	%	N	%	N	%	N	%	N	%	N	%
ELA																
3	590	590	0	0.00	584	98.98	6	1.02	0	0.00	0	0.00	0	0.00	0	0.00
4	579	579	0	0.00	579	100.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
5	513	508	5	0.97	500	98.43	6	1.18	1	0.20	1	0.20	0	0.00	0	0.00
6	520	518	2	0.38	511	98.65	7	1.35	0	0.00	0	0.00	0	0.00	0	0.00
7	486	478	8	1.65	468	97.91	8	1.67	2	0.42	0	0.00	0	0.00	0	0.00
8	557	553	4	0.72	547	98.92	3	0.54	3	0.54	0	0.00	0	0.00	0	0.00
Mathematics																
3	541	540	1	0.18	538	99.63	2	0.37	0	0.00	0	0.00	0	0.00	0	0.00
4	418	418	0	0.00	418	100.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
5	432	432	0	0.00	431	99.77	1	0.23	0	0.00	0	0.00	0	0.00	0	0.00
6	538	537	1	0.19	537	100.0	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
7	465	457	8	1.72	452	98.91	5	1.09	0	0.00	0	0.00	0	0.00	0	0.00
8	435	435	0	0.00	431	99.08	4	0.92	0	0.00	0	0.00	0	0.00	0	0.00

5.2.4 Score Precision and Reliability

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

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

- Precision of ability estimation (how well the engine recovered students' true ability based on the item pool):
  - Bias: Shows the difference between true and final estimated theta.
  - P-value for the z-test: Determines if the difference of bias between the true and final estimated theta is statistically different. If the p-value is larger than 0.05, there is no statistical difference of bias between the true and final estimated theta.
  - Mean Standard Error (MSE): Provides the average squared bias across the population of examinees. 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 the respective confidence interval in terms of theta.

Table 5.5 presents the results of the precision ability estimation from the pre-administration simulations. The mean biases across all students are small, ranging from 0.00 to 0.01 for the overall scores of both ELA and Mathematics. The p-value supports the null-hypothesis that there is not a significant difference between the simulated students' true and final estimated thetas. The MSE is also relatively small, showing that the CBE typically recovered a value near the student's true theta.

Table 5.5: Mean Bias of the Ability Estimation (True - Estimated) – Simulation

		Bias		P-Value for		95%	99%
Grade	Reporting Category	Mean	SE	Z-Test	MSE	Coverage	Coverage
ELA							
3	Reading Vocabulary	-0.04	0.01	0.00	1.23	0.91	0.04
	Reading Comprehension	0.00	0.00	0.95	0.32	4.58	0.81
	Writing Skills	0.01	0.01	0.05	0.69	2.09	0.06
	Overall	0.01	0.00	0.40	0.17	5.07	1.19
4	Reading Vocabulary	-0.08	0.01	0.00	1.22	0.96	0.04
	Reading Comprehension	0.00	0.00	0.64	0.33	4.40	0.63
	Writing Skills	-0.03	0.01	0.00	0.70	2.16	0.14
	Overall	0.00	0.00	1.00	0.18	5.10	1.08
5	Reading Vocabulary	-0.05	0.01	0.00	1.25	0.79	0.03
	Reading Comprehension	0.00	0.00	0.43	0.32	4.34	0.52
	Writing Skills	0.00	0.01	0.52	0.66	1.98	0.06
	Overall	0.00	0.00	0.77	0.17	5.07	0.92
6	Reading Vocabulary	-0.03	0.01	0.00	1.16	0.60	0.00
	Reading Comprehension	0.00	0.00	0.58	0.32	4.21	0.64
	Writing Skills	-0.01	0.01	0.22	0.65	1.63	0.05
	Overall	0.00	0.00	0.61	0.16	4.94	1.06
7	Reading Vocabulary	-0.04	0.01	0.00	1.13	0.94	0.04
	Reading Comprehension	-0.01	0.00	0.40	0.35	3.95	0.51
	Writing Skills	0.00	0.00	0.71	0.56	2.35	0.10
	Overall	0.00	0.00	0.75	0.16	4.73	0.84
8	Reading Vocabulary	-0.06	0.01	0.00	1.25	0.79	0.04
	Reading Comprehension	0.00	0.00	0.64	0.32	4.30	0.68
	Writing Skills	0.04	0.01	0.00	0.66	1.76	0.09
	Overall	0.00	0.00	0.79	0.17	4.77	0.83
Mathematics							
3	Number	-0.02	0.00	0.00	0.49	3.16	0.30
	Algebra	-0.05	0.01	0.00	1.09	1.22	0.06
	Geometry	-0.04	0.01	0.00	0.78	2.22	0.16
	Data	0.00	0.01	0.99	1.06	1.14	0.09
	Overall	0.00	0.00	0.98	0.18	5.11	1.11
4	Number	-0.01	0.00	0.39	0.50	3.08	0.30
	Algebra	0.01	0.01	0.10	0.90	1.44	0.07
	Geometry	-0.01	0.01	0.16	0.91	1.41	0.06
	Data	0.04	0.01	0.00	1.14	1.08	0.03
	Overall	0.00	0.00	0.55	0.17	5.12	0.87
5	Number	-0.01	0.00	0.13	0.51	3.31	0.35
	Algebra	0.00	0.01	0.87	0.92	1.92	0.08
	Geometry	-0.03	0.01	0.00	0.95	1.54	0.09
	Data	-0.04	0.01	0.00	1.12	1.19	0.08
	Overall	0.00	0.00	0.75	0.18	5.34	1.07
6	Number	-0.02	0.01	0.05	0.77	2.01	0.13
	Algebra	-0.01	0.00	0.17	0.51	3.28	0.28
	Geometry	-0.01	0.01	0.29	1.10	0.98	0.03
	Data	0.03	0.01	0.00	1.09	1.13	0.08
	Overall	0.00	0.00	0.66	0.17	5.03	1.06
	Number	0.03	0.01	0.00	0.91	1.41	0.04
	Algebra	0.01	0.00	0.19	0.50	3.30	0.30

Table 5.5: Mean Bias of the Ability Estimation (True - Estimated) – Simulation, cont.

7	Geometry	0.04	0.01	0.00	0.92	1.43	0.09
	Data	0.06	0.01	0.00	1.10	1.31	0.07
	Overall	0.00	0.00	0.70	0.17	5.36	1.16
8	Number	0.02	0.01	0.01	0.79	2.01	0.14
	Algebra	0.01	0.01	0.14	0.76	2.04	0.12
	Geometry	0.02	0.01	0.04	0.66	2.83	0.18
	Data	-0.02	0.01	0.01	1.11	0.76	0.04
	Overall	0.00	0.00	0.93	0.17	5.15	1.12

Table 5.6 and Table 5.7 present the score precision and reliability estimates for the simulation and engine evaluation studies, respectively, including the average number of items administered, the standard deviation (SD) of the estimated theta, the mean SEM, the root mean square error (RMSE), and a marginal reliability coefficient. For both studies, the SD, mean SEM, and RMSE are relatively small. The marginal reliability for the simulations ranges from 0.84 to 0.86 for ELA and 0.88 to 0.90 for Mathematics, whereas for engine evaluation ranges from 0.84 to 0.88 for ELA and 0.89 to 0.92 for Mathematics . These results indicate that, overall, the score precision is relatively good.

Table 5.6: Score Precision and Reliability – Simulation

Grade	Reporting Category	Average #Items	SD of Estimated Theta	Mean SEM	RMSE	Reliability
ELA						
3	Reading Vocabulary	4	1.48	1.19	1.23	0.31
	Reading Comprehension	13	1.12	0.56	0.56	0.75
	Writing Skills	6	1.27	0.81	0.82	0.58
	Overall	23	1.04	0.40	0.41	0.85
4	Reading Vocabulary	4	1.48	1.24	1.30	0.22
	Reading Comprehension	13	1.15	0.56	0.56	0.76
	Writing Skills	6	1.29	0.81	0.83	0.59
	Overall	23	1.06	0.41	0.41	0.85
5	Reading Vocabulary	4	1.46	1.22	1.27	0.25
	Reading Comprehension	13	1.10	0.56	0.56	0.74
	Writing Skills	6	1.26	0.79	0.80	0.59
	Overall	23	1.01	0.40	0.40	0.84
6	Reading Vocabulary	4	1.43	1.10	1.14	0.36
	Reading Comprehension	13	1.08	0.55	0.56	0.73
	Writing Skills	6	1.24	0.77	0.79	0.60
	Overall	23	0.99	0.39	0.39	0.84
7	Reading Vocabulary	4	1.47	1.11	1.16	0.38
	Reading Comprehension	12	1.15	0.58	0.58	0.74
	Writing Skills	7	1.25	0.73	0.74	0.65
	Overall	23	1.05	0.40	0.40	0.86
8	Reading Vocabulary	4	1.44	1.24	1.29	0.20
	Reading Comprehension	13	1.08	0.55	0.56	0.74
	Writing Skills	6	1.25	0.79	0.81	0.58
	Overall	23	1.00	0.40	0.40	0.84

Table 5.6: Score Precision and Reliability – Simulation, cont.

Mathematics						
3	Number	9	1.45	0.69	0.70	0.77
	Algebra	4	1.64	1.10	1.14	0.52
	Geometry	6	1.54	0.88	0.90	0.66
	Data	4	1.65	1.10	1.13	0.53
	Overall	23	1.33	0.41	0.41	0.90
4	Number	9	1.39	0.69	0.70	0.74
	Algebra	5	1.52	0.95	0.98	0.59
	Geometry	5	1.53	0.95	0.97	0.60
	Data	4	1.61	1.10	1.13	0.51
	Overall	23	1.24	0.41	0.41	0.89
5	Number	9	1.44	0.70	0.71	0.76
	Algebra	5	1.59	0.95	0.97	0.63
	Geometry	5	1.56	1.04	1.11	0.50
	Data	4	1.58	1.14	1.20	0.43
	Overall	23	1.32	0.42	0.42	0.90
6	Number	6	1.50	0.87	0.88	0.65
	Algebra	9	1.41	0.70	0.71	0.75
	Geometry	4	1.59	1.08	1.11	0.52
	Data	4	1.60	1.08	1.11	0.52
	Overall	23	1.26	0.41	0.41	0.90
7	Number	5	1.50	0.95	0.97	0.58
	Algebra	9	1.34	0.69	0.70	0.73
	Geometry	5	1.49	0.97	0.99	0.56
	Data	4	1.55	1.09	1.12	0.48
	Overall	23	1.20	0.41	0.41	0.88
8	Number	6	1.55	0.88	0.89	0.67
	Algebra	6	1.54	0.86	0.88	0.67
	Geometry	7	1.51	0.79	0.80	0.72
	Data	4	1.66	1.10	1.14	0.53
	Overall	23	1.32	0.41	0.41	0.90

Table 5.7: Score Precision and Reliability - Engine Evaluation

Grade	Reporting Category	Average #Items	SD of Estimated Theta	Mean SEM	RMSE	Reliability
ELA						
3	Reading Vocabulary	4	1.68	1.20	1.25	0.45
	Reading Comprehension	13	1.35	0.57	0.57	0.82
	Writing Skills	6	1.30	0.83	0.84	0.58
	Overall	23	1.20	0.41	0.41	0.88
4	Reading Vocabulary	4	1.61	1.24	1.29	0.35
	Reading Comprehension	13	1.30	0.56	0.57	0.81
	Writing Skills	6	1.27	0.81	0.82	0.58
	Overall	23	1.16	0.41	0.41	0.87
5	Reading Vocabulary	4	1.58	1.21	1.26	0.36
	Reading Comprehension	13	1.27	0.56	0.57	0.80

Table 5.7: Score Precision and Reliability - Engine Evaluation, cont.

	Writing Skills	6	1.27	0.80	0.81	0.59
	Overall	23	1.13	0.41	0.41	0.87
6	Reading Vocabulary	4	1.50	1.12	1.16	0.40
	Reading Comprehension	13	1.26	0.56	0.57	0.80
	Writing Skills	6	1.24	0.78	0.80	0.58
	Overall	23	1.09	0.39	0.40	0.87
7	Reading Vocabulary	4	1.46	1.10	1.15	0.37
	Reading Comprehension	12	1.23	0.58	0.59	0.77
	Writing Skills	7	1.17	0.73	0.74	0.60
	Overall	23	1.05	0.40	0.40	0.86
8	Reading Vocabulary	4	1.51	1.24	1.29	0.26
	Reading Comprehension	13	1.17	0.56	0.56	0.77
	Writing Skills	6	1.16	0.78	0.80	0.52
	Overall	23	1.02	0.40	0.40	0.84
Mathematics						
3	Number	9	1.70	0.72	0.73	0.82
	Algebra	4	1.70	1.10	1.14	0.55
	Geometry	6	1.58	0.88	0.90	0.68
	Data	4	1.74	1.13	1.16	0.55
	Overall	23	1.44	0.41	0.42	0.92
4	Number	9	1.55	0.71	0.72	0.78
	Algebra	5	1.69	0.99	1.02	0.64
	Geometry	5	1.68	0.99	1.02	0.64
	Data	4	1.75	1.15	1.19	0.54
	Overall	23	1.35	0.41	0.42	0.91
5	Number	9	1.54	0.71	0.72	0.78
	Algebra	5	1.63	0.98	1.00	0.62
	Geometry	5	1.62	1.03	1.07	0.56
	Data	4	1.64	1.14	1.19	0.48
	Overall	23	1.31	0.41	0.42	0.90
6	Number	6	1.59	0.88	0.90	0.68
	Algebra	9	1.54	0.72	0.73	0.78
	Geometry	4	1.78	1.12	1.16	0.58
	Data	4	1.66	1.14	1.18	0.49
	Overall	23	1.34	0.41	0.41	0.90
7	Number	5	1.56	0.97	1.00	0.59
	Algebra	9	1.44	0.71	0.73	0.75
	Geometry	5	1.57	1.01	1.04	0.56
	Data	4	1.64	1.12	1.16	0.50
	Overall	23	1.24	0.41	0.41	0.89
8	Number	6	1.63	0.91	0.93	0.67
	Algebra	6	1.63	0.89	0.91	0.69
	Geometry	7	1.58	0.82	0.83	0.72
	Data	4	1.56	1.11	1.15	0.45
	Overall	23	1.31	0.41	0.41	0.90

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

Table 5.8: SEM by Deciles -Simulation

	Proficiency Score Distribution										
Grade	**Decile1**	**Decile2**	**Decile3**	**Decile4**	**Decile5**	**Decile6**	**Decile7**	**Decile8**	**Decile9**	**Decile10**	**Overall**
ELA											
3	0.43	0.41	0.40	0.40	0.39	0.39	0.40	0.40	0.41	0.43	0.40
4	0.42	0.40	0.39	0.39	0.39	0.39	0.40	0.41	0.42	0.47	0.41
5	0.43	0.41	0.40	0.39	0.39	0.38	0.38	0.38	0.40	0.44	0.40
6	0.42	0.39	0.38	0.37	0.37	0.37	0.38	0.38	0.39	0.42	0.39
7	0.44	0.40	0.39	0.38	0.38	0.38	0.38	0.39	0.40	0.43	0.40
8	0.44	0.41	0.40	0.39	0.39	0.38	0.38	0.39	0.40	0.43	0.40
Mathematics											
3	0.43	0.41	0.40	0.40	0.40	0.40	0.40	0.40	0.41	0.46	0.41
4	0.44	0.42	0.41	0.40	0.40	0.40	0.40	0.40	0.40	0.42	0.41
5	0.42	0.41	0.40	0.40	0.39	0.39	0.40	0.40	0.42	0.54	0.42
6	0.43	0.41	0.41	0.40	0.40	0.40	0.40	0.40	0.40	0.42	0.41
7	0.44	0.42	0.41	0.41	0.40	0.40	0.39	0.39	0.39	0.41	0.41
8	0.44	0.42	0.41	0.41	0.40	0.40	0.39	0.39	0.39	0.43	0.41

Table 5.9: SEM by Deciles -Engine Evaluation

	Proficiency Score Distribution										
Grade	**Decile1**	**Decile2**	**Decile3**	**Decile4**	**Decile5**	**Decile6**	**Decile7**	**Decile8**	**Decile9**	**Decile10**	**Overall**
ELA											
3	0.48	0.43	0.41	0.40	0.39	0.39	0.40	0.40	0.41	0.42	0.41
4	0.45	0.41	0.39	0.39	0.39	0.39	0.39	0.40	0.42	0.46	0.41
5	0.47	0.43	0.41	0.40	0.39	0.38	0.38	0.38	0.39	0.44	0.41
6	0.47	0.40	0.38	0.37	0.37	0.37	0.37	0.38	0.39	0.42	0.39
7	0.47	0.41	0.39	0.38	0.38	0.37	0.38	0.38	0.39	0.43	0.40
8	0.47	0.41	0.40	0.39	0.38	0.38	0.38	0.38	0.39	0.44	0.40
Mathematics											
3	0.44	0.42	0.41	0.40	0.40	0.40	0.40	0.40	0.41	0.48	0.41
4	0.47	0.43	0.41	0.41	0.40	0.40	0.40	0.40	0.40	0.43	0.41
5	0.44	0.42	0.41	0.40	0.39	0.39	0.39	0.39	0.41	0.50	0.41
6	0.46	0.42	0.41	0.41	0.40	0.40	0.40	0.40	0.40	0.42	0.41
7	0.48	0.43	0.42	0.41	0.41	0.40	0.40	0.39	0.39	0.42	0.41
8	0.46	0.43	0.42	0.41	0.40	0.40	0.39	0.39	0.39	0.42	0.41

6. Psychometric Analyses

During the Spring 2021 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 Phase I Pilot 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 for field test items
- Science field test analyses
- Common item linking between NSCAS and MAP Growth for ELA and Mathematics

6.1 Number of Student Included in the Analyses

Table 6.1 presents the number of students included in the post-administration analyses presented in this section (i.e., classical analyses, DIF, IRT calibration, equating, and scaling). As in the 2018 and 2019 technical reports, 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

Grade	Test ID	N
ELA		
3	5296	21,796
4	5297	21,723
5	5298	22,232
6	5299	22,308
7	5300	22,106
8	5301	20,708
Mathematics		
3	5302	21,776
4	5303	21,689
5	5304	22,199
6	5305	22,288
7	5306	22,071
8	5307	20,672
Science		
5	5268	22,201
8	5269	20,693

6.2 Classical Item Analyses

This section summarizes the p-values and item-total correlations for operational and field test items. 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.

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

Table 6.2: Summary P-Values: Operational Items

						#Items by P-Value Range									
Grade	**#Items**	**Mean**	**SD**	**Min**	**Max**	**≤0.1**	**≤0.2**	**≤0.3**	**≤0.4**	**≤0.5**	**≤0.6**	**≤0.7**	**≤0.8**	**≤0.9**	**>0.9**
ELA															
3	590	0.488	0.118	0.064	0.927	2	7	20	92	188	194	64	19	3	1
4	579	0.538	0.130	0.076	1.000	1	0	10	66	156	184	100	43	15	4
5	508	0.524	0.129	0.000	0.970	1	7	11	49	147	170	81	34	5	3
6	518	0.519	0.121	0.137	0.885	0	2	18	64	151	144	106	26	7	0
7	478	0.520	0.127	0.000	0.924	1	0	12	65	140	145	80	25	8	2
8	553	0.550	0.134	0.000	0.987	2	4	7	57	117	177	125	46	12	6
Mathematics															
3	540	0.531	0.088	0.030	0.843	2	1	4	27	131	266	100	7	2	0
4	418	0.476	0.084	0.000	0.785	1	0	10	55	187	148	12	5	0	0
5	432	0.530	0.097	0.250	1.000	0	0	6	28	121	195	69	9	2	2
6	537	0.488	0.092	0.164	0.844	0	3	20	62	192	214	43	2	1	0
7	457	0.442	0.095	0.142	0.807	0	4	29	103	212	87	17	4	1	0
8	435	0.457	0.093	0.000	0.738	1	2	15	92	194	104	24	3	0	0

Table 6.3: Summary P-Values: Field Test Items

						#Items by P-Value Range									
Grade	**#Items**	**Mean**	**SD**	**Min**	**Max**	**≤0.1**	**≤0.2**	**≤0.3**	**≤0.4**	**≤0.5**	**≤0.6**	**≤0.7**	**≤0.8**	**≤0.9**	**>0.9**
ELA															
3	184	0.490	0.162	0.083	0.877	1	6	15	32	50	36	23	14	7	0
4	185	0.522	0.174	0.038	0.918	1	5	14	21	41	48	26	15	13	1
5	186	0.508	0.167	0.142	0.935	0	2	23	32	27	45	25	27	4	1
6	173	0.486	0.167	0.123	0.911	0	7	20	28	38	36	25	12	6	1
7	180	0.542	0.165	0.099	0.925	1	1	11	24	34	50	26	19	13	1
8	227	0.569	0.178	0.168	0.957	0	6	13	22	38	46	45	34	19	4
Mathematics															
3	231	0.510	0.203	0.012	0.961	4	11	24	38	37	38	29	30	17	3
4	150	0.528	0.164	0.159	0.858	0	1	16	18	30	33	27	17	8	0
5	182	0.554	0.179	0.137	0.972	0	4	13	15	35	43	29	32	6	5
6	231	0.511	0.205	0.054	0.914	4	10	28	31	33	48	30	26	18	3
7	226	0.446	0.214	0.022	0.926	12	19	26	44	31	36	25	19	13	1
8	157	0.400	0.201	0.030	0.860	5	24	24	33	24	19	14	10	4	0
Science															
5	58	0.548	0.191	0.076	0.929	1	2	1	6	14	11	13	3	5	2
8	51	0.388	0.221	0.009	0.812	6	6	7	10	6	6	4	5	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

						#Items by Item-Total Correlation Range						
Grade	**#Items**	**Mean**	**SD**	**Min**	**Max**	**≤0.1**	**≤0.2**	**≤0.3**	**≤0.4**	**≤0.5**	**≤0.6**	**>0.6**
ELA												
3	590	0.392	0.089	0.010	0.906	2	7	60	263	201	47	10
4	579	0.385	0.085	0.000	0.781	3	2	66	287	183	26	12
5	508	0.380	0.084	-0.188	0.648	3	6	56	246	166	29	2
6	518	0.390	0.084	0.087	0.730	1	5	64	225	169	47	7
7	478	0.383	0.081	0.000	0.767	1	2	63	229	149	29	5
8	553	0.395	0.091	0.000	0.815	2	7	57	235	196	47	9
Mathematics												
3	540	0.399	0.078	0.200	0.754	0	0	47	250	194	36	13
4	418	0.396	0.085	0.000	0.691	1	3	37	187	152	24	14
5	432	0.422	0.100	0.000	1.000	1	2	35	152	154	73	15
6	537	0.395	0.084	0.146	0.688	0	2	50	261	162	51	11
7	457	0.386	0.080	0.104	0.622	0	3	57	212	151	29	5
8	435	0.389	0.079	0.000	0.647	1	1	45	210	144	28	6

Table 6.5: Summary Item-Total Correlations: Field Test Items

						#Items by Item-Total Correlation Range						
Grade	**#Items**	**Mean**	**SD**	**Min**	**Max**	**≤0.1**	**≤0.2**	**≤0.3**	**≤0.4**	**≤0.5**	**≤0.6**	**>0.6**
ELA												
3	184	0.325	0.137	-0.143	0.614	13	23	34	57	39	16	2
4	185	0.315	0.143	-0.149	0.562	17	18	44	53	41	12	0
5	186	0.309	0.134	-0.136	0.569	13	30	34	56	46	7	0
6	173	0.306	0.136	-0.159	0.602	12	27	36	44	48	5	1
7	180	0.319	0.110	-0.007	0.552	7	20	44	71	31	7	0
8	227	0.318	0.122	-0.109	0.545	14	22	47	80	60	4	0
Mathematics												
3	231	0.390	0.125	-0.073	0.631	4	15	29	55	87	36	5
4	150	0.419	0.119	-0.188	0.629	2	4	13	34	60	35	2
5	182	0.403	0.114	0.094	0.618	2	7	21	57	54	37	4
6	231	0.365	0.111	-0.036	0.590	6	9	44	71	84	17	0
7	226	0.366	0.120	-0.170	0.618	5	16	35	77	69	21	3
8	157	0.369	0.119	-0.015	0.616	6	5	33	40	52	19	2
Science												
5	58	0.441	0.117	0.119	0.629	0	3	3	12	19	20	1
8	51	0.403	0.119	0.090	0.644	1	2	7	12	18	9	2

6.2.3 Item Suppression

Based on the item analysis conducted using the Spring 2021 results and the flagging criteria presented in Table 6.6 and Table 6.7 for multiple-choice (MC) and partial-credit (i.e., non-MC) items, 43 MC items and 19 non-MC items from the adaptive assessments were identified for content and psychometric review.

After the content and psychometric team reviewed these flagged items, NWEA recommended suppressing no items from the 2021 scoring and removing 14 items (14 ELA and no Mathematics items) from the future item pool. All recommendations were approved by NDE. There was one Grade 5 ELA item (11194960) that did not have step parameters, so the engine suppressed the item from use in scoring.

Table 6.6: Flagging Criteria for MC Items

Flag Type*	Criterion
low item-total	< 0.20
high item-total for a distractor	> 0.05

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

Table 6.7: Flagging Criteria for non-MC 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.

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 6.8 presents the focal and reference groups for the NSCAS DIF analyses.

Table 6.8: 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 administer 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 provide 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):

$$Model1 : logit(P) = \beta_0 + \beta_1 X + \beta_2 E$$
$$Model2 : logit(P) = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 E$$
$$Model3 : logit(P) = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG + \beta_4 E$$

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:

$$\chi^2 = [-2\ ln\ L(Model2)] - [-2\ ln\ L(Model3)]$$

.

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

$$\chi^2 = [-2\ ln\ L(Model1)] - [-2\ ln\ L(Model2)]$$

.

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

$$\chi^2 = [-2\ ln\ L(Model1)] - [-2\ ln\ L(Model3)]$$

.

The effect size is also used to avoid practically trivial but statistically significant results (French & Miller, 1996). Effect size is indicated by the difference of the Nagelkerke ΔR^2 between two models (Gómez-Benito, Hidalgo, & Padilla, 2009). Table 6.9 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 6.9: 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:

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

where;

- R_{rm} 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):

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

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

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

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:

$$SMD = \sum p_{FK} m_{FK} - \sum p_{RK} m_{RK} \tag{6.4}$$

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 item score of the focal group students 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 of the reference group students 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 6.10 and Table 6.11 present the Educational Testing Service (ETS) DIF categories for

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

Table 6.10: MH DIF Categories for Dichotomous Items

DIF Category	Level of DIF	Definition*
A	Negligible	MH χ^2 test is not significant at 0.05 level or $\mid$ MH D-DIF $\mid < 1.0$
B	Moderate	MH χ^2 test is significant at 0.05 level and $1.0 \leq \mid$ MH D-DIF $\mid < 1.5$
C	Strong	MH χ^2 test is significant at 0.05 level and $\mid$ MH D-DIF $\mid \geq 1.5$

[*] | MH D-DIF |= Absolute value of the Mantel-Haenszel delta difference.

Table 6.11: MH DIF Categories for Polytomous Items

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

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

6.3.3 DIF Results

Tables 6.12, 6.13, and 6.14 present the number of operational items assigned to each DIF category for DIF, UIDIF, and NUIDIF, respectively. Tables 6.15, 6.16, and 6.17 present the number of field test items assigned to each category for DIF, UIDIF, and NUIDIF, respectively. Table 6.18 presents the number of items assigned to each MH DIF category for Science field test items. For both LR and MH DIF, raw scores were used for matching criterion. Male was the reference group for gender, and white was the reference group for ethnicity. DIF was not conducted if the sample size for either group was less than 250. The + sign next to the DIF category indicates that the item is in favor of the reference group, and the - sign indicates that the item is in favor of the focal group. As shown in the tables, most items were categorized as DIF Category A (negligible DIF).

Table 6.12: LR DIF Results: Operational Items

		#Items by DIF Category							
Grade	Focal Group	Total	A	B	B+	B-	C	C+	C-
ELA									
3	Female	351	351	0	0	0	0	0	0
	Black or African American	9	9	0	0	0	0	0	0
	Hispanic	78	77	1	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	2	2	0	0	0	0	0	0

Table 6.12: LR DIF Results: Operational Items, cont.

4	Female	411	408	2	0	1	0	0	0
	Black or African American	2	2	0	0	0	0	0	0
	Hispanic	81	81	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
5	Female	380	379	1	0	0	0	0	0
	Black or African American	12	12	0	0	0	0	0	0
	Hispanic	86	86	0	0	0	0	0	0
	Asian	2	2	0	0	0	0	0	0
	Two or More Races	7	7	0	0	0	0	0	0
6	Female	349	347	2	0	0	0	0	0
	Black or African American	9	9	0	0	0	0	0	0
	Hispanic	99	99	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	3	3	0	0	0	0	0	0
7	Female	299	297	2	0	0	0	0	0
	Black or African American	20	20	0	0	0	0	0	0
	Hispanic	72	72	0	0	0	0	0	0
	Asian	2	2	0	0	0	0	0	0
	Two or More Races	6	6	0	0	0	0	0	0
8	Female	319	316	1	0	2	0	0	0
	Black or African American	10	10	0	0	0	0	0	0
	Hispanic	66	66	0	0	0	0	0	0
	Asian	1	1	0	0	0	0	0	0
	Two or More Races	4	4	0	0	0	0	0	0
Mathematics									
3	Female	474	470	4	0	0	0	0	0
	Black or African American	4	4	0	0	0	0	0	0
	Hispanic	36	35	1	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	1	1	0	0	0	0	0	0
4	Female	414	413	0	0	1	0	0	0
	Black or African American	5	5	0	0	0	0	0	0
	Hispanic	134	134	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
5	Female	406	399	6	0	0	1	0	0
	Black or African American	6	6	0	0	0	0	0	0
	Hispanic	133	133	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
6	Female	500	496	3	0	1	0	0	0
	Black or African American	2	2	0	0	0	0	0	0
	Hispanic	106	106	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0

Table 6.12: LR DIF Results: Operational Items, cont.

	Two or More Races	0	0	0	0	0	0	0	0
7	Female	441	437	2	0	1	0	0	1
	Black or African American	12	12	0	0	0	0	0	0
	Hispanic	115	114	1	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	2	2	0	0	0	0	0	0
8	Female	426	424	2	0	0	0	0	0
	Black or African American	6	6	0	0	0	0	0	0
	Hispanic	109	109	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0

Table 6.13: LR UIDIF Results: Operational Items

		#Items by DIF Category					
Grade	**Focal Group**	**Total**	**A**	**B+**	**B-**	**C+**	**C-**
ELA							
3	Female	351	351	0	0	0	0
	Black or African American	9	9	0	0	0	0
	Hispanic	78	77	0	1	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	2	2	0	0	0	0
4	Female	411	409	1	1	0	0
	Black or African American	2	2	0	0	0	0
	Hispanic	81	81	0	0	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0
5	Female	380	379	0	1	0	0
	Black or African American	12	12	0	0	0	0
	Hispanic	86	86	0	0	0	0
	Asian	2	2	0	0	0	0
	Two or More Races	7	7	0	0	0	0
6	Female	349	347	1	1	0	0
	Black or African American	9	9	0	0	0	0
	Hispanic	99	99	0	0	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	3	3	0	0	0	0
7	Female	299	298	0	1	0	0
	Black or African American	20	20	0	0	0	0
	Hispanic	72	72	0	0	0	0
	Asian	2	2	0	0	0	0
	Two or More Races	6	6	0	0	0	0
8	Female	319	317	0	2	0	0
	Black or African American	10	10	0	0	0	0
	Hispanic	66	66	0	0	0	0
	Asian	1	1	0	0	0	0

Table 6.13: LR UIDIF Results: Operational Items, cont.

	Two or More Races	4	4	0	0	0	0
Mathematics							
3	Female	474	471	0	3	0	0
	Black or African American	4	4	0	0	0	0
	Hispanic	36	35	0	1	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	1	1	0	0	0	0
4	Female	414	413	0	1	0	0
	Black or African American	5	5	0	0	0	0
	Hispanic	134	134	0	0	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0
5	Female	406	399	0	6	0	1
	Black or African American	6	6	0	0	0	0
	Hispanic	133	133	0	0	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0
6	Female	500	497	0	3	0	0
	Black or African American	2	2	0	0	0	0
	Hispanic	106	106	0	0	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0
7	Female	441	437	1	2	0	1
	Black or African American	12	12	0	0	0	0
	Hispanic	115	114	0	1	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	2	2	0	0	0	0
8	Female	426	424	0	2	0	0
	Black or African American	6	6	0	0	0	0
	Hispanic	109	109	0	0	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0

Table 6.14: LR NUIDIF Results: Operational Items

		#Items by DIF Category							
Grade	**Focal Group**	**Total**	**A**	**B**	**B+**	**B-**	**C**	**C+**	**C-**
ELA									
3	Female	351	351	0	0	0	0	0	0
	Black or African American	9	9	0	0	0	0	0	0
	Hispanic	78	78	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	2	2	0	0	0	0	0	0
4	Female	411	411	0	0	0	0	0	0
	Black or African American	2	2	0	0	0	0	0	0
	Hispanic	81	81	0	0	0	0	0	0

Table 6.14: LR NUIDIF Results: Operational Items, cont.

	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
5	Female	380	380	0	0	0	0	0	0
	Black or African American	12	12	0	0	0	0	0	0
	Hispanic	86	86	0	0	0	0	0	0
	Asian	2	2	0	0	0	0	0	0
	Two or More Races	7	7	0	0	0	0	0	0
6	Female	349	349	0	0	0	0	0	0
	Black or African American	9	9	0	0	0	0	0	0
	Hispanic	99	99	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	3	3	0	0	0	0	0	0
7	Female	299	299	0	0	0	0	0	0
	Black or African American	20	20	0	0	0	0	0	0
	Hispanic	72	72	0	0	0	0	0	0
	Asian	2	2	0	0	0	0	0	0
	Two or More Races	6	6	0	0	0	0	0	0
8	Female	319	319	0	0	0	0	0	0
	Black or African American	10	10	0	0	0	0	0	0
	Hispanic	66	66	0	0	0	0	0	0
	Asian	1	1	0	0	0	0	0	0
	Two or More Races	4	4	0	0	0	0	0	0
Mathematics									
3	Female	474	474	0	0	0	0	0	0
	Black or African American	4	4	0	0	0	0	0	0
	Hispanic	36	36	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	1	1	0	0	0	0	0	0
4	Female	414	414	0	0	0	0	0	0
	Black or African American	5	5	0	0	0	0	0	0
	Hispanic	134	134	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
5	Female	406	406	0	0	0	0	0	0
	Black or African American	6	6	0	0	0	0	0	0
	Hispanic	133	133	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
6	Female	500	500	0	0	0	0	0	0
	Black or African American	2	2	0	0	0	0	0	0
	Hispanic	106	106	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
7	Female	441	441	0	0	0	0	0	0
	Black or African American	12	12	0	0	0	0	0	0
	Hispanic	115	115	0	0	0	0	0	0

Table 6.14: LR NUIDIF Results: Operational Items, cont.

	Asian	0	0	0	0	0	0	0	0
	Two or More Races	2	2	0	0	0	0	0	0
8	Female	426	426	0	0	0	0	0	0
	Black or African American	6	6	0	0	0	0	0	0
	Hispanic	109	109	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0

Table 6.15: LR DIF Results: Field Test Items

		#Items by DIF Category							
Grade	Focal Group	Total	A	B	B+	B-	C	C+	C-
ELA									
3	Female	184	184	0	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	13	13	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
4	Female	185	184	1	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	13	13	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
5	Female	186	186	0	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	41	41	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
6	Female	173	173	0	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	38	38	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
7	Female	180	179	1	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	16	16	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
8	Female	115	115	0	0	0	0	0	0
	Black or African American	1	1	0	0	0	0	0	0
	Hispanic	14	14	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
Mathematics									
	Female	231	230	1	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0

Table 6.15: LR DIF Results: Field Test Items, cont.

3	Hispanic	1	1	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
4	Female	150	150	0	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	1	1	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
5	Female	182	181	1	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	0	0	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
6	Female	231	231	0	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	0	0	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
7	Female	156	155	1	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	8	8	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
8	Female	157	157	0	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	0	0	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0

Table 6.16: LR UIDIF Results: Field Test Items

		#Items by DIF Category					
Grade	Focal Group	Total	A	B+	B-	C+	C-
ELA							
3	Female	184	184	0	0	0	0
	Black or African American	0	0	0	0	0	0
	Hispanic	13	13	0	0	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0
4	Female	185	185	0	0	0	0
	Black or African American	0	0	0	0	0	0
	Hispanic	13	13	0	0	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0
	Female	186	186	0	0	0	0
	Black or African American	0	0	0	0	0	0

Table 6.16: LR UIDIF Results: Field Test Items, cont.

5	Hispanic	41	41	0	0	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0
6	Female	173	173	0	0	0	0
	Black or African American	0	0	0	0	0	0
	Hispanic	38	38	0	0	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0
7	Female	180	180	0	0	0	0
	Black or African American	0	0	0	0	0	0
	Hispanic	16	16	0	0	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0
8	Female	115	115	0	0	0	0
	Black or African American	1	1	0	0	0	0
	Hispanic	14	14	0	0	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0
Mathematics							
3	Female	231	231	0	0	0	0
	Black or African American	0	0	0	0	0	0
	Hispanic	1	1	0	0	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0
4	Female	150	150	0	0	0	0
	Black or African American	0	0	0	0	0	0
	Hispanic	1	1	0	0	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0
5	Female	182	182	0	0	0	0
	Black or African American	0	0	0	0	0	0
	Hispanic	0	0	0	0	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0
6	Female	231	231	0	0	0	0
	Black or African American	0	0	0	0	0	0
	Hispanic	0	0	0	0	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0
7	Female	156	155	0	1	0	0
	Black or African American	0	0	0	0	0	0
	Hispanic	8	8	0	0	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0
	Female	157	157	0	0	0	0
	Black or African American	0	0	0	0	0	0

Table 6.16: LR UIDIF Results: Field Test Items, cont.

8	Hispanic	0	0	0	0	0	0
	Asian	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0

Table 6.17: LR NUIDIF Results: Field Test Items

		#Items by DIF Category							
Grade	Focal Group	Total	A	B	B+	B-	C	C+	C-
ELA									
3	Female	184	184	0	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	13	13	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
4	Female	185	185	0	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	13	13	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
5	Female	186	186	0	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	41	41	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
6	Female	173	173	0	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	38	38	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
7	Female	180	180	0	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	16	16	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
8	Female	115	115	0	0	0	0	0	0
	Black or African American	1	1	0	0	0	0	0	0
	Hispanic	14	14	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
Mathematics									
3	Female	231	231	0	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	1	1	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
	Female	150	150	0	0	0	0	0	0

Table 6.17: LR NUIDIF Results: Field Test Items, cont.

4	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	1	1	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
5	Female	182	182	0	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	0	0	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
6	Female	231	231	0	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	0	0	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
7	Female	156	156	0	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	8	8	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0
8	Female	157	157	0	0	0	0	0	0
	Black or African American	0	0	0	0	0	0	0	0
	Hispanic	0	0	0	0	0	0	0	0
	Asian	0	0	0	0	0	0	0	0
	Two or More Races	0	0	0	0	0	0	0	0

Table 6.18: MH DIF Results: Science Field Test Items

			#Items by DIF Category				
Grade	**Focal Group**	**Total**	**A**	**B+**	**B-**	**C+**	**C-**
Science							
5	Female	58	58	0	0	0	0
	Black or African American	58	53	2	2	0	1
	Hispanic	58	57	0	1	0	0
	Asian	4	4	0	0	0	0
	Two or More Races	58	57	0	1	0	0
8	Female	51	49	0	1	0	1
	Black or African American	51	46	0	2	0	3
	Hispanic	51	48	0	2	0	1
	Asian	0	0	0	0	0	0
	Two or More Races	49	0	1	0	0	1

6.4 IRT Calibration

6.4.1 Calibration Methods

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 filed test items of ELA and Mathematics onto the NSCAS scale. For all content areas, item parameter estimations were implemented using WINSTEPS 4.8.0.0 (Linacre, 2021) that used joint maximum likelihood estimation (MLE) (Wright & 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:

$$p(\mu_{ij} = 1 \mid \theta_j, b_i) = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}} \tag{6.5}$$

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

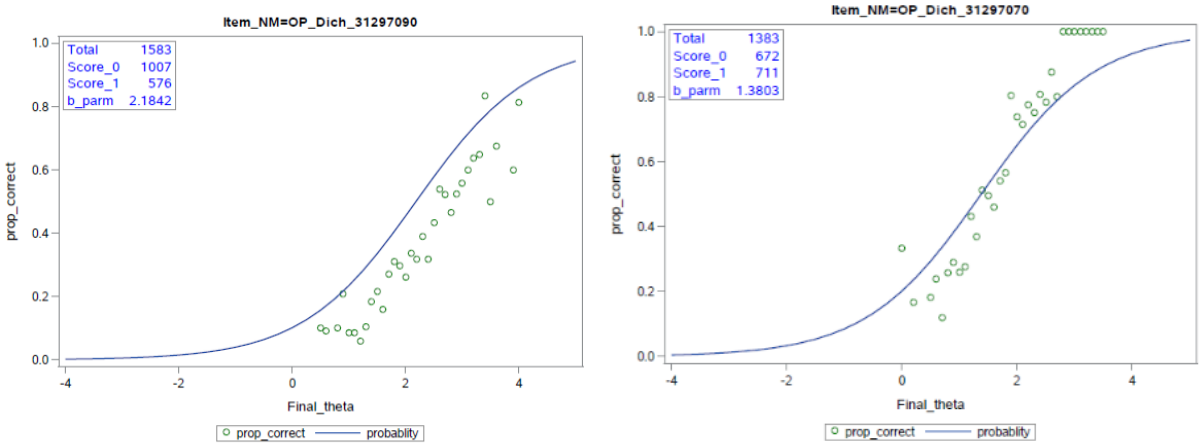
$$p(\mu_{ij} = k \mid \theta_j) = \frac{e^{\sum_{\mu=1}^{k} Da_i(\theta_j - b_i + d_{i\mu})}}{\sum_{v=1}^{m_i} e^{\sum_{\mu=1}^{k} Da_i(\theta_j - b_i + d_{i\mu})}} \tag{6.6}$$

where k is the score on the item, m_i is the total number of score categories for the item, $d_i u$ is the threshold parameter for the threshold between scores μ and μ-1, and θ_j and b_i are the person and item parameters, respectively.

Field test items were calibrated onto the NSCAS scale, following the steps below.

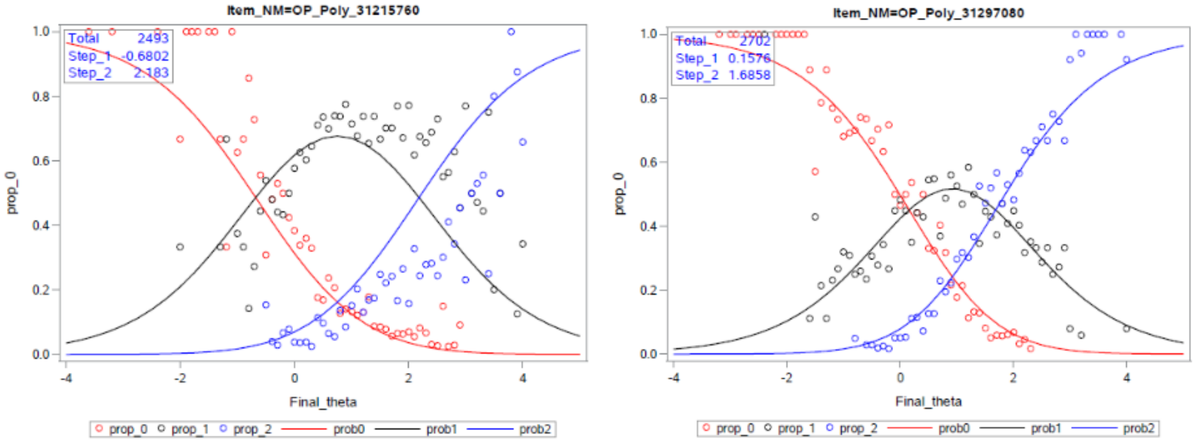
1. Determine which NSCAS operational items perform the best with the empirical data to be used as anchor items. In other words, compare the item characteristic curve (ICC) created by the existing item parameters for each item to the distribution of student responses. If the item parameters hold, the ICC curve should be very close to the distribution of student responses (i.e., the ICC line should be sitting on top of the student responses).
2. Identify field test items with flags from CIA and exclude from calibration.
3. Calibrate the field test items to the NSCAS scale while fixing NSCAS anchor items from Step 1 and excluding field test items from Step 2.
4. Review ICCs from step 3 and identify additional items to exclude
5. Identify items with very high b-parameter or step parameters (i.e., if parameter estimate $\geq$ 4.25)
6. Identify items with reversed step parameters (i.e., Step2 parameter is lower than Step1 parameter)
7. Calibration NSCAS FT items and Create ICCs with new item parameters, excluding additional NSCAS FT items from step 4, step 5, and step 6
8. Review ICCs from step 7
9. Combine items identify in step 2, step4, step 5, and step 6 (i.e., Data Review items)
10. If Data Review decision is to keep any flagged items from Step 9, calibrate them while fixing Operational items from Step 1 and NSCAS FT item parameters from Step 7.

Figure 6.1: Example Plot of ICC and Student Responses - Dichotomous Item



(a) Not Selected as an Anchor **(b) Selected as an Anchor**

Figure 6.2: Example Plot of ICC and Student Responses - Polytomous Item



(a) Not Selected as an Anchor **(b) Selected as an Anchor**

6.4.2 Calibration Results

The first step of the field test item calibration was to determine the NSCAS anchor items by reviewing and comparing plots of the ICCs and the distribution of student responses for each item. Figure F.1 and Figure 6.2 present example plots of ICC and student responses for selected items. One dichotomous and one polytomous item examples are included for either case of anchors or non-anchors to highlight how these plots were used for selecting anchors. Table 6.19 presents the total number of NSCAS operational items and the number of anchor items used in calibrating the field test items.

Table 6.20 and Table 6.21 present the summary IRT item statistics across all operational and field test items, respectively. Operational item parameter means increase by grade for ELA and Mathematics, as can be expected for vertical scales.

Table 6.19: Number of NSCAS Anchor Items used for MAP Growth Calibration

	#NSCAS Items	
Grade	**Operational**	**Anchor**
ELA		
3	590	63
4	579	78
5	508	65
6	518	67
7	478	95
8	553	90
Mathematics		
3	540	69
4	418	62
5	432	69
6	537	86
7	457	70
8	435	77

Table 6.20: Summary IRT Item Statistics: Operational Items

Grade	#Items	#Parameters	Mean	SD	Min.	Max.	Range (Max.- Min.)
ELA							
3	590	629	-0.722	1.143	-3.773	3.431	7.205
4	579	630	-0.521	1.098	-3.326	3.677	7.003
5	507	539	-0.292	1.148	-3.023	4.268	7.291
6	518	565	-0.089	1.113	-3.088	2.988	6.076
7	478	511	0.015	0.988	-2.442	2.808	5.250
8	553	592	0.172	1.137	-2.341	5.255	7.596
Mathematics							
3	540	579	-0.781	1.257	-4.877	6.297	11.174
4	418	465	0.280	1.169	-2.612	3.908	6.520
5	432	480	0.223	1.155	-4.468	3.695	8.163
6	537	597	0.697	1.304	-3.653	5.479	9.131
7	457	498	1.219	1.207	-2.005	4.950	6.955
8	435	477	1.367	1.287	-1.780	5.641	7.421

Table 6.21: Summary IRT Item Statistics: Field Test Items

Grade	#Items	#Parameters	Mean	SD	Min.	Max.	Range (Max.- Min.)
ELA							
3	125	156	-0.510	1.101	-3.063	2.434	5.498
4	121	152	-0.291	1.082	-2.889	2.401	5.290
5	112	151	-0.181	1.118	-4.090	2.388	6.477
6	98	117	0.207	0.990	-2.277	2.934	5.211
7	117	155	0.278	1.105	-2.190	2.862	5.052
8	147	181	0.348	0.959	-1.951	2.739	4.691
Mathematics							
3	179	204	-0.573	1.171	-2.981	2.988	5.968
4	126	153	0.111	1.147	-2.243	3.620	5.863
5	148	170	0.233	1.108	-2.078	3.268	5.346
6	184	206	0.390	1.225	-2.399	3.878	6.277
7	150	172	0.716	1.119	-1.844	3.984	5.828
8	107	133	1.265	1.184	-1.387	4.007	5.394
Science							
5	58	59	-0.27	1.13	-2.96	2.87	5.83
8	51	59	0.71	1.40	-1.73	5.20	6.93

6.5 Science Field Test

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.

6.5.1 Design

Table 6.22 presents the field test form design for each grade. Each grade has six test forms, each with 3-4 tasks and 4-8 associated prompts. Each test form has the same number of prompts for each grade, making the test lengths equal across forms. Each task is included on at least two test forms per grade to ensure a sufficient number of responses per task for item calibration and to allow an evaluation of how the prompts of the task are likely to function operationally. These common tasks across forms also serve as anchor sets to equate prompts across forms. For example, Task 2135 in Grade 5 is common on Forms D and E.

The order of prompts within a task is fixed, but the order of tasks on a form varies across students to reduce task position effect that can alter the quality of the data due to factors such as fatigue. For example, students might be tired at the end of a test and will not do as well as the beginning, so task positions vary across students (e.g., a task can appear early on a form for some students but in a late position for others) to ensure an even opportunity for full student engagement.

Table 6.22: Spring 2021 NSCAS Science Field Test Form Design

Task Code	#Prompts	Form A	Form B	Form C	Form D	Form E	Form F
Grade 5							
2135	7				X	X	
2136	6			X			X
2139	4		X		X		
2142	4		X	X	X		
2143	8	X				X	
2144	4		X	X			
2145	5				X	X	
2146	6	X		X			
2147	6	X					X
2149	8		X				X
	Total #Prompts	20	20	20	20	20	20
	Total #Tasks	3	4	4	4	3	3
Grade 8							
2133	5	X				X	
2150	6				X	X	
2151	5		X	X			
2154	6		X				X
2155	5			X			X
2156	6		X		X		
2158	6					X	X
2160	7	X		X			
2161	5	X			X		
	Total #Prompts	17	17	17	17	17	17
	Total #Tasks	3	3	3	3	3	3

6.5.2 Constraint-Based Engine

The pre-administration engine simulation and post-administration engine evaluation verified that the engine's population exposure control worked as intended to ensure that each test form would be administered to a representative sample of Nebraska students as defined by gender and ethnicity demographic characteristics. The engine also administered the fixed forms as intended. Prompts within a task were administered in a fixed pre-specified order, and the position of tasks on a form varied across students to reduce the risk of data quality issues due to task position effect. Detailed information regarding the simulation study can be found in the full report (NWEA, 2020b, 2021a).

6.5.3 Analyses and Calibration

Science field test items were analyzed and flagged using the same flagging criteria for ELA and Mathematics field test items (see Section 2.13). To determine the measurement model for the newly developed Nebraska science assessment based on the Next Generation Science Standards (NGSS), the following three analysis was conducted.

- Correlation between DCI, SEP and CCC
- Principal Component Analysis (PCA)

- Parallel Analysis

This 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 given the Principal components and parallel analysis. Then, 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, and
- Three-parameter logistic (3PL) for dichotomous items and general partial credit model (GPCM) for polytomous items.

Based on the fit statistics results, NWEA recommended the 1PL and PCM combination model approach, as this combination model not only fit the data well, but also provided more reasonable item difficulty parameters. NDE decided to move forward with the 1PL and PCM combination model and will reassess calibration model after the operational test in 2022. The summary IRT item statistics using the 1PL and PCM is included in Table 6.21.

6.6 Common Item Linking Between NSCAS and MAP Growth (ELA and Mathematics)

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 must be provided between the Nebraska Student-Centered Assessment System (NSCAS) and MAP Growth scales. Whereas equipercentile linking was used to produce the Rasch Unit (RIT) scores for the Spring 2021 Phase 1 Pilot administration, NWEA has been investigating various linking approaches for the Winter Pilot and beyond.

6.6.1 Embedded MAP Growth Items

To conduct the common item linking study, a set of MAP Growth items were selected and embedded at the end of the NSCAS Spring 2021 Phase 1 Pilot test forms for ELA and mathematics. NSCAS and MAP Growth use different item players, which means ELA reading passages are formatted differently. Mathematics items have different calculator rules regarding when calculators can be used and what calculator types can be used. Item display settings such as color, text font, and layout are also different. Therefore, a subset of items from the MAP Growth tests, that are similar in formatting to the NSCAS items, were selected for the common item linking study by the NWEA Content and Psychometric Solutions teams. These MAP Growth linking items were then placed at the end of the Spring 2021 Phase 1 Pilot test forms. Table 6.23 presents the number of embedded MAP Growth items selected for the item pool for each grade. These items did not contribute to operational scores.

To demonstrate how the MAP Growth items were administered during the Spring 2021 Phase 1 Pilot, NWEA ran the 2021 simulations with these MAP Growth linking items. The following constraints were imposed for the MAP Growth items:

- The total number of MAP Growth linking items for each student is 5.
- Each student gets MAP Growth linking items at the end of the test.
- MAP Growth linking items are not included for calculating student scores.

- The maximum number of passages is 1.
- The minimum number of items per passage is 3.
- The maximum number of items per reporting category is 2 or 3.
- The targeted minimum number of students for each MAP Growth item is 750.
- Students are pseudo-randomly assigned to each MAP Growth item.

Table 6.23: Number of Embedded MAP Growth Items in the Spring 2021 Phase 1 Pilot

	#Embedded MAP Growth Items			
	ELA			Mathematics
Grade	Reading	Language Usage	Total	
ELA_RD				
3	89	61	150	150
4	113	40	153	150
5	112	40	152	150
6	110	40	150	150
7	88	61	149	150
8	106	40	146	150
Total	618	282	900	900

6.6.2 Data

Student responses from the 2021 administrations of both the Pilot and MAP Growth assessments were then used to link the following NSCAS and MAP Growth assessments.

- ELA_RD = NSCAS ELA, MAP Growth Reading
- MA_MA = NSCAS Mathematics, MAP Growth Mathematics

Data from the NSCAS Spring 2021 Phase 1 Pilot assessments in ELA and mathematics were used to calibrate the embedded MAP Growth items in the common item linking study and compare achievement level distributions based on students' NSCAS scores and linked RIT scores. The Spring 2021 NSCAS and the Spring 2021 MAP Growth results from Nebraska students were merged by students to compare the RIT and linked RIT scores. To merge the data, each student's NSCAS testing record was matched to their MAP Growth score using their student ID. Only students who took both the MAP Growth and NSCAS assessments in Spring 2021 were included in the study sample. This merged data were also used to run the 2021 equipercentile linking. About 13,000 or more students were merged per grade, with 65-85% NSCAS students and 93-94% MAP Growth students merged. Demographics of the merged students are representative of the Nebraska population.

6.6.3 Linking Procedure

Common item linking was conducted following the steps below using the NSCAS Spring 2021 Phase 1 Pilot data. Steps 1-7 refer to the IRT common item linking procedure, whereas Step 8 refers to the equipercentile linking procedure.

1. Determine the NSCAS anchor items. Determine which NSCAS operational items perform the best with the empirical data to be used as anchor items. In other words, compare the

item characteristic curve (ICC) created by the existing item parameters for each item to the distribution of student responses. If the item parameters hold, the ICC curve should be very close to the distribution of student responses (i.e., the ICC line should be sitting on top of the student responses).

2. Calibrate the embedded MAP Growth items. Calibrate the embedded MAP Growth items to the NSCAS scale while fixing NSCAS anchor items from Step 1. The result is newly calibrated item parameters for the embedded MAP Growth items.
3. Verify the newly calibrated MAP Growth item parameters. Plot all MAP Growth items again (i.e., compare the ICCs to the distribution of student responses) to verify that their calibrated item parameters align with the distribution of student responses. Remove MAP Growth items flagged for low item-total correction (<0.2) or positive distractor correlation (>0.05). Use the remaining items to obtain the transformation constants in Step 4.
4. Obtain the transformation constants for each grade using the item difficulty parameter estimates between two sets of MAP Growth items (i.e., between MAP Growth bank values and calibration results using the combined MAP Growth and NSCAS data from Step 3). The MeanSigma (MS) transformation constants were obtained using the STUIRT software (Kim & Kolen, 2004). The Mean/Sigma (MS) method uses the means and the standard deviations of the b-parameter estimates.
5. Bring NSCAS items onto the RIT scale. Apply each set of transformation constants to the NSCAS items to bring them onto the RIT scale.
6. Identify cuts on the RIT scale. Apply each set of transformation constants to the NSCAS cuts to identify the IRT linked RIT cuts on the RIT scale.
7. Calculate the IRT linked RIT scores for each student by applying each set of transformation constants to the NSCAS student theta. For ELA, obtain one more set of scores that uses only Reading Vocabulary and Reading Comprehension items. This step is needed because MAP Growth Reading corresponds to these two reporting categories, whereas MAP Growth Language Usage corresponds to the third reporting category of Writing. Conduct scoring in WINSTEPS while fixing all NSCAS item parameter estimates to their RIT scale (obtained in Step 5). After implementing the scoring runs, round students' theta estimates to one digit to be consistent with the NWEA constraint-based engine.
8. Calculate the linked RIT scores based on equipercentile linking. The reported linked RIT scores for the Spring 2021 Phase 1 Pilot were based on the conversion tables from the equipercentile linking based on the 2019 data (NWEA, 2020c). Another set of equipercentile linked RIT scores were then obtained for this study following the same equipercentile linking procedure using the 2021 data to create a new conversion table. Thus, there are two sets of linked RIT scores: equipercentile linking based on 2019 data and equipercentile linking based on 2021 data.

6.6.4 Linking Results

The first step of the common item linking procedure was to determine the NSCAS anchor items. The NSCAS anchor items selected were used for calibrating both field testing items and MAP Growth items, as shown in Table 6.19.

Once the embedded MAP Growth items were calibrated while fixing the NSCAS anchor items, their item parameters were verified to ensure that they align with the distribution of student responses. Items were removed if they had a low item-total correlation (<0.2) or positive distractor correlation (>0.05). The remaining items were then used to obtain the transformation constants using STUIRT.

Table 6.24 presents these results, including the number of embedded MAP Growth items removed from the analysis and the number of items used in STUIRT to obtain the transformation constants.

The NSCAS ELA assessments include three reporting categories: Reading Vocabulary, Reading Comprehension, and Writing Skills. However, MAP Growth Reading only includes the first two reporting categories, while MAP Growth Language Usage includes the writing items. To better match the construct of the NSCAS ELA and MAP Growth Reading assessments, NWEA computed the IRT linked RIT for ELA using only the two reporting categories of Reading Vocabulary and Reading Comprehension.

Furthermore, based on the 2021 NSCAS data, there was a larger than expected number of students with low linked RIT scores who received the LOSS+2 minimum score. Further investigation showed that while most of these students responded to all 35 items, they had very low raw scores and had shorter test duration than the general population of students taking the test. Based on these results, NWEA believes that there is a possible student engagement issue for these scores and decided to remove them from all subsequent analyses.

Table 6.25 presents the descriptive statistics of the IRT linked RIT (MS) based on only two reporting categories for ELA_RD and all reporting categories for MA_MA, as well as the Fall 2020 RIT and the Spring 2021 RIT. To see if the IRT linked RIT means fall within the ±1 standard error of measurement (SEM) of the RIT means, Table 6.26 presents the mean SEM for the RIT scores from Spring 2021 merged data. Table 6.27 presents the achievement level distributions, including the distributions for NSCAS for comparison. The percentage of students at each achievement level are very similar between IRT linked RIT (MS) and equipercentile linked RIT using 2019 data that were part of the reported scores for the Spring 2021 Phase 1 Pilot.

The results indicate that IRT linked RIT (MS) scores are comparable looking at the overall population. NWEA recommended that IRT linked RIT with the MS transformation be used for the Nebraska through-year assessments, using items from the two reading reporting categories only for ELA (i.e., Reading Vocabulary and Reading Comprehension) and all items for mathematics.

Table 6.24: Number of Embedded MAP Growth Items used for Transformation

Grade	#Embedded Items (MAP Growth)	#Removed Items (MAP Growth)	#Included Items in STUIRT (MAP Growth)	Correlation between Two sets of Item Parameter Estimates
ELA_RD				
3	89	1	88	0.93
4	113	2	111	0.93
5	112	7	105	0.93
6	110	5	105	0.91
7	88	7	81	0.93
8	106	6	100	0.89
MA_MA				
3	150	4	146	0.93
4	150	6	144	0.88
5	150	6	144	0.92
6	150	11	139	0.93
7	150	10	140	0.94
8	150	29	121	0.90

Table 6.25: Descriptive Statistics of RIT and Linked RIT Scores

	RIT (Fall 2020)*					RIT (Spring 2021)*					IRT Linked RIT (MS)			
Grade	**N**	**Mean**	**SD**	**Min**	**Max**	**N**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
ELA_RD														
3	16,719	189.71	15.82	140	239	18,442	198.98	15.76	135	245	196.61	13.18	137	238
4	13,995	199.24	15.08	145	249	15,462	206.08	15.28	140	260	203.59	12.06	154	255
5	14,209	206.62	14.70	147	250	15,761	211.71	14.93	145	262	209.44	11.74	161	255
6	14,333	212.07	14.31	152	254	16,242	215.00	15.08	156	261	213.74	11.40	165	269
7	13,183	215.66	14.52	155	261	14,873	217.58	15.45	154	267	215.07	11.73	167	256
8	11,935	219.37	14.76	154	267	13,503	221.27	15.51	151	274	219.19	11.87	174	264
MA_MA														
3	14,106	188.78	12.76	121	250	15,609	202.49	14.22	138	266	203.95	13.98	171	256
4	14,122	199.78	13.54	134	256	15,548	211.21	15.64	139	269	216.23	16.78	171	281
5	14,379	209.23	14.39	135	310	15,897	219.38	17.21	144	289	223.59	17.05	174	292
6	13,951	215.48	14.12	141	276	15,687	223.27	16.78	146	288	226.88	16.40	180	294
7	12,725	222.44	15.38	146	283	14,345	227.94	17.85	138	307	231.05	16.61	185	303
8	11,722	228.39	16.50	146	297	13,316	232.81	19.15	136	316	237.72	17.52	187	310

* The Fall 2020 RIT results used merged data from Fall 2020 MAP Growth, Spring 2021 MAP Growth, and Spring 2021 NSCAS. The Spring 2021 RIT results used merged data from Spring 2021 MAP Growth and NSCAS MAP Growth. The merged Spring 2021 data were also used for the recommended IRT linked RIT (MS).

Table 6.26: Mean SEM

	RIT (Spring 2021)				IRT Linked RIT (MS)			
Grade	**Mean**	**SEM**	**Mean-1SEM**	**Mean+1SEM**	**Mean**	**SEM**	**Mean-1SEM**	**Mean+1SEM**
ELA_RD								
3	198.98	3.36	195.62	202.34	196.61	5.04	191.58	201.65
4	206.08	3.37	202.71	209.45	203.59	5.01	198.58	208.60
5	211.71	3.40	208.31	215.11	209.44	4.99	204.45	214.43
6	215.00	3.36	211.64	218.37	213.74	4.84	208.91	218.58
7	217.58	3.38	214.21	220.96	215.07	5.10	209.98	220.17
8	221.27	3.40	217.88	224.67	219.19	5.05	214.13	224.24
MA_MA								
3	202.49	2.91	199.58	205.40	203.95	4.12	199.83	208.07
4	211.21	2.92	208.29	214.13	216.23	4.11	212.12	220.35
5	219.38	2.96	216.42	222.35	223.59	4.13	219.46	227.72
6	223.27	2.91	220.36	226.19	226.88	4.09	222.79	230.97
7	227.94	2.92	225.03	230.86	231.05	4.12	226.93	235.16
8	232.81	2.92	229.89	235.73	237.72	4.09	233.63	241.81

Table 6.27: NSCAS vs. Linked RIT Achievement Level Distributions

		NSCAS			IRT Linked RIT (MS)			Equipercentile Linked RIT (2019 Data)		
Grade	N (Before Merge)	%Dev	%OT	%CCR	%Dev	%OT	%CCR	%Dev	%OT	%CCR
ELA_RD										
3	21,621	49.5	36.1	14.4	48.4	34.0	17.6	48.6	36.1	15.3
4	21,551	45.9	36.8	17.3	42.6	35.7	21.7	44.9	37.1	18.0
5	22,046	53.8	31.5	14.8	52.9	29.9	17.2	51.8	33.2	15.0
6	22,157	54.0	30.2	15.8	51.9	28.2	19.9	54.0	29.3	16.7
7	21,960	55.1	35.9	9.0	52.6	35.8	11.6	54.6	35.5	9.8
8	20,572	49.1	37.9	13.0	49.0	37.1	13.9	47.3	38.6	14.0
MA_MA										
3	21,482	52.2	38.3	9.5	51.0	38.5	10.5	49.5	40.3	10.2
4	21,605	54.2	37.7	8.2	52.5	39.1	8.4	51.9	39.5	8.5
5	22,130	54.3	38.2	7.6	54.0	38.5	7.6	52.7	39.2	8.0
6	22,167	52.7	39.2	8.1	52.5	39.0	8.5	51.4	39.9	8.7
7	22,017	53.7	38.4	7.9	53.3	38.6	8.1	53.0	39.1	7.9
8	20,611	54.5	37.8	7.7	52.4	39.3	8.3	54.5	37.8	7.7

6.6.5 Further Considerations

Although NWEA is recommending the IRT linked RIT with the MS transformation, there are areas of further consideration. First, Table 6.25 shows that the tails of the distribution are pulled in with the linked RIT as compared to the RIT. One possible reason for this is that NSCAS uses only on-grade items, while MAP Growth uses both on- and off-grade items. Including off-grade items in the through-year assessment may move student scores at both tails closer to that of the MAP Growth distribution. Also, the NSCAS LOSS may need to be adjusted to be lower, and the NSCAS HOSS may need to be higher when the new scale is set in 2022. The updates to the LOSS and HOSS are more needed considering approximately 100 students were piled at the calculated LOSS in 2021. Second, the administration dates may need to be considered as well. Using 30 days between one test's end and the other test's start date, approximately 70% of students took both MAP Growth Reading and NSCAS ELA and 80% of students took MAP Growth Mathematics and NSCAS Mathematics in Spring 2019 and Spring 2021. If data with this much time between administrations are used, it may impact linking and scoring results. Students taking both tests within 30 days would be recommended, considering that a subset of the data (i.e., 30-day data) for the common person linking produced mixed results. Lastly, the construct differences between NSCAS ELA and MAP Growth Reading still exist. MAP Growth Reading items are more stand-alone items, while all NSCAS reading items are associated with passages. Furthermore, in general, NSCAS has more items per passage. All MAP Growth passages have at least one item associated, and only 50% of students see passages with three items while the minimum number of items per passage is set to four for NSCAS.

6.7 Scaling

Science was a field test and the test did not produce a student score in 2021. Scaling for Science will be set in 2022. For ELA and Mathematics, NSCAS Phase I Pilot reports provide both NSCAS scale score and linked RIT score which was converted from the NSCAS scale score.

6.7.1 NSCAS Scale Score

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:

$$SS = (\theta \times A) + B \tag{6.7}$$

$$SSCSEM = (\theta_{CSEM} \times A) \tag{6.8}$$

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, 2013):

$$\theta_{CSEM} = CSEM(\theta_j) = \frac{1}{\sqrt{I(\theta_j)}} \tag{6.9}$$

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

$$I(\theta_j) = \sum_i \frac{p\prime_{ij}(\theta_j)^2}{p_{ij}(\theta_j)q_{ij}(\theta_j)} \tag{6.10}$$

where $p\prime_{ij}(\theta_j)$ is the derivative of $p_{ij}(\theta_j)$ and $q_{ij}(\theta_j) = 1 - p_{ij}(\theta_j)$. 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$\times$Min(SSCSEM).
4. The HOSS may be high enough that SSCSEM (Penultimate HOSS) $<$ 5$\times$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 6.28 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 6.29 summarizes the cut score implementation, or the conversions of student ability (theta) to scale scores that were used for scoring. Specifically, the table presents the calculations of the slopes and intercepts for all grades of the scale score conversions, including the cut scores set during standard setting.

Table 6.28: 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
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

Table 6.29: Cut Scores and Conversion of Theta to Scale Scores

	Cut Scores		Conversion		Cuts (Theta*)	
Grade	**On Track**	**CCR**	**Slope(A)**	**Intercept(B)**	**On Track**	**CCR**
ELA						
3	2477	2557	72.47244	2500	-0.3193	0.7867
4	2500	2582	72.47244	2500	-0.0024	1.1291
5	2531	2599	72.47244	2500	0.4309	1.3599
6	2543	2603	72.47244	2500	0.5970	1.4212
7	2556	2630	72.47244	2500	0.7741	1.7938
8	2561	2632	72.47244	2500	0.8389	1.8146
Mathematics						
3	1190	1286	54.92622	1200	-0.1821	1.5657
4	1222	1317	54.92622	1200	0.4005	2.1301
5	1236	1331	54.92622	1200	0.6554	2.3850
6	1244	1342	54.92622	1200	0.8011	2.5853
7	1247	1346	54.92622	1200	0.8557	2.6581
8	1264	1365	54.92622	1200	1.1652	3.0040

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

6.7.2 Linked RIT Score

For ELA and Mathematics, NSCAS Phase I Pilot reports provide both NSCAS scale score and linked RIT score. Calculated NSCAS scale scores were converted to linked RIT scores, using the conversion tables created from the equipercentile linking based on the 2019 data (NWEA, 2020c). Table 6.30 presents score range for both scores.

Table 6.30: Score Range (LOSS and HOSS) for NSCAS scale score and linked RIT score

	NSCAS Scale Score			Linked RIT Score		
Grade	**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

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

7. Standard Setting

No standard setting was held in 2020-2021. 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 will take place for the new NSCAS Science assessment following the first operational administration.

7.1 Overview

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

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

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

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

7.2 ID Matching Method

The *Standards* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 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 RALDs.

7.3 Meeting Process

The meetings included an overview of the NSCAS and meeting goals, training, ID Matching training, multiple rounds of judgments, RALD 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 RALDs and OIB, completed the item matching activity, and recommended cut scores.
- Round 2: Panelists reviewed the dispersion of their Round 1 recommendations, reviewed benchmark cut score ranges, and revisited their cut scores.
- Round 3: Panelists reviewed impact data, discussed their Round 2 recommendations, and revisited their cut scores.
- Round 4: Panelists reviewed impact data, discussed their Round 3 recommendations, and recommended final cut scores.
- Vertical Articulation: In a cross-grade activity, a small group of panelists examined the system of cut scores and impact data to ensure coherence across the grades.

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

- Round 1: Panelists experienced the adaptive student assessment, studied the RALDs and OIB, studied the placement of the 2017 cut scores, and recommended cut scores.
- Round 2: Panelists reviewed impact data, discussed their Round 1 recommendations, and recommended final cut scores.
- Vertical Articulation: In a cross-grade activity, a small group of panelists examined the system of cut scores and impact data to ensure coherence across the grades.

7.4 RALD Revision

The ID Matching method requires clear RALDs that describe the KSAs of a student at a particular achievement level. Using those RALDs to identify a cut score ensures alignment of the assessment system and allows educators to focus on the RALDs 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 RALDs. The training and presenter were the same as was given to the original set of teachers who reviewed the Mathematics RALDs during their original development process. While the training given to participants was the same regarding the framework of RALD 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 (see Section 2.6.2).

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

7.5 Final Results

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

Table 7.1: Final Approved Cut Scores and Impact Data -ELA and Mathematics

		Cut Scores		Impact Data			
Content Area	**Grade**	**On Track**	**CCR**	**Developing**	**On Track**	**CCR**	**On Track + CCR**
ELA	3	2477	2557	46.7	37.3	15.9	53.2
	4	2500	2582	43.4	40.5	16.1	56.6
	5	2531	2599	48.6	35.3	16.1	51.4
	6	2543	2603	52.4	30.4	17.2	47.6
	7	2556	2630	52.4	32.7	14.9	47.6
	8	2561	2632	49.0	37.1	13.9	51.0
Mathematics	3	1190	1286	50.2	39.5	10.3	49.8
	4	1222	1317	50.2	39.4	10.4	49.8
	5	1236	1331	49.5	41.1	9.4	50.5
	6	1244	1342	45.2	44.6	10.3	54.9
	7	1247	1346	50.6	39.2	10.2	49.4
	8	1264	1365	49.4	41.1	9.5	50.6

8. Test Results

All students who took the online forms of the 2021 NSCAS Phase I Pilot were included in the test results. In 2021, students requiring a paper or Spanish assessment were exempt from taking the 2021 NSCAS assessments, therefore there were no paper-pencil or Spanish assessment 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. Results presented in this section are not from the state student data file that NDE received and may therefore differ slightly from the official state summary report due to ongoing resolution of student demographics and NTCs 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)). 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. Starting in 2021, new rule was applied for ELL students, with additional LEP status of Monitor: students having LEP status of 1 or 4 (i.e., ’Yes EL’ or ’Monitor’) are considered as ELL, while students having LEP status of 2 or 3 (i.e., ’Not EL’ or ’Formerly EL’) are considered as non-ELL.

As shown in these tables, 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. Among the students across grades, about 46% to 49% are eligible for FRL, 7-16% have LEP status, and 13-16% belong to at least one SPED category. For all three of these programs/categories, the participation rate is slightly lower for upper-grade students. In terms of the test accommodations, the calculator is used by most students (80% or higher for Grades 6-8 in Mathematics). In general, the answer choice eliminator was the most-used tool and TTS was the least-used tool across all grades and content areas. These percentages are very similar to last year.

Table 8.1: Number of Students Tested by Demographics - Grade 3

		ELA		Mathematics	
Demographic Sub-Group		N	%	N	%
Total N-Count		21,796	100.00	21,776	100.00
Gender	Female	10,631	48.78	10,617	48.76
	Male	11,165	51.22	11,159	51.24
Ethnicity	AI/AN	287	1.32	285	1.31
	Asian	698	3.20	696	3.20
	Black or African American	1,311	6.02	1,308	6.01
	Hispanic	4,221	19.37	4,216	19.36
	NH/PI	36	0.17	36	0.17
	White	14,233	65.31	14,223	65.32
	Two or More Races	1,007	4.62	1,010	4.64
FRL	Yes	10,820	49.65	10,818	49.68
	No	10,973	50.35	10,956	50.32
LEP	Yes	3,542	16.25	3,538	16.25
	No	18,251	83.75	18,236	83.75
SPED	Yes	3,601	16.52	3,579	16.44
	No	18,195	83.48	18,197	83.56
Universal Features & Accommodations	Answer Choice Eliminator	9,312	42.72	9,267	42.56
	Highlighter	9,587	43.99	7,649	35.13
	Line Reader	11,059	50.74	5,651	25.95
	Notepad	7,255	33.29	6,855	31.48
	Text-to-Speech (TTS)	3,663	16.81	3,486	16.01
	Zoom	5,411	24.83	3,144	14.44
	Ruler	-	-	5,060	23.24

*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.

Table 8.2: Number of Students Tested by Demographics - Grade 4

Demographic Sub-Group		ELA N	ELA %	Mathematics N	Mathematics %
Total N-Count		21,723	100.00	21,689	100.00
Gender	Female	10,577	48.69	10,562	48.70
	Male	11,146	51.31	11,127	51.30
Ethnicity	AI/AN	255	1.17	254	1.17
	Asian	658	3.03	655	3.02
	Black or African American	1,251	5.76	1,247	5.75
	Hispanic	4,288	19.74	4,281	19.74
	NH/PI	35	0.16	36	0.17
	White	14,288	65.78	14,270	65.80
	Two or More Races	947	4.36	945	4.36
FRL	Yes	10,734	49.42	10,725	49.45
	No	10,988	50.58	10,963	50.55
LEP	Yes	3,380	15.56	3,378	15.58
	No	18,342	84.44	18,310	84.42
SPED	Yes	3,672	16.90	3,642	16.79
	No	18,051	83.10	18,047	83.21
Universal Features & Accommodations	Answer Choice Eliminator	9,768	44.97	10,374	47.83
	Highlighter	8,797	40.50	7,161	33.02
	Line Reader	10,385	47.81	5,125	23.63
	Notepad	7,181	33.06	7,897	36.41
	Text-to-Speech (TTS)	3,480	16.02	3,068	14.15
	Zoom	5,424	24.97	2,809	12.95
	Calculator (basic)	-	-	160	0.74
	Protractor	-	-	5,879	27.11
	Reference Sheet	-	-	10,110	46.61

*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.

Table 8.3: Number of Students Tested by Demographics - Grade 5

Demographic Sub-Group		ELA N	ELA %	Mathematics N	Mathematics %	Science N	Science %
Total N-Count		22,232	100.00	22,199	100.00	22,201	100.00
Gender	Female	10,776	48.47	10,753	48.44	10,751	48.43
	Male	11,456	51.53	11,446	51.56	11,450	51.57
Ethnicity	AI/AN	281	1.26	279	1.26	280	1.26
	Asian	636	2.86	635	2.86	635	2.86
	Black or African American	1,358	6.11	1,353	6.10	1,355	6.10
	Hispanic	4,400	19.79	4,392	19.79	4,385	19.75
	NH/PI	34	0.15	34	0.15	33	0.15
	White	14,547	65.44	14,536	65.49	14,548	65.54
	Two or More Races	973	4.38	968	4.36	962	4.33
FRL	Yes	11,069	49.80	11,069	49.87	11,051	49.79
	No	11,160	50.20	11,128	50.13	11,146	50.21
LEP	Yes	3,337	15.01	3,332	15.01	3,323	14.97
	No	18,892	84.99	18,865	84.99	18,875	85.03
SPED	Yes	3,553	15.98	3,532	15.91	3,557	16.02
	No	18,679	84.02	18,667	84.09	18,644	83.98
Universal Features & Accommodations	Answer Choice Eliminator	9,317	41.91	10,320	46.49	5,808	26.16
	Highlighter	6,897	31.02	4,767	21.47	2,956	13.31
	Line Reader	9,212	41.44	3,504	15.78	2,922	13.16
	Notepad	6,107	27.47	6,637	29.90	3,175	14.30
	Text-to-Speech (TTS)	3,233	14.54	2,671	12.03	2,911	13.11
	Zoom	4,583	20.61	2,046	9.22	1,783	8.03
	Calculator (basic)	-	-	249	1.12	1,633	7.36
	Reference Sheet	-	-	11,667	52.56	-	-

*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.

Table 8.4: Number of Students Tested by Demographics - Grade 6

		ELA		Mathematics	
Demographic Sub-Group		N	%	N	%
Total N-Count		22,308	100.00	22,288	100.00
Gender	Female	10,853	48.65	10,849	48.68
	Male	11,455	51.35	11,439	51.32
Ethnicity	AI/AN	287	1.29	285	1.28
	Asian	584	2.62	586	2.63
	Black or African American	1,308	5.86	1,307	5.86
	Hispanic	4,511	20.22	4,508	20.23
	NH/PI	33	0.15	33	0.15
	White	14,670	65.76	14,656	65.76
	Two or More Races	914	4.10	913	4.10
FRL	Yes	10,931	49.00	10,939	49.08
	No	11,376	51.00	11,349	50.92
LEP	Yes	3,050	13.67	3,049	13.68
	No	19,257	86.33	19,239	86.32
SPED	Yes	3,428	15.37	3,419	15.34
	No	18,880	84.63	18,869	84.66
Universal Features & Accommodations	Answer Choice Eliminator	8,046	36.07	11,494	51.57
	Highlighter	5,541	24.84	4,258	19.10
	Line Reader	7,881	35.33	3,629	16.28
	Notepad	5,148	23.08	7,541	33.83
	Text-to-Speech (TTS)	2,497	11.19	1,838	8.25
	Zoom	4,212	18.88	1,854	8.32
	Calculator (basic)	-	-	16,017	71.86
	Reference Sheet	-	-	13,119	58.86

*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.

Table 8.5: Number of Students Tested by Demographics - Grade 7

		ELA		Mathematics	
Demographic Sub-Group		N	%	N	%
Total N-Count		22,106	100.00	22,071	100.00
Gender	Female	10,677	48.30	10,657	48.29
	Male	11,429	51.70	11,414	51.71
Ethnicity	AI/AN	269	1.22	268	1.21
	Asian	593	2.68	593	2.69
	Black or African American	1,279	5.79	1,276	5.78
	Hispanic	4,172	18.88	4,168	18.89
	NH/PI	35	0.16	35	0.16
	White	14,834	67.11	14,814	67.13
	Two or More Races	921	4.17	915	4.15
FRL	Yes	10,398	47.04	10,391	47.09
	No	11,705	52.96	11,677	52.91
LEP	Yes	2,312	10.46	2,309	10.46
	No	19,791	89.54	19,760	89.54
SPED	Yes	3,197	14.46	3,182	14.42
	No	18,909	85.54	18,889	85.58
Universal Features & Accommodations	Answer Choice Eliminator	6,707	30.34	9,351	42.37
	Highlighter	4,031	18.23	3,045	13.80
	Line Reader	5,921	26.78	3,004	13.61
	Notepad	3,701	16.74	6,339	28.72
	Text-to-Speech (TTS)	1,838	8.31	1,237	5.60
	Zoom	2,977	13.47	1,636	7.41
	Calculator (basic)	-	-	913	4.14
	Calculator (scientific)	-	-	18,073	81.89
	Reference Sheet	-	-	12,152	55.06

*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.

Table 8.6: Number of Students Tested by Demographics - Grade 8

Demographic Sub-Group		ELA N	ELA %	Mathematics N	Mathematics %	Science N	Science %
Total N-Count		20,708	100.00	20,672	100.00	20,693	100.00
Gender	Female	9,895	47.78	9,886	47.82	9,893	47.81
	Male	10,813	52.22	10,786	52.18	10,800	52.19
Ethnicity	AI/AN	279	1.35	276	1.34	280	1.35
	Asian	500	2.41	498	2.41	500	2.42
	Black or African American	1,197	5.78	1,195	5.78	1,195	5.78
	Hispanic	3,948	19.07	3,940	19.06	3,934	19.01
	NH/PI	38	0.18	39	0.19	38	0.18
	White	13,963	67.43	13,944	67.46	13,961	67.47
	Two or More Races	783	3.78	779	3.77	783	3.78
FRL	Yes	9,578	46.25	9,576	46.33	9,569	46.25
	No	11,130	53.75	11,095	53.67	11,122	53.75
LEP	Yes	1,549	7.48	1,554	7.52	1,545	7.47
	No	19,159	92.52	19,117	92.48	19,146	92.53
SPED	Yes	2,754	13.30	2,733	13.22	2,770	13.39
	No	17,954	86.70	17,939	86.78	17,923	86.61
Universal Features & Accommodations	Answer Choice Eliminator	5,014	24.21	8,613	41.67	2,505	12.11
	Highlighter	2,570	12.41	2,039	9.86	693	3.35
	Line Reader	3,752	18.12	2,249	10.88	629	3.04
	Notepad	2,171	10.48	4,561	22.06	824	3.98
	Text-to-Speech (TTS)	1,327	6.41	779	3.77	938	4.53
	Zoom	2,075	10.02	1,660	8.03	773	3.74
	Calculator (basic)	-	-	94	0.45	372	1.80
	Calculator (scientific)	-	-	16,796	81.25	-	-
	Reference Sheet	-	-	10,033	48.53	-	-

*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.

8.2 Administration Mode (Online vs. Paper-Pencil)

Table 8.7 shows the number of students who took the 2021 NSCAS assessments. The 2021 NSCAS assessments were administered online. Students requiring a paper or Spanish assessment were exempt from taking the 2021 NSCAS assessments, therefore there were no paper-pencil or Spanish assessments.

Table 8.7: Number of Students Tested by Administration Mode

Grade	Total Students	Online Students
ELA		
3	21,776	21,776
4	21,711	21,711
5	22,214	22,214
6	22,294	22,294
7	22,085	22,085
8	20,685	20,685
Mathematics		
3	21,761	21,761
4	21,675	21,675
5	22,187	22,187
6	22,274	22,274
7	22,048	22,048
8	20,657	20,657
Science		
5	22,201	22,201
8	20,693	20,693

8.3 Testing Time

Table 8.8, Table 8.9, and Table 8.10 present the number of minutes students took to complete the Spring 2021 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 20-120 minutes, the Mathematics test in 20-100 minutes, and the Science test in 10-60 minutes.

Table 8.8: Testing Time in Minutes - ELA

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
Time	N	%	N	%	N	%	N	%	N	%	N	%
<10 minutes	61	0.3	43	0.2	36	0.2	40	0.2	41	0.2	56	0.3
10 - <20	473	2.2	317	1.5	242	1.1	243	1.1	275	1.2	286	1.4
20 - <30	1,743	8.0	1,282	5.9	1,184	5.3	1,215	5.4	1,159	5.2	1,462	7.1
30 - <40	3,465	15.9	2,962	13.6	2,884	13.0	2,943	13.2	3,202	14.5	3,787	18.3
40 - <50	3,954	18.2	4,111	18.9	4,256	19.1	4,531	20.3	4,724	21.4	4,858	23.5
50 - <60	3,807	17.5	4,115	19.0	4,366	19.6	4,631	20.8	4,839	21.9	4,242	20.5
60 - <70	2,931	13.5	3,238	14.9	3,395	15.3	3,491	15.7	3,336	15.1	2,677	12.9
70 - <80	1,926	8.8	2,217	10.2	2,346	10.6	2,192	9.8	1,967	8.9	1,516	7.3
80 - <90	1,335	6.1	1,388	6.4	1,403	6.3	1,288	5.8	1,108	5.0	791	3.8
90 - <100	802	3.7	822	3.8	838	3.8	767	3.4	605	2.7	448	2.2
100 - <110	494	2.3	467	2.2	466	2.1	406	1.8	359	1.6	236	1.1
110 - <120	296	1.4	295	1.4	319	1.4	209	0.9	181	0.8	143	0.7
120 - <130	172	0.8	166	0.8	188	0.8	141	0.6	134	0.6	90	0.4
130 - <140	123	0.6	123	0.6	117	0.5	80	0.4	61	0.3	33	0.2
140 - <150	58	0.3	63	0.3	64	0.3	36	0.2	36	0.2	26	0.1
150 - <160	41	0.2	31	0.1	48	0.2	33	0.1	20	0.1	14	0.1

Table 8.8: Testing Time in Minutes - ELA, cont.

160 - <170	32	0.1	25	0.1	26	0.1	17	0.1	13	0.1	11	0.1
170 - <180	18	0.1	19	0.1	17	0.1	12	0.1	10	0.0	5	0.0
>=180 minutes	53	0.2	30	0.1	30	0.1	25	0.1	23	0.1	18	0.1
Total	21,784	100.0	21,714	100.0	22,225	100.0	22,300	100.0	22,093	100.0	20,699	100.0

Table 8.9: Testing Time in Minutes - Mathematics

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
Time	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
<10 minutes	30	0.1	44	0.2	20	0.1	37	0.2	55	0.2	45	0.2
10 - <20	682	3.1	404	1.9	328	1.5	286	1.3	291	1.3	316	1.5
20 - <30	3,769	17.3	2,344	10.8	2,441	11.0	1,346	6.0	1,237	5.6	1,295	6.3
30 - <40	5,909	27.2	4,723	21.8	5,252	23.7	3,328	14.9	2,980	13.5	3,361	16.3
40 - <50	4,803	22.1	4,913	22.7	5,334	24.0	4,679	21.0	4,396	19.9	4,541	22.0
50 - <60	2,941	13.5	3,666	16.9	3,883	17.5	4,344	19.5	4,259	19.3	4,146	20.1
60 - <70	1,614	7.4	2,339	10.8	2,224	10.0	3,203	14.4	3,460	15.7	2,884	14.0
70 - <80	886	4.1	1,314	6.1	1,170	5.3	2,044	9.2	2,113	9.6	1,791	8.7
80 - <90	474	2.2	794	3.7	630	2.8	1,238	5.6	1,328	6.0	1,019	4.9
90 - <100	269	1.2	436	2.0	396	1.8	723	3.2	782	3.5	519	2.5
100 - <110	166	0.8	262	1.2	219	1.0	424	1.9	434	2.0	312	1.5
110 - <120	74	0.3	179	0.8	126	0.6	248	1.1	268	1.2	163	0.8
120 - <130	50	0.2	89	0.4	74	0.3	158	0.7	156	0.7	104	0.5
130 - <140	36	0.2	52	0.2	34	0.2	88	0.4	117	0.5	66	0.3
140 - <150	22	0.1	37	0.2	29	0.1	56	0.3	75	0.3	42	0.2
150 - <160	11	0.1	30	0.1	16	0.1	34	0.2	44	0.2	21	0.1
160 - <170	9	0.0	16	0.1	7	0.0	17	0.1	16	0.1	8	0.0
170 - <180	5	0.0	10	0.0	5	0.0	8	0.0	21	0.1	9	0.0
>=180 minutes	13	0.1	28	0.1	10	0.0	19	0.1	26	0.1	24	0.1
Total	21,763	100.0	21,680	100.0	22,198	100.0	22,280	100.0	22,058	100.0	20,666	100.0

Table 8.10: Testing Time in Minutes - Science

	Grade 5		Grade 8	
Time	N	%	N	%
<10 minutes	207	0.9	351	1.7
10 - <20	5,051	22.8	6,534	31.6
20 - <30	9,125	41.1	9,222	44.6
30 - <40	4,995	22.5	3,307	16.0
40 - <50	1,848	8.3	875	4.2
50 - <60	640	2.9	256	1.2
60 - <70	206	0.9	80	0.4
70 - <80	77	0.3	33	0.2
80 - <90	31	0.1	16	0.1
90 - <100	11	0.0	11	0.1
100 - <110	6	0.0	4	0.0
110 - <120	2	0.0	0	0.0
120 - <130	1	0.0	1	0.0
130 - <140	0	0.0	1	0.0
140 - <150	1	0.0	1	0.0
150 - <160	0	0.0	0	0.0
160 - <170	0	0.0	0	0.0
170 - <180	0	0.0	0	0.0
>=180 minutes	0	0.0	1	0.0
Total	22,201	100.0	20,693	100.0

8.4 Achievement Level Distributions

Table 8.11 presents the achievement level distributions for the Spring 2021 NSCAS assessments. Appendix D provides the achievement level distributions by demographic group. For ELA, 46-55% of students are at Developing and 44-53% of students are at On Track or CCR Benchmark. For Mathematics, 52-54% of students are at Developing and 45-47% of students are at On Track or CCR Benchmark.

Table 8.11: Achievement Level Distributions

		Level 3*		Level 2*		Level 1*		Level 2 + Level 1	
Grade	**Total N**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
ELA									
3	21,776	10,856	49.9	7,807	35.9	3,113	14.3	10,920	50.1
4	21,711	10,046	46.3	7,935	36.5	3,730	17.2	11,665	53.7
5	22,214	12,015	54.1	6,943	31.3	3,256	14.7	10,199	45.9
6	22,294	12,096	54.3	6,697	30.0	3,501	15.7	10,198	45.7
7	22,085	12,215	55.3	7,886	35.7	1,984	9.0	9,870	44.7
8	20,685	10,213	49.4	7,803	37.7	2,669	12.9	10,472	50.6
Mathematics									
3	21,761	11,495	52.8	8,222	37.8	2,044	9.4	10,266	47.2
4	21,675	11,770	54.3	8,143	37.6	1,762	8.1	9,905	45.7
5	22,187	12,068	54.4	8,447	38.1	1,672	7.5	10,119	45.6
6	22,274	11,786	52.9	8,684	39.0	1,804	8.1	10,488	47.1
7	22,048	11,842	53.7	8,457	38.4	1,749	7.9	10,206	46.3
8	20,657	11,287	54.6	7,784	37.7	1,586	7.7	9,370	45.4

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

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. Table 8.13 presents the descriptive statistics for the raw scores of Science field test by form. 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

				Percentiles						
Grade	**N-Count**	**Mean**	**SD**	**P5**	**P10**	**P25**	**P50**	**P75**	**P90**	**P95**
ELA										
3	21,776	2467.07	87.30	2304	2346	2412	2477	2531	2571	2594
4	21,711	2501.13	84.00	2346	2389	2451	2507	2562	2602	2625
5	22,214	2514.52	81.92	2365	2404	2463	2523	2570	2616	2638
6	22,294	2526.95	79.31	2376	2417	2483	2536	2582	2619	2641
7	22,085	2537.68	76.09	2393	2435	2494	2547	2589	2625	2647
8	20,685	2555.23	74.19	2418	2461	2514	2562	2604	2641	2665
Mathematics										
3	21,761	1183.17	78.88	1052	1081	1129	1184	1235	1282	1314
4	21,675	1212.60	74.40	1091	1117	1162	1213	1261	1307	1337
5	22,187	1228.97	72.08	1113	1141	1182	1227	1274	1317	1349
6	22,274	1237.62	73.71	1113	1144	1191	1238	1282	1332	1364
7	22,048	1245.78	68.31	1138	1165	1203	1241	1285	1332	1369
8	20,657	1259.15	71.79	1145	1170	1212	1256	1304	1352	1382

Table 8.13: Raw Score Descriptive Statistics for Science Fixed Forms

					Percentiles						
Grade	Form	N-Count	Mean	SD	P5	P10	P25	P50	P75	P90	P95
5	A	4,233	11.70	4.29	4	6	8	12	15	17	18
	B	3,056	9.15	4.14	3	4	6	9	12	15	16
	C	3,580	11.59	4.50	4	5	8	12	15	17	18
	D	4,280	10.95	3.55	5	6	8	11	14	15	16
	E	3,001	12.88	3.92	6	7	10	13	16	18	18
	F	4,051	10.02	4.61	3	4	6	10	14	16	18
8	A	3,067	7.93	2.97	3	4	6	8	10	12	13
	B	4,242	7.78	4.16	1	2	5	7	11	14	15
	C	3,900	10.42	4.49	3	5	7	10	14	16	18
	D	3,352	6.91	3.11	2	3	5	7	9	11	12
	E	3,069	4.88	2.86	1	1	3	4	7	9	10
	F	3,063	8.22	3.90	2	3	5	8	11	14	15

8.6 Reporting Category Correlations

For the Spring 2021 assessments, reporting category correlations were not calculated because reporting category scores were not reported.

8.7 Correlations with MAP Growth

Table 8.14 presents the correlation coefficients between MAP Growth and NSCAS scores for students who took both tests in Spring 2021. As shown in the table, the correlation coefficients range from 0.78 to 0.82 for ELA/Reading, 0.75 to 0.79 for ELA/Language Usage, and 0.85 to 0.88 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 8.14: Correlation and Descriptive Statistics of NSCAS and MAP Growth Scores

			NSCAS*				MAP Growth*			
Grade	**N**	**r**	**Mean**	**SD**	**Min.**	**Max.**	**Mean**	**SD**	**Min.**	**Max.**
ELA/Reading										
3	18,441	0.81	2468	85.67	2220	2749	199	15.76	135	245
4	15,462	0.82	2500	82.70	2253	2838	206	15.28	140	260
5	15,760	0.81	2515	80.10	2283	2791	212	14.93	145	262
6	16,242	0.81	2525	77.29	2293	2863	215	15.08	156	261
7	14,872	0.79	2534	75.14	2303	2792	218	15.45	154	267
8	13,503	0.78	2553	72.58	2310	2788	221	15.51	151	274
ELA/Language Usage										
3	3,925	0.76	2478	78.15	2223	2699	201	12.41	151	241
4	4,055	0.77	2504	76.02	2254	2749	207	11.88	153	248
5	4,064	0.76	2520	73.00	2283	2724	213	11.69	150	247
6	5,120	0.79	2533	74.51	2293	2799	216	12.73	148	253
7	5,069	0.77	2544	70.10	2304	2767	219	12.93	157	263
8	4,924	0.75	2561	69.79	2313	2787	222	13.17	137	275
Mathematics										
3	15,608	0.88	1184	77.05	1000	1470	202	14.22	138	266
4	15,548	0.86	1213	73.98	1013	1500	211	15.64	139	269
5	15,897	0.87	1227	70.82	1020	1510	219	17.21	144	289
6	15,687	0.86	1236	71.75	1033	1530	223	16.78	146	288
7	14,344	0.85	1244	68.09	1040	1540	228	17.85	138	307
8	13,316	0.86	1257	70.83	1050	1550	233	19.15	136	316

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

8.8 Score Differences

To evaluate student's annual progress toward college and career readiness, the data were merged with the previous year's data by students who advanced by one grade using student ID and grade (i.e., students who repeated a grade or skipped a grade were not included). The Spring 2020 NSCAS testing was cancelled due to COVID-19 and the Spring 2021 NSCAS Phase I Pilot assessments are different from the Spring 2019 NSCAS General Summative assessments. Therefore, score differences were not analyzed.

9. Reliability

The *Standards* refer to reliability as the ”consistency of scores across replications of a testing procedure” (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 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 2021 NSCAS assessments was examined through analysis of measurement error in simulated and operational conditions, as follows:

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

Combined, these data provide several ways of looking at the reliability of the NSCAS 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 and Table 9.2 present marginal reliabilities of scale scores by grade and reporting category for ELA and Mathematics, respectively. Marginal reliability estimates for the total scores are well above 0.80 (0.84 or higher), which is typically considered the minimally acceptable level of reliability. Because reliability estimates 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 9.3, 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.

Table 9.1: Marginal Reliability of Scale Scores–ELA

Grade	N	Total Score	Reading Vocabulary	Reading Comprehension	Writing Skills
3	21784	0.88	0.45	0.82	0.58
4	21714	0.88	0.35	0.81	0.58
5	22225	0.87	0.36	0.80	0.59
6	22300	0.87	0.40	0.80	0.58
7	22093	0.86	0.38	0.77	0.60
8	20699	0.84	0.26	0.77	0.52

Table 9.2: Marginal Reliability of Scale Scores–Mathematics

Grade	N	Total Score	Number	Algebra	Geomery	Data
3	21763	0.92	0.82	0.55	0.68	0.55
4	21680	0.91	0.78	0.64	0.64	0.54
5	22198	0.90	0.78	0.62	0.56	0.47
6	22280	0.91	0.68	0.78	0.58	0.49
7	22058	0.89	0.59	0.75	0.56	0.50
8	20666	0.90	0.67	0.69	0.72	0.45

Table 9.3: Marginal Reliability: Variance

				Deciles									
Grade	**N**	**Variance**	**Overall**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
ELA													
3	21784	7629.19	0.88	0.84	0.87	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.88
4	21714	7057.88	0.88	0.85	0.88	0.88	0.89	0.89	0.89	0.89	0.88	0.87	0.84
5	22225	6713.09	0.87	0.83	0.86	0.87	0.88	0.88	0.89	0.89	0.89	0.88	0.85
6	22300	6292.27	0.87	0.81	0.86	0.88	0.88	0.89	0.88	0.88	0.88	0.87	0.85
7	22093	5793.66	0.86	0.80	0.85	0.86	0.87	0.87	0.87	0.87	0.87	0.86	0.83
8	20699	5515.28	0.84	0.79	0.84	0.85	0.86	0.86	0.86	0.86	0.86	0.85	0.82
Mathematics													
3	21763	6222.97	0.92	0.90	0.91	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.89
4	21680	5539.05	0.91	0.88	0.90	0.90	0.91	0.91	0.91	0.92	0.91	0.91	0.90
5	22198	5202.88	0.90	0.88	0.90	0.90	0.91	0.91	0.91	0.91	0.91	0.90	0.85
6	22280	5435.94	0.91	0.88	0.90	0.90	0.91	0.91	0.91	0.91	0.91	0.91	0.90
7	22058	4670.35	0.89	0.85	0.88	0.88	0.89	0.89	0.90	0.90	0.90	0.90	0.89
8	20666	5158.11	0.90	0.87	0.89	0.90	0.90	0.90	0.91	0.91	0.91	0.91	0.89

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 9.4 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), 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 9.5 then presents the overall and by-decile CSEM. The overall CSEM is slightly higher for ELA (from 28.6 to 29.8) than for Mathematics (from 22.7 to 22.8). CSEM is also relatively similar in the middle (between Deciles 2 and 9), which is consistent with reliability results. Appendix F presents scatterplots for scale score CSEM for each content area and grade.

Table 9.4: CSEMs at the Proficient Cut Scores

	Level 3 - Level 2 Cut Scores			Level 2 - Level 1 Cut Scores		
Grade	**N**	**Mean CSEM**	**SD**	**N**	**Mean CSEM**	**SD**
ELA						
3	2247	28.5	1.0	1473	29.4	1.0
4	2178	28.0	0.9	1757	30.0	0.8
5	2488	27.8	0.7	1352	28.4	0.8
6	2880	27.0	0.9	1888	28.2	0.7
7	2803	27.3	0.7	1223	29.2	0.9
8	2699	27.9	0.6	1443	28.7	0.9
Mathematics						
3	2317	21.7	0.8	895	23.3	0.9
4	2411	21.7	0.9	849	22.4	0.9
5	2646	21.3	0.8	746	23.7	0.8
6	2737	22.0	0.9	802	22.0	1.1
7	2945	22.0	0.9	677	21.7	0.9
8	2609	21.8	0.8	758	21.9	0.9

Note: Level 3 = Developing, Level 2 = On Track, Level 1 = CCR Benchmark.

Table 9.5: Mean CSEMs by Deciles

		Mean CSEM by Deciles									
Grade	**Mean CSEM**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
ELA											
3	29.8	34.8	30.9	29.5	28.8	28.5	28.5	28.6	28.9	29.3	30.5
4	29.6	32.5	29.6	28.6	28.1	28.0	28.1	28.4	29.1	30.2	33.5
5	29.4	33.8	30.9	29.8	28.7	28.1	27.8	27.4	27.5	28.4	31.9
6	28.6	34.1	29.3	27.7	27.1	26.9	27.0	27.4	27.8	28.2	30.6
7	28.9	33.8	29.6	28.4	27.9	27.4	27.3	27.4	27.8	28.4	31.1
8	29.2	33.8	29.9	28.7	28.2	28.0	27.9	27.8	28.0	28.6	31.6
Mathematics											
3	22.8	24.4	23.1	22.6	22.1	21.8	21.7	21.7	21.9	22.6	26.1
4	22.8	26.0	23.8	23.0	22.4	22.0	21.7	21.7	21.8	22.0	23.5
5	22.7	24.5	23.1	22.6	21.9	21.5	21.3	21.3	21.5	22.4	27.3
6	22.7	25.5	23.5	22.9	22.4	22.2	22.0	21.9	21.7	21.7	23.0
7	22.7	26.0	23.7	23.2	22.8	22.3	21.9	21.6	21.4	21.4	22.9
8	22.7	25.4	23.6	23.1	22.7	22.3	21.8	21.6	21.3	21.4	23.2

9.3 Classification Accuracy

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

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

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

Table 9.6 and Table 9.7 present the classification accuracy results by grade, achievement level, and reporting category. Overall classification accuracy ranges from 0.794 (ELA Grade 8) to 0.877 (Mathematics Grade 4). 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 results suggests that accurate level classifications are being made. Overall classification accuracy by achievement level ranges from 0.637 (ELA Grade 6 On Track) to 0.921 (Mathematics Grade 4 Developing).

Table 9.6: Classification Accuracy by Achievement Level and Reporting Category - ELA

	Achievement			Expected Proportion			Class	Overall
Grade	Level	N	%	L3	L2	L1	Acc.	Class. Acc.
Overall								
3	Developing	10859	0.50	0.45	0.05	0.00	0.902	0.820
	On Track	7807	0.36	0.06	0.26	0.05	0.718	
	CCR Benchmark	3113	0.14	0.00	0.03	0.11	0.790	
4	Developing	10047	0.46	0.41	0.05	0.00	0.892	0.812
	On Track	7935	0.37	0.06	0.27	0.05	0.726	
	CCR Benchmark	3730	0.17	0.00	0.04	0.13	0.779	
5	Developing	12016	0.54	0.49	0.05	0.00	0.900	0.818
	On Track	6943	0.31	0.06	0.21	0.04	0.684	

Table 9.6: Classification Accuracy - ELA, cont.

	CCR Benchmark	3256	0.15	0.00	0.03	0.12	0.796	
6	Developing	12097	0.54	0.48	0.06	0.00	0.890	0.797
	On Track	6697	0.30	0.06	0.19	0.05	0.637	
	CCR Benchmark	3501	0.16	0.00	0.03	0.12	0.783	
7	Developing	12217	0.55	0.49	0.06	0.00	0.892	0.812
	On Track	7886	0.36	0.07	0.25	0.04	0.703	
	CCR Benchmark	1984	0.09	0.00	0.02	0.07	0.756	
8	Developing	10217	0.49	0.43	0.06	0.00	0.877	0.794
	On Track	7803	0.38	0.07	0.26	0.05	0.690	
	CCR Benchmark	2669	0.13	0.00	0.03	0.10	0.783	
Reading Vocabulary								
3	Developing	11441	0.53	0.43	0.08	0.00	0.823	0.691
	On Track	5264	0.24	0.08	0.10	0.07	0.393	
	CCR Benchmark	5040	0.23	0.00	0.05	0.16	0.703	
4	Developing	10636	0.49	0.40	0.07	0.00	0.822	0.673
	On Track	5654	0.26	0.08	0.10	0.08	0.379	
	CCR Benchmark	5394	0.25	0.00	0.05	0.17	0.687	
5	Developing	12510	0.56	0.47	0.07	0.00	0.826	0.694
	On Track	4615	0.21	0.07	0.07	0.06	0.341	
	CCR Benchmark	5060	0.23	0.00	0.04	0.16	0.689	
6	Developing	11969	0.54	0.44	0.07	0.00	0.814	0.680
	On Track	5066	0.23	0.08	0.07	0.08	0.317	
	CCR Benchmark	5253	0.24	0.00	0.04	0.17	0.725	
7	Developing	11648	0.53	0.43	0.08	0.00	0.813	0.673
	On Track	6011	0.27	0.09	0.11	0.08	0.397	
	CCR Benchmark	4406	0.20	0.00	0.04	0.14	0.680	
8	Developing	10635	0.51	0.42	0.08	0.00	0.809	0.661
	On Track	4865	0.24	0.08	0.08	0.08	0.328	
	CCR Benchmark	5178	0.25	0.00	0.05	0.17	0.672	
Reading Comprehension								
3	Developing	10918	0.50	0.44	0.06	0.00	0.882	0.780
	On Track	7187	0.33	0.07	0.21	0.06	0.627	
	CCR Benchmark	3671	0.17	0.00	0.04	0.13	0.775	
4	Developing	9710	0.45	0.39	0.06	0.00	0.868	0.767
	On Track	7613	0.35	0.07	0.22	0.06	0.638	
	CCR Benchmark	4388	0.20	0.00	0.05	0.16	0.767	
5	Developing	12077	0.54	0.48	0.06	0.00	0.879	0.772
	On Track	6514	0.29	0.07	0.17	0.06	0.580	
	CCR Benchmark	3623	0.16	0.00	0.04	0.12	0.761	
6	Developing	12068	0.54	0.47	0.07	0.00	0.874	0.762
	On Track	5899	0.27	0.07	0.14	0.06	0.525	
	CCR Benchmark	4327	0.19	0.00	0.04	0.15	0.773	
7	Developing	12519	0.57	0.50	0.07	0.00	0.877	0.771
	On Track	6839	0.31	0.07	0.18	0.06	0.587	
	CCR Benchmark	2727	0.12	0.00	0.03	0.09	0.748	
	Developing	10361	0.50	0.43	0.07	0.00	0.860	

Table 9.6: Classification Accuracy - ELA, cont.

8	On Track	7200	0.35	0.08	0.20	0.07	0.586	0.749
	CCR Benchmark	3124	0.15	0.00	0.04	0.11	0.755	
Writing Skills								
3	Developing	11073	0.51	0.42	0.08	0.00	0.831	0.704
	On Track	6815	0.31	0.09	0.15	0.08	0.489	
	CCR Benchmark	3851	0.18	0.00	0.04	0.13	0.723	
4	Developing	10606	0.49	0.40	0.08	0.00	0.824	0.697
	On Track	7257	0.33	0.09	0.17	0.08	0.500	
	CCR Benchmark	3835	0.18	0.00	0.04	0.13	0.718	
5	Developing	11403	0.51	0.43	0.08	0.00	0.829	0.707
	On Track	6340	0.29	0.08	0.13	0.07	0.469	
	CCR Benchmark	4439	0.20	0.00	0.04	0.15	0.735	
6	Developing	12322	0.55	0.46	0.08	0.00	0.835	0.712
	On Track	5695	0.26	0.08	0.11	0.07	0.434	
	CCR Benchmark	4267	0.19	0.00	0.04	0.14	0.728	
7	Developing	12093	0.55	0.45	0.09	0.00	0.825	0.707
	On Track	7435	0.34	0.09	0.17	0.08	0.504	
	CCR Benchmark	2541	0.12	0.00	0.03	0.09	0.739	
8	Developing	9725	0.47	0.38	0.08	0.00	0.804	0.682
	On Track	7066	0.34	0.10	0.17	0.08	0.482	
	CCR Benchmark	3886	0.19	0.00	0.04	0.14	0.739	

Table 9.7: Classification Accuracy by Achievement Level and Reporting Category - Mathematics

	Achievement			Expected Proportion			Class	Overall
Grade	**Level**	**N**	%	**L3**	**L2**	**L1**	**Acc.**	**Class. Acc.**
Overall								
3	Developing	11496	0.53	0.49	0.04	0.00	0.919	0.874
	On Track	8222	0.38	0.05	0.31	0.02	0.823	
	CCR Benchmark	2044	0.09	0.00	0.02	0.08	0.830	
4	Developing	11772	0.54	0.50	0.04	0.00	0.921	0.877
	On Track	8143	0.38	0.05	0.31	0.02	0.822	
	CCR Benchmark	1762	0.08	0.00	0.01	0.07	0.840	
5	Developing	12072	0.54	0.50	0.05	0.00	0.910	0.871
	On Track	8447	0.38	0.05	0.31	0.02	0.822	
	CCR Benchmark	1672	0.08	0.00	0.01	0.06	0.840	
6	Developing	11788	0.53	0.48	0.05	0.00	0.909	0.869
	On Track	8684	0.39	0.05	0.32	0.02	0.821	
	CCR Benchmark	1804	0.08	0.00	0.01	0.07	0.840	
7	Developing	11844	0.54	0.48	0.06	0.00	0.890	0.862
	On Track	8457	0.38	0.05	0.32	0.02	0.823	
	CCR Benchmark	1749	0.08	0.00	0.01	0.07	0.861	
8	Developing	11289	0.55	0.49	0.05	0.00	0.905	0.869
	On Track	7784	0.38	0.05	0.31	0.02	0.825	
	CCR Benchmark	1586	0.08	0.00	0.01	0.06	0.831	

Table 9.7: Classification Accuracy - Mathematics, cont.

Number								
3	Developing	11557	0.53	0.47	0.06	0.00	0.879	0.807
	On Track	7174	0.33	0.07	0.23	0.04	0.688	
	CCR Benchmark	3026	0.14	0.00	0.03	0.11	0.813	
4	Developing	11809	0.55	0.48	0.07	0.00	0.875	0.801
	On Track	7473	0.35	0.07	0.24	0.04	0.693	
	CCR Benchmark	2391	0.11	0.00	0.03	0.09	0.773	
5	Developing	11673	0.53	0.46	0.07	0.00	0.867	0.794
	On Track	7966	0.36	0.07	0.25	0.04	0.694	
	CCR Benchmark	2546	0.12	0.00	0.03	0.09	0.774	
6	Developing	11609	0.52	0.44	0.08	0.00	0.843	0.757
	On Track	8159	0.37	0.08	0.24	0.05	0.642	
	CCR Benchmark	2501	0.11	0.00	0.03	0.08	0.741	
7	Developing	11355	0.52	0.42	0.09	0.00	0.818	0.736
	On Track	7979	0.36	0.08	0.22	0.06	0.616	
	CCR Benchmark	2691	0.12	0.00	0.03	0.09	0.746	
8	Developing	11381	0.55	0.47	0.08	0.00	0.846	0.770
	On Track	6841	0.33	0.07	0.22	0.05	0.650	
	CCR Benchmark	2433	0.12	0.00	0.03	0.09	0.754	
Algebra								
3	Developing	11323	0.52	0.44	0.08	0.00	0.839	0.724
	On Track	7306	0.34	0.08	0.19	0.07	0.554	
	CCR Benchmark	3118	0.14	0.00	0.04	0.10	0.706	
4	Developing	11349	0.52	0.44	0.08	0.00	0.842	0.745
	On Track	7475	0.35	0.08	0.21	0.06	0.597	
	CCR Benchmark	2843	0.13	0.00	0.03	0.10	0.748	
5	Developing	11721	0.53	0.44	0.09	0.00	0.830	0.742
	On Track	7837	0.35	0.08	0.22	0.06	0.609	
	CCR Benchmark	2627	0.12	0.00	0.03	0.09	0.754	
6	Developing	11460	0.52	0.44	0.07	0.00	0.860	0.795
	On Track	8485	0.38	0.07	0.27	0.04	0.703	
	CCR Benchmark	2327	0.10	0.00	0.02	0.08	0.808	
7	Developing	11659	0.53	0.45	0.08	0.00	0.849	0.788
	On Track	8372	0.38	0.08	0.27	0.04	0.703	
	CCR Benchmark	2010	0.09	0.00	0.02	0.07	0.791	
8	Developing	11072	0.54	0.45	0.08	0.00	0.843	0.772
	On Track	7080	0.34	0.07	0.23	0.04	0.662	
	CCR Benchmark	2501	0.12	0.00	0.03	0.09	0.769	
Geometry								
3	Developing	11619	0.53	0.45	0.08	0.00	0.850	0.759
	On Track	7774	0.36	0.08	0.22	0.06	0.627	
	CCR Benchmark	2365	0.11	0.00	0.03	0.08	0.743	
4	Developing	12578	0.58	0.50	0.08	0.00	0.855	0.759
	On Track	6802	0.31	0.07	0.19	0.05	0.596	
	CCR Benchmark	2289	0.11	0.00	0.03	0.08	0.717	
	Developing	12589	0.57	0.48	0.09	0.00	0.842	

Table 9.7: Classification Accuracy - Mathematics, cont.

5	On Track	7177	0.32	0.08	0.19	0.06	0.580	0.741
	CCR Benchmark	2404	0.11	0.00	0.03	0.08	0.694	
6	Developing	11919	0.54	0.45	0.09	0.00	0.832	0.733
	On Track	7237	0.33	0.08	0.19	0.06	0.569	
	CCR Benchmark	3114	0.14	0.00	0.03	0.10	0.736	
7	Developing	12034	0.55	0.45	0.10	0.00	0.817	0.746
	On Track	7412	0.34	0.08	0.21	0.05	0.626	
	CCR Benchmark	2577	0.12	0.00	0.03	0.09	0.761	
8	Developing	11575	0.56	0.48	0.08	0.00	0.854	0.787
	On Track	7066	0.34	0.07	0.23	0.04	0.681	
	CCR Benchmark	2013	0.10	0.00	0.02	0.08	0.784	
Data								
3	Developing	11194	0.52	0.43	0.08	0.00	0.831	0.728
	On Track	7294	0.34	0.08	0.20	0.06	0.582	
	CCR Benchmark	3265	0.15	0.00	0.04	0.11	0.700	
4	Developing	11160	0.52	0.42	0.09	0.00	0.812	0.721
	On Track	7122	0.33	0.08	0.19	0.06	0.571	
	CCR Benchmark	3376	0.16	0.00	0.04	0.12	0.737	
5	Developing	11322	0.51	0.42	0.08	0.00	0.828	0.706
	On Track	7504	0.34	0.09	0.18	0.07	0.541	
	CCR Benchmark	3351	0.15	0.00	0.04	0.10	0.662	
6	Developing	12547	0.56	0.47	0.09	0.00	0.826	0.731
	On Track	7219	0.32	0.08	0.19	0.06	0.571	
	CCR Benchmark	2504	0.11	0.00	0.03	0.08	0.723	
7	Developing	11349	0.52	0.41	0.10	0.00	0.794	0.717
	On Track	8032	0.37	0.09	0.22	0.06	0.597	
	CCR Benchmark	2652	0.12	0.00	0.03	0.09	0.750	
8	Developing	9798	0.48	0.39	0.09	0.00	0.811	0.696
	On Track	8415	0.41	0.10	0.23	0.08	0.564	
	CCR Benchmark	2436	0.12	0.00	0.03	0.08	0.686	

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):

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

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:

$$SEM = SD\sqrt{1 - \hat{\alpha}} \tag{9.2}$$

where SD represents the standard deviation of the raw score distribution and $\hat{\alpha}$ represents Cronbach's alpha. The overall SEM is expressed in raw score units and is a test-level statistic. Table 9.8 presents Cronbach's alpha reliability coefficients and the SEMs for the Science fixed forms. Table 9.9 presents Cronbach's alpha reliability coefficients by demographics for the Science fixed forms, along with the SEMs.

Table 9.8: Cronbach's Alpha for Science Fixed Forms

Grade	Form	#Items	N	Reliability	SEM
5	A	20	4,233	0.81	1.87
	B	20	3,056	0.78	1.94
	C	20	3,580	0.81	1.96
	D	20	4,280	0.73	1.84
	E	20	3,001	0.80	1.75
	F	20	4,051	0.82	1.96
8	A	17	3,067	0.70	1.63
	B	17	4,242	0.74	2.12
	C	17	3,900	0.77	2.15
	D	17	3,352	0.68	1.76
	E	17	3,069	0.69	1.59
	F	17	3,063	0.75	1.95

Table 9.9: Cronbach’s Alpha by Demographics for Science Fixed Forms

Form	Demographic Sub-Group*		#Items	N	Reliability	SEM
Grade 5						
A	Overall	Overall	20	4,233	0.81	1.87
	Gender	Female	20	2,057	0.79	1.91
		Male	20	2,176	0.82	1.87
	Ethnicity	AI/AN	20	57	0.65	2.03
		Asian	20	113	0.82	1.85
		Black or African American	20	236	0.70	1.99
		Hispanic	20	802	0.76	1.99
		NH/PI	20	3	0.82	1.71
		White	20	2,840	0.79	1.85
		Two or More Races	20	182	0.80	1.91
	FRL	Yes	20	2,088	0.78	1.97
		No	20	2,145	0.78	1.80
	LEP	Yes	20	590	0.75	1.97
		No	20	3,643	0.80	1.88
	SPED	Yes	20	666	0.76	1.98
		No	20	3,567	0.79	1.87
B	Overall	Overall	20	3,056	0.78	1.94
	Gender	Female	20	1,491	0.76	1.94
		Male	20	1,565	0.80	1.92
	Ethnicity	AI/AN	20	32	0.66	2.01
		Asian	20	88	0.82	1.90
		Black or African American	20	196	0.76	1.85
		Hispanic	20	640	0.70	1.93
		NH/PI	20	3	0.91	1.83
		White	20	1,961	0.78	1.91
		Two or More Races	20	136	0.77	1.92
	FRL	Yes	20	1,561	0.74	1.94
		No	20	1,495	0.77	1.93
	LEP	Yes	20	496	0.71	1.92
		No	20	2,560	0.78	1.93
	SPED	Yes	20	485	0.74	1.86
		No	20	2,571	0.77	1.94
C	Overall	Overall	20	3,580	0.81	1.96
	Gender	Female	20	1,721	0.79	1.98
		Male	20	1,859	0.82	1.97
	Ethnicity	AI/AN	20	37	0.69	2.05
		Asian	20	109	0.83	1.96
		Black or African American	20	239	0.79	1.97
		Hispanic	20	712	0.77	2.04
		NH/PI	20	5	0.85	1.89
		White	20	2,331	0.78	1.99
		Two or More Races	20	146	0.77	2.00
	FRL	Yes	20	1,788	0.78	2.00
		No	20	1,791	0.77	1.96

Table 9.9: Cronbach's Alpha by Demographics for Science Fixed Forms, cont.

	LEP	Yes	20	557	0.76	2.02
		No	20	3,022	0.80	1.97
	SPED	Yes	20	607	0.76	2.00
		No	20	2,973	0.79	1.97
D	Overall	Overall	20	4,280	0.73	1.84
	Gender	Female	20	2,068	0.71	1.85
		Male	20	2,212	0.74	1.86
	Ethnicity	AI/AN	20	50	0.72	1.89
		Asian	20	121	0.74	1.86
		Black or African American	20	252	0.71	1.88
		Hispanic	20	846	0.72	1.89
		NH/PI	20	6	0.76	1.87
		White	20	2,829	0.69	1.83
		Two or More Races	20	176	0.73	1.82
	FRL	Yes	20	2,149	0.71	1.89
		No	20	2,131	0.67	1.81
	LEP	Yes	20	610	0.72	1.90
		No	20	3,670	0.71	1.85
	SPED	Yes	20	676	0.71	1.88
		No	20	3,604	0.71	1.83
E	Overall	Overall	20	3,001	0.80	1.75
	Gender	Female	20	1,456	0.79	1.77
		Male	20	1,545	0.81	1.74
	Ethnicity	AI/AN	20	42	0.64	1.90
		Asian	20	84	0.83	1.77
		Black or African American	20	195	0.78	1.86
		Hispanic	20	609	0.78	1.84
		NH/PI	20	10	0.83	1.87
		White	20	1,914	0.78	1.70
		Two or More Races	20	147	0.76	1.80
	FRL	Yes	20	1,473	0.79	1.82
		No	20	1,527	0.76	1.66
	LEP	Yes	20	480	0.78	1.85
		No	20	2,521	0.79	1.74
	SPED	Yes	20	475	0.79	1.90
		No	20	2,526	0.78	1.71
F	Overall	Overall	20	4,051	0.82	1.96
	Gender	Female	20	1,958	0.81	1.95
		Male	20	2,093	0.83	1.95
	Ethnicity	AI/AN	20	62	0.74	1.92
		Asian	20	120	0.84	1.92
		Black or African American	20	237	0.77	1.95
		Hispanic	20	776	0.79	1.94
		NH/PI	20	6	0.66	2.08
		White	20	2,673	0.81	1.93
		Two or More Races	20	175	0.83	1.92

Table 9.9: Cronbach’s Alpha by Demographics for Science Fixed Forms, cont.

	FRL	Yes	20	1,992	0.79	1.96
		No	20	2,057	0.80	1.94
	LEP	Yes	20	590	0.78	1.91
		No	20	3,459	0.82	1.94
	SPED	Yes	20	648	0.79	1.92
		No	20	3,403	0.81	1.96
Grade 8						
A	Overall	Overall	17	3,067	0.70	1.63
	Gender	Female	17	1,475	0.69	1.62
		Male	17	1,592	0.71	1.63
	Ethnicity	AI/AN	17	53	0.69	1.60
		Asian	17	74	0.74	1.61
		Black or African American	17	198	0.67	1.61
		Hispanic	17	601	0.66	1.63
		NH/PI	17	7	0.71	1.73
		White	17	2,005	0.67	1.62
		Two or More Races	17	129	0.62	1.64
	FRL	Yes	17	1,427	0.68	1.61
		No	17	1,640	0.65	1.64
	LEP	Yes	17	241	0.63	1.60
		No	17	2,826	0.68	1.64
	SPED	Yes	17	386	0.69	1.55
		No	17	2,681	0.66	1.64
B	Overall	Overall	17	4,242	0.74	2.12
	Gender	Female	17	1,993	0.74	2.11
		Male	17	2,249	0.75	2.09
	Ethnicity	AI/AN	17	49	0.82	1.75
		Asian	17	93	0.78	2.09
		Black or African American	17	196	0.75	1.78
		Hispanic	17	780	0.69	1.96
		NH/PI	17	11	0.80	2.05
		White	17	2,978	0.72	2.15
		Two or More Races	17	133	0.71	2.08
	FRL	Yes	17	1,993	0.73	1.98
		No	17	2,247	0.71	2.17
	LEP	Yes	17	329	0.66	1.76
		No	17	3,911	0.73	2.14
	SPED	Yes	17	602	0.73	1.83
		No	17	3,640	0.72	2.15
C	Overall	Overall	17	3,900	0.77	2.15
	Gender	Female	17	1,866	0.77	2.15
		Male	17	2,034	0.78	2.11
	Ethnicity	AI/AN	17	48	0.82	2.10
		Asian	17	99	0.80	2.16
		Black or African American	17	198	0.75	1.93
		Hispanic	17	717	0.76	2.02

Table 9.9: Cronbach’s Alpha by Demographics for Science Fixed Forms, cont.

		NH/PI	17	7	-2.40	2.09
		White	17	2,692	0.74	2.19
		Two or More Races	17	139	0.75	2.12
	FRL	Yes	17	1,809	0.76	2.07
		No	17	2,091	0.74	2.16
	LEP	Yes	17	291	0.68	1.97
		No	17	3,609	0.76	2.17
	SPED	Yes	17	506	0.75	1.94
		No	17	3,394	0.75	2.16
D	Overall	Overall	17	3,352	0.68	1.76
	Gender	Female	17	1,605	0.67	1.74
		Male	17	1,747	0.70	1.74
	Ethnicity	AI/AN	17	40	0.66	1.64
		Asian	17	75	0.70	1.72
		Black or African American	17	203	0.69	1.61
		Hispanic	17	637	0.67	1.71
		NH/PI	17	5	0.20	2.06
		White	17	2,260	0.65	1.78
		Two or More Races	17	132	0.72	1.72
	FRL	Yes	17	1,552	0.67	1.69
		No	17	1,800	0.64	1.77
	LEP	Yes	17	233	0.60	1.59
		No	17	3,119	0.67	1.76
	SPED	Yes	17	451	0.64	1.60
		No	17	2,901	0.66	1.76
E	Overall	Overall	17	3,069	0.69	1.59
	Gender	Female	17	1,479	0.65	1.60
		Male	17	1,590	0.72	1.58
	Ethnicity	AI/AN	17	47	0.48	1.38
		Asian	17	74	0.70	1.61
		Black or African American	17	203	0.54	1.44
		Hispanic	17	601	0.64	1.50
		NH/PI	17	7	0.43	1.78
		White	17	2,014	0.68	1.63
		Two or More Races	17	123	0.72	1.61
	FRL	Yes	17	1,391	0.62	1.50
		No	17	1,678	0.67	1.67
	LEP	Yes	17	234	0.59	1.37
		No	17	2,835	0.68	1.62
	SPED	Yes	17	411	0.55	1.38
		No	17	2,658	0.68	1.62
F	Overall	Overall	17	3,063	0.75	1.95
	Gender	Female	17	1,475	0.74	1.94
		Male	17	1,588	0.76	1.96
	Ethnicity	AI/AN	17	43	0.63	1.82
		Asian	17	85	0.82	1.94

Table 9.9: Cronbach’s Alpha by Demographics for Science Fixed Forms, cont.

		Black or African American	17	197	0.70	1.84
		Hispanic	17	598	0.71	1.92
		White	17	2,012	0.74	1.94
		Two or More Races	17	127	0.72	1.91
	FRL	Yes	17	1,397	0.72	1.92
		No	17	1,666	0.73	1.95
	LEP	Yes	17	217	0.72	1.84
		No	17	2,846	0.74	1.97
	SPED	Yes	17	414	0.71	1.80
		No	17	2,649	0.73	1.97

*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.

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” (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 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).

10.1 Intended Purposes and Uses of Test Scores

The purposes of the NSCAS assessment are as follows:

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

As the *Standards* note, ”validation is the joint responsibility of the test developer and the test user. . . the test user is ultimately responsible for evaluating the evidence in the particular setting in which the test is to be used” (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 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 (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p.21-22).”

The *Standards* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p.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” (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p.15). Evidence based on internal structure refer to the psychometric analyses of ”the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based” (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p.16). Evidence based on relations to other variables refers to traditional forms of criterion-related validity evidence such as predictive and concurrent validity, and evidence based on validity and consequences of testing refers to the evaluation of the intended and unintended consequences associated with a testing program.

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

10.3 Evidentiary Validity Framework

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

Table 10.1: Sources of Validity Evidence for Each NSCAS Test Purpose

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

Table 10.2: Sources of Validity Evidence based on Test Content

Test Purpose	Summary of Evidence	Tech Report Sections
1. Measure and report Nebraska students' depth of achievement regarding Nebraska's standards.	• Bias is minimized through Universal Design and accessibility resources. • TOS, 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 in ELA and Mathematics to be on track for achieving college readiness.	• Nebraska's College and Career Ready Standards are based on skills leading to college and career readiness across grades. • TOS, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity.	2

Table 10.2: Sources of Validity Evidence based on Test Content, cont.

3. Measure students' annual progress toward college and career readiness in ELA and Mathematics.	• Nebraska's College and Career Ready Standards are based on skills leading to college and career readiness across grades. • TOS, passage specifications and item specifications are aligned to grade-level content, process skills, and associated cognitive complexity.	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.	• TOS, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity. • TOS 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.	• Bias is minimized through Universal Design and accessibility resources. • DIF analysis completed for all items across all required subgroups. • 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 Nebraska's standards.	• Bias is minimized through Universal Design and accessibility resources. • TOS, 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 in ELA and Mathematics to be on track for achieving college readiness.	• TOS, 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 in ELA and Mathematics.	• TOS, 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.	• TOS, 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.	• Bias is minimized through Universal Design and accessibility resources. • DIF analysis completed for all items across all required subgroups. • 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 Nebraska’s standards.	• 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 in ELA and Mathematics to be on track for achieving college readiness.	• Scale is vertically articulated. • Achievement levels were vertically articulated.	6,7
3. Measure students’ annual progress toward college and career readiness in ELA and Mathematics.	• 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.	• 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.	• 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 Nebraska’s standards.	• Correlations with MAP Growth are high.	8
2. Report if student achievement is sufficient academic proficiency in ELA and Mathematics to be on track for achieving college readiness.		
3. Measure students’ annual progress toward college and career readiness in ELA and Mathematics.		

Table 10.5: Sources of Validity Evidence based on Other Variables, cont.

4. Inform teachers how student thinking differs along different areas of the scale as represented by the ALDs as information to support instructional planning.		
5. Assess students' construct relevant achievement in ELA, Mathematics, and Science for all students and subgroups of students.		

10.4 Interpretive Argument Claims

The test scores 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 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 and ELA cut score review process, including the methodology, identification of workshop participants, and implementation process, and ALD development and validation
Assessments were accessible to all students and fair across student subgroups.	3. Test Administration and Security 6. Psychometric Analyses	Accommodation policy and implementation, sensitivity review, availability of translations, and DIF analyses

10.5 NSCAS Validity Argument

The test development and technical quality of the Spring 2021 NSCAS Phase I Pilot 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 TOS during the constraint-based engine simulation studies and post-hoc analyses.

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

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

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

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

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

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Ed.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Council, N. R., et al. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

CRESST. (2015). *Simulation-based evaluation of the Smarter Balanced summative assessments. national center for research on evaluation, standards, and student testing.* (Tech. Rep.).

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. *Belmont, CA: Wadsworth Group/Thompson Learning*.

Dorans, N. J., & Schmitt, A. P. (1991). Constructed response and differential item functioning: A pragmatic approach. *ETS Research Report Series*, *1991*(2), i–49.

Drane, W., Torton, S., & Scott, M. (2021). *Data forensics report* (Tech. Rep.).

EdMetric. (2018a). *Nebraska student-centered assessment system – English language arts cut score review technical report* (Tech. Rep.).

EdMetric. (2018b). *Nebraska student-centered assessment system – mathematics standard setting technical report* (Tech. Rep.).

EdMetric. (2019). *Alignment study for nebraska student-centered assessment system, mathematics grades 3—8.* (Tech. Rep.).

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: Foundationa, methods, and innovations* (2nd ed., pp. 103–130). 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*(3), 315–332.

Fu, J., & Monfils, L. (2016). LRDIF_ES: A SAS macro for logistic regression tests for differential item functioning of dichotomous and polytomous items. *ETS Research Memorandum Series, ETS RM-16-17*.

Gómez-Benito, J., Hidalgo, M. D., & Padilla, J.-L. (2009). Efficacy of effect size measures in logistic regression: an application for detecting DIF. *Methodology–European Journal of Research Methods for the Behavioral and Social Sciences*, *5*(1), 18–25.

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media.

Holland, P., & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel procedure. In W. H. . B. HI (Ed.), *Test Validity* (pp. 129–145).

Huff, K., Warner, Z., & Schweid, J. (2016). Large-scale standards-based assessments of educational achievement. In A. A. Rupp & J. Leighton (Eds.), *The handbook of cognition assessment: Frameworks, methodologies, and applications* (pp. 399–426). John Wiley & Sons.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kim, S., & Kolen, M. J. (2004). *STUIRT: A computer program for scale transformation under unidimensional item response theory models.* University of Iowa, IA.

Linacre, J. (2021). *Winsteps® Rasch measurement computer program (version 4.8.0.0) [computer software].* Portland, OR.

Linn, R. L. (2006). Following the standards: Is it time for another revision? *Educational Measurement: Issues and Practice*, *25*(3), 54–56.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23.

Nebraska Department of Education. (2018). *Nebraska Student-Centered Assessment System (NSCAS) summative & alternate accessibility manual.*

Nebraska Department of Education. (2019). *Nebraska Student-Centered Assessment System (NSCAS) summative & alternate accessibility manual.*

NWEA. (2020a). *Constraint-based engine scientific approach and methodology (confidential).*

NWEA. (2020b, October). *Constraint-based engine simulation report for the spring 2021 NSCAS science field test* (Tech. Rep.). Portland, OR.

NWEA. (2020c, September). *Linking study report: Predicting performance on NSCAS general summative assessments based on NWEA MAP Growth scores* (Tech. Rep.). Portland, OR.

NWEA. (2021a, May). *Constraint-based engine evaluation report for the spring 2021 NSCAS science field test* (Tech. Rep.). Portland, OR.

NWEA. (2021b, January). *Constraint-based engine simulation report for the spring 2021 NSCAS ELA and Mathematics assessments.* Portland, OR.

NWEA. (2021c, May). *Constraint-based engine simulation report for the spring 2021 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.

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.

Rasch, G. (1960,1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Pædagogiske Institut.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement*, *18*(3), 229—244.

SBAC. (2016). *Smarter Balanced Assessment Consortium: 2014–15* (Tech. Rep.). CA.

Schneider, M. C., Huff, K. L., Egan, K. L., Gaines, M. L., & Ferrara, S. (2013). Relationships among item cognitive complexity, contextual demands, and item difficulty: Implications for achievement-level descriptors. *Educational Assessment*, *18*(2), 99–121.

Schneider, M. C., & Johnson, R. L. (2018). *Creating and implementing student learning objectives to support student learning and teacher evaluation*. Under contract. Taylor and Francis.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Mmeasurement*, *27*(4), 361–370.

USDE. (2018). *A state's guide to the U.S. Department of Education's assessment peer review process.* (Tech. Rep.). Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education.

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 in mathematics and science education. (Council of Chief State School Officers and National Institute for Science EducationResearch Monograph No. 6).

Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states: A study of the State Collaborative on Assessment & Student Standards (SCASS), Technical Issues in Large-Scale Assessment (TILSA).* Council of Chief State School Officers.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, *20*(1), 7–25.

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

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA press.

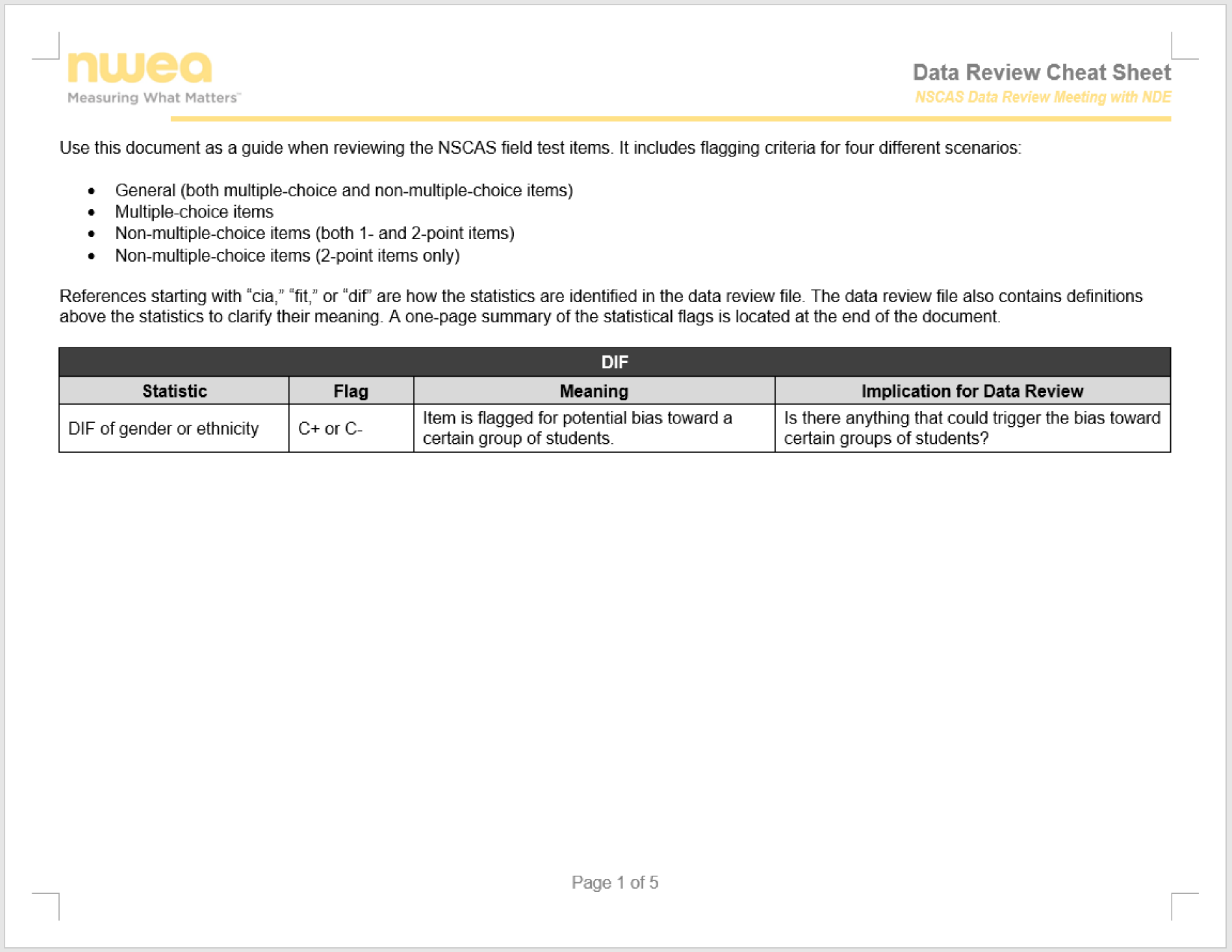
Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression as a unitary framework for binary and likert-type (ordinal) item scores.* Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

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

Zwick, R., Thayer, D. T., & Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement*, *18*(2), 121–140.

Appendix A. Data Review Cheat Sheet

Figure A.1: Data Review Cheat Sheet



Data Review Cheat Sheet

NSCAS Data Review Meeting with NDE

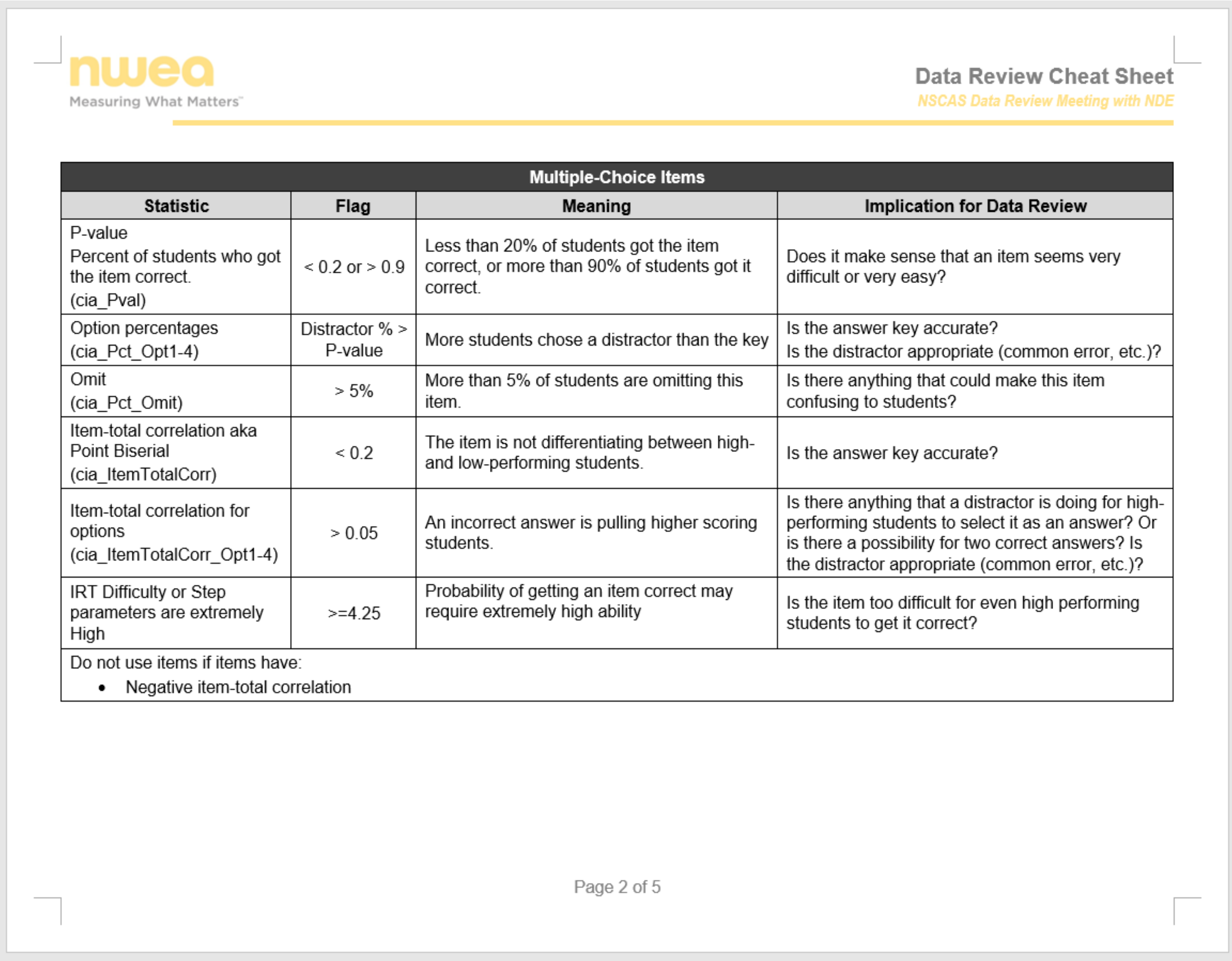
Use this document as a guide when reviewing the NSCAS field test items. It includes flagging criteria for four different scenarios:

- 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



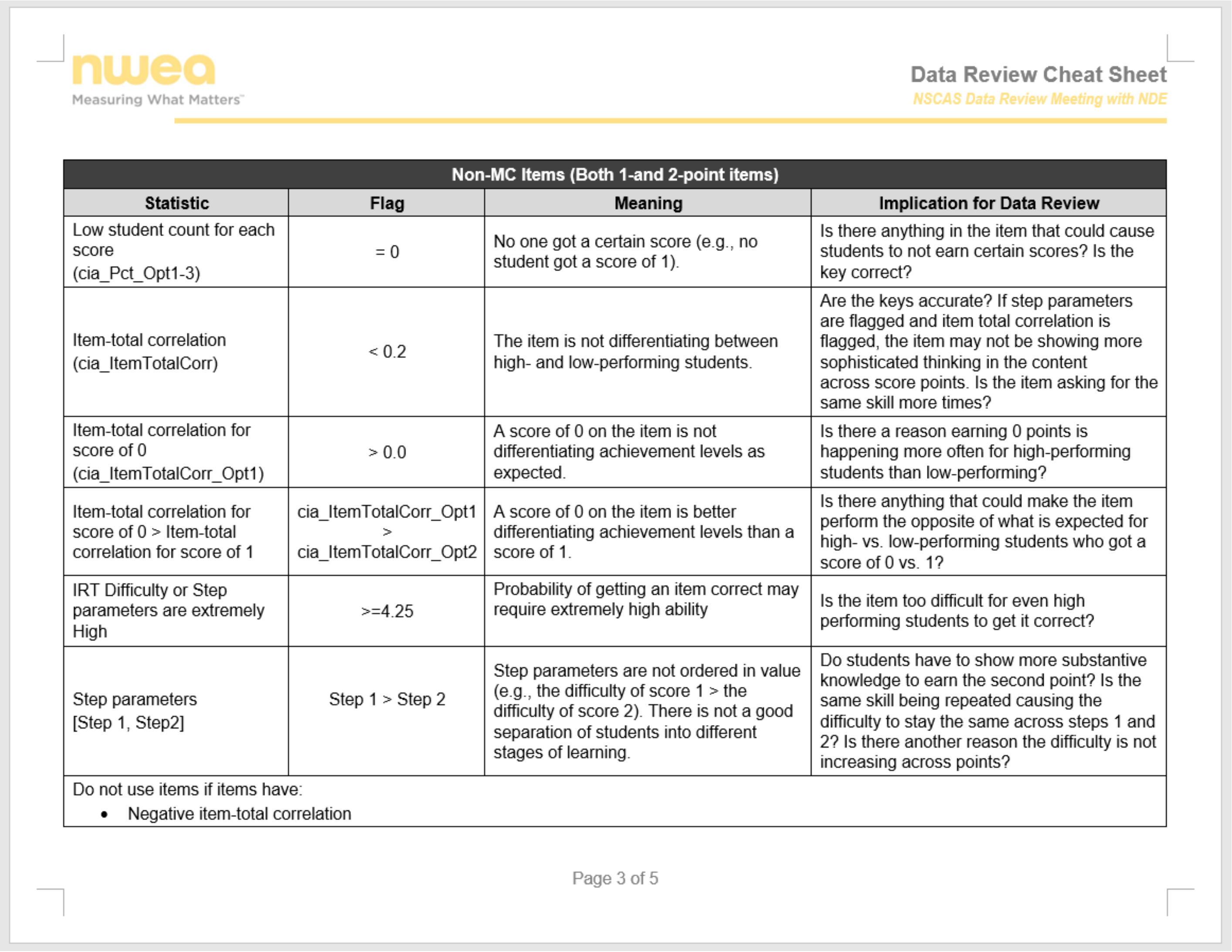
nwea
Measuring What Matters™

Data Review Cheat Sheet

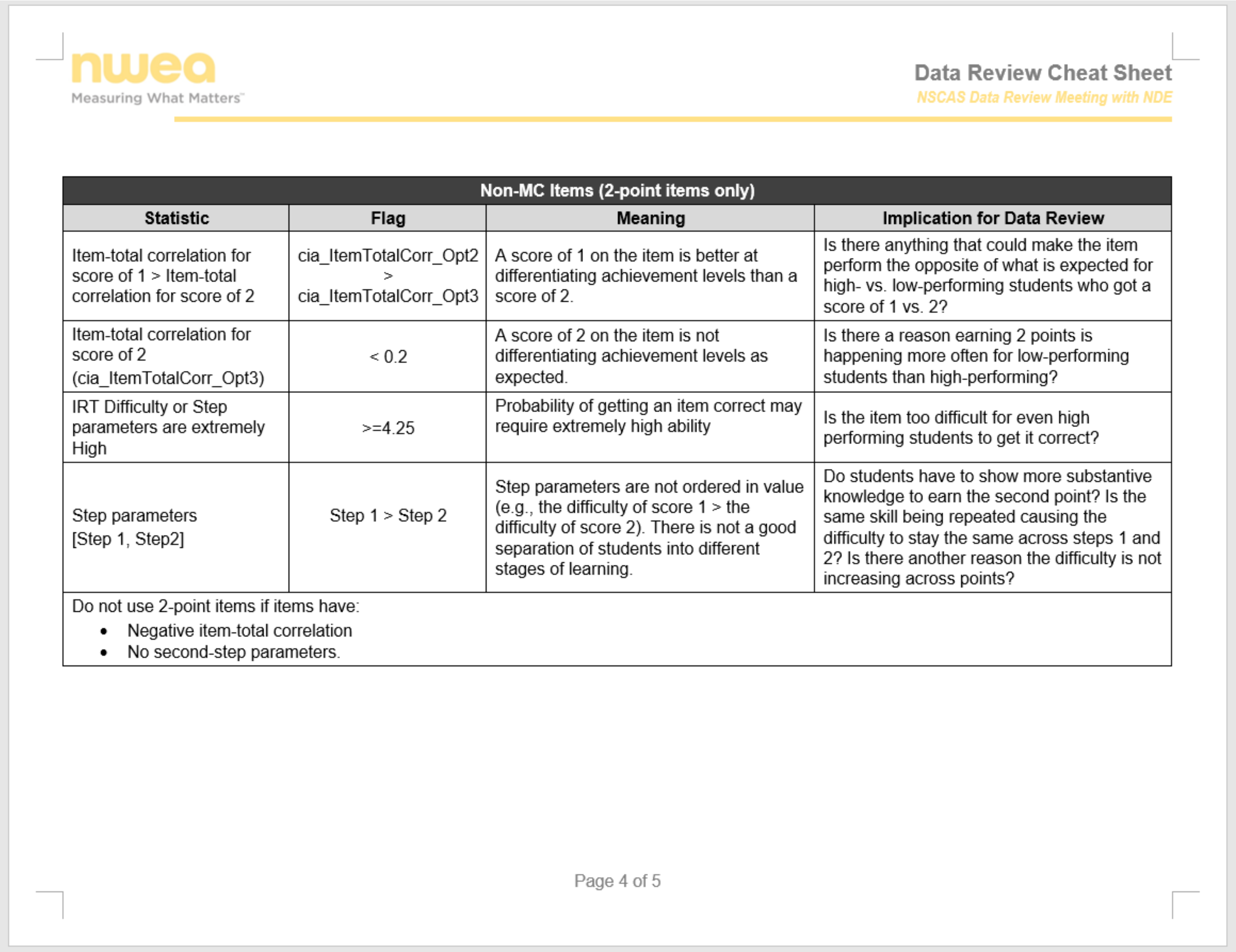
NSCAS Data Review Meeting with NDE

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: • Negative item-total correlation			

Page 2 of 5



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: • Negative item-total correlation			



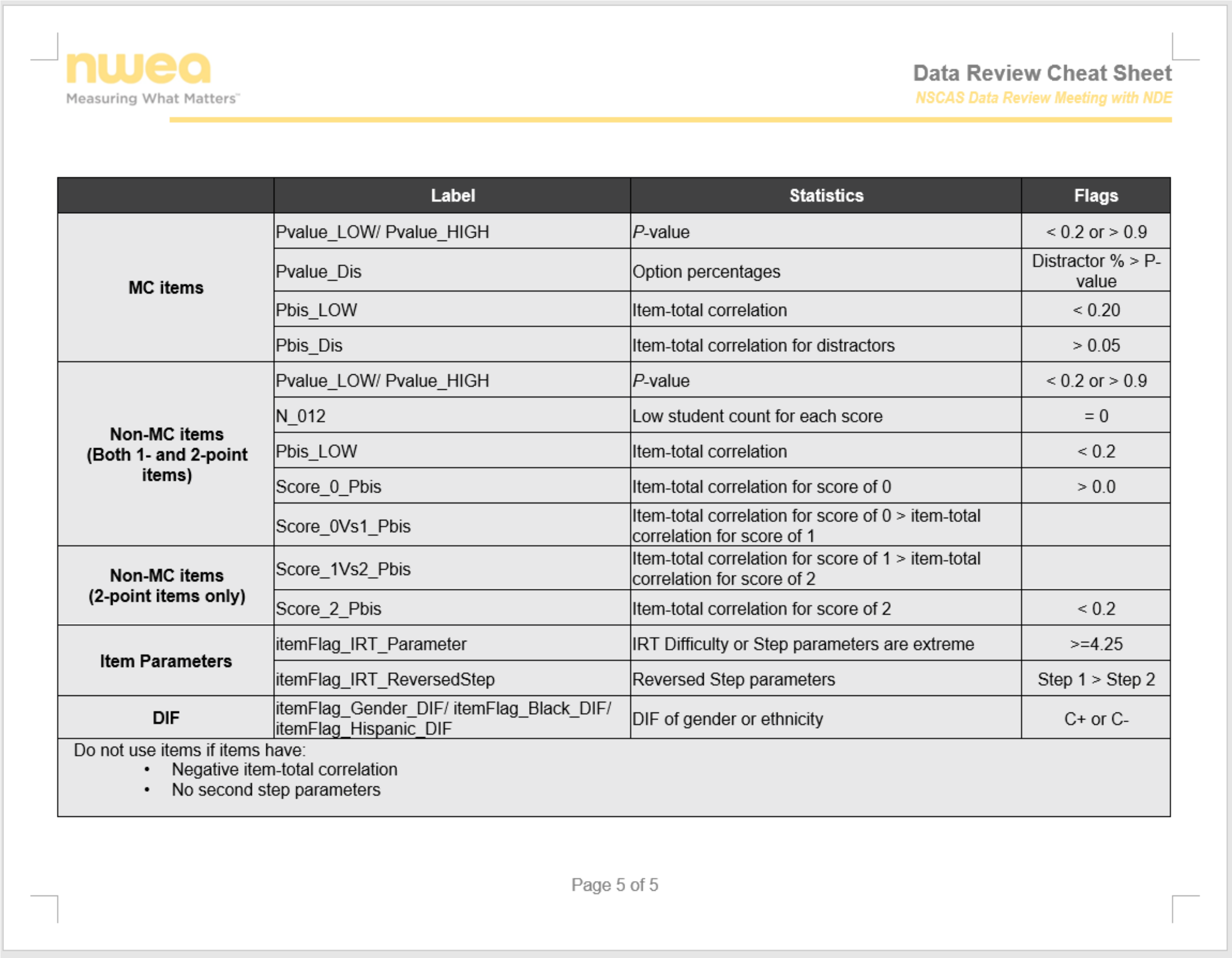
nwea
Measuring What Matters™

Data Review Cheat Sheet
NSCAS Data Review Meeting with NDE

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	cia_ItemTotalCorr_Opt2 > 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: • Negative item-total correlation • No second-step parameters.			

Page 4 of 5



	Label	Statistics	Flags
MC items	Pvalue_LOW/ Pvalue_HIGH	*P*-value	< 0.2 or > 0.9
	Pvalue_Dis	Option percentages	Distractor % > P-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	*P*-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:

- Negative item-total correlation
- No second step parameters

Appendix B. Summary P-Values by Item Types

Table B.1: Summary P-Values by Item Type: Operational Items

							#Items by P-Value Range									
Grade	Item Type	#Items	Mean	SD	Min	Max	≤0.1	≤0.2	≤0.3	≤0.4	≤0.5	≤0.6	≤0.7	≤0.8	≤0.9	>0.9
ELA																
3	Choice	515	0.493	0.113	0.124	0.927	0	3	16	84	158	178	56	16	3	1
	Choice Multiple	23	0.496	0.112	0.261	0.735	0	0	1	2	10	4	5	1	0	0
	Composite	23	0.401	0.137	0.121	0.629	0	3	2	5	9	2	2	0	0	0
	Gap Match	28	0.478	0.139	0.090	0.736	1	1	1	1	11	10	1	2	0	0
	Hot Text	1	0.064	.	0.064	0.064	1	0	0	0	0	0	0	0	0	0
4	Choice	504	0.539	0.134	0.076	1.000	1	0	10	58	140	148	86	42	15	4
	Choice Multiple	32	0.579	0.076	0.379	0.776	0	0	0	1	4	19	7	1	0	0
	Composite	18	0.497	0.097	0.373	0.657	0	0	0	2	8	4	4	0	0	0
	Gap Match	23	0.500	0.088	0.303	0.627	0	0	0	5	4	11	3	0	0	0
	Hot Text	2	0.563	0.023	0.547	0.580	0	0	0	0	0	2	0	0	0	0
5	Choice	450	0.530	0.122	0.150	0.970	0	3	9	42	130	158	69	31	5	3
	Choice Multiple	18	0.524	0.192	0.000	0.800	1	0	1	1	2	7	4	2	0	0
	Composite	16	0.448	0.102	0.162	0.650	0	1	0	2	9	3	1	0	0	0
	Gap Match	24	0.467	0.188	0.103	0.799	0	3	1	4	6	2	7	1	0	0
6	Choice	453	0.527	0.119	0.223	0.885	0	0	11	55	130	128	96	26	7	0
	Choice Multiple	31	0.450	0.094	0.274	0.643	0	0	3	6	12	8	2	0	0	0
	Composite	16	0.490	0.119	0.137	0.683	0	1	0	1	6	6	2	0	0	0
	Gap Match	16	0.462	0.143	0.261	0.622	0	0	4	2	3	1	6	0	0	0
	Hot Text	2	0.358	0.305	0.142	0.574	0	1	0	0	0	1	0	0	0	0
7	Choice	427	0.524	0.129	0.000	0.924	1	0	11	54	122	132	72	25	8	2
	Choice Multiple	26	0.438	0.078	0.269	0.559	0	0	1	7	13	5	0	0	0	0
	Composite	10	0.534	0.118	0.301	0.668	0	0	0	1	3	2	4	0	0	0
	Gap Match	15	0.519	0.103	0.319	0.640	0	0	0	3	2	6	4	0	0	0
8	Choice	491	0.562	0.126	0.078	0.987	1	0	4	41	103	164	116	45	11	6
	Choice Multiple	34	0.436	0.120	0.114	0.655	0	2	1	10	12	6	3	0	0	0
	Composite	15	0.405	0.181	0.000	0.613	1	1	2	4	2	2	3	0	0	0
	Gap Match	13	0.535	0.171	0.146	0.839	0	1	0	2	0	5	3	1	1	0
Mathematics																
3	Choice	411	0.532	0.078	0.262	0.836	0	0	2	24	100	207	73	4	1	0
	Choice Multiple	21	0.569	0.109	0.401	0.843	0	0	0	0	5	9	4	2	1	0
	Composite	30	0.503	0.089	0.249	0.689	0	0	1	0	14	11	4	0	0	0
	Gap Match	23	0.538	0.129	0.030	0.666	1	0	0	1	2	13	6	0	0	0
	Graphic Gap Match	23	0.525	0.136	0.070	0.713	1	0	1	1	2	11	6	1	0	0
	Hot Text	5	0.445	0.192	0.148	0.644	0	1	0	1	0	2	1	0	0	0
	Text Entry	27	0.545	0.079	0.402	0.683	0	0	0	0	8	13	6	0	0	0
4	Choice	268	0.482	0.075	0.271	0.775	0	0	2	29	129	96	8	4	0	0
	Choice Multiple	25	0.474	0.099	0.222	0.669	0	0	2	3	8	11	1	0	0	0
	Composite	36	0.400	0.104	0.000	0.533	1	0	4	11	17	3	0	0	0	0
	Gap Match	17	0.484	0.072	0.332	0.585	0	0	0	2	8	7	0	0	0	0
	Graphic Gap Match	23	0.513	0.060	0.395	0.610	0	0	0	1	10	11	1	0	0	0

Table B.1: Summary P-Values by Item Type: Operational Item, cont.

	Hot Text	14	0.429	0.103	0.253	0.607	0	0	2	3	5	3	1	0	0	0
	Text Entry	35	0.497	0.088	0.346	0.785	0	0	0	6	10	17	1	1	0	0
5	Choice	302	0.534	0.092	0.266	1.000	0	0	3	14	87	143	46	6	1	2
	Choice Multiple	25	0.521	0.076	0.394	0.684	0	0	0	1	10	10	4	0	0	0
	Composite	39	0.473	0.127	0.250	0.791	0	0	3	9	9	12	4	2	0	0
	Gap Match	18	0.535	0.051	0.460	0.611	0	0	0	0	5	10	3	0	0	0
	Graphic Gap Match	14	0.605	0.069	0.479	0.712	0	0	0	0	2	5	6	1	0	0
	Hot Text	8	0.473	0.065	0.372	0.564	0	0	0	1	4	3	0	0	0	0
	Text Entry	26	0.549	0.113	0.332	0.827	0	0	0	3	4	12	6	0	1	0
6	Choice	374	0.500	0.076	0.291	0.757	0	0	3	34	149	155	32	1	0	0
	Choice Multiple	40	0.410	0.119	0.177	0.602	0	2	7	9	9	12	1	0	0	0
	Composite	38	0.420	0.122	0.164	0.661	0	1	7	11	8	9	2	0	0	0
	Gap Match	27	0.508	0.088	0.237	0.680	0	0	1	2	7	14	3	0	0	0
	Graphic Gap Match	12	0.533	0.049	0.448	0.622	0	0	0	0	3	8	1	0	0	0
	Hot Text	15	0.467	0.146	0.257	0.844	0	0	2	3	5	3	1	0	1	0
	Text Entry	31	0.510	0.091	0.302	0.739	0	0	0	3	11	13	3	1	0	0
7	Choice	329	0.449	0.086	0.231	0.807	0	0	14	76	155	68	14	1	1	0
	Choice Multiple	23	0.397	0.161	0.142	0.768	0	2	5	6	5	3	0	2	0	0
	Composite	27	0.370	0.117	0.186	0.706	0	1	7	9	7	2	0	1	0	0
	Gap Match	20	0.444	0.053	0.340	0.545	0	0	0	4	13	3	0	0	0	0
	Graphic Gap Match	7	0.426	0.076	0.310	0.536	0	0	0	2	4	1	0	0	0	0
	Hot Text	15	0.417	0.132	0.155	0.636	0	1	3	2	6	1	2	0	0	0
	Text Entry	36	0.473	0.065	0.306	0.630	0	0	0	4	22	9	1	0	0	0
8	Choice	287	0.471	0.080	0.285	0.738	0	0	3	49	143	74	16	2	0	0
	Choice Multiple	16	0.388	0.090	0.171	0.535	0	1	0	9	3	3	0	0	0	0
	Composite	30	0.362	0.125	0.000	0.665	1	1	5	12	9	1	1	0	0	0
	Gap Match	33	0.441	0.088	0.245	0.613	0	0	3	7	15	7	1	0	0	0
	Graphic Gap Match	9	0.435	0.086	0.281	0.558	0	0	1	3	3	2	0	0	0	0
	Hot Text	27	0.428	0.098	0.238	0.649	0	0	3	8	10	5	1	0	0	0
	Text Entry	33	0.506	0.088	0.352	0.703	0	0	0	4	11	12	5	1	0	0

Table B.2: Summary P-Values by Item Type: Field Test Items

							#Items by P-Value Range									
Grade	**Item Type**	**#Items**	**Mean**	**SD**	**Min**	**Max**	**≤0.1**	**≤0.2**	**≤0.3**	**≤0.4**	**≤0.5**	**≤0.6**	**≤0.7**	**≤0.8**	**≤0.9**	**>0.9**
ELA																
3	Choice	124	0.519	0.165	0.157	0.877	0	2	9	21	29	23	20	13	7	0
	Choice Multiple	17	0.482	0.080	0.324	0.594	0	0	0	3	6	8	0	0	0	0
	Composite	20	0.331	0.147	0.083	0.560	1	4	5	3	5	2	0	0	0	0
	Gap Match	16	0.479	0.140	0.237	0.758	0	0	1	4	5	3	2	1	0	0
	Hot Text	7	0.473	0.093	0.394	0.672	0	0	0	1	5	0	1	0	0	0
4	Choice	122	0.536	0.182	0.159	0.918	0	3	10	14	26	26	21	9	12	1
	Choice Multiple	23	0.535	0.097	0.354	0.716	0	0	0	1	7	11	2	2	0	0
	Composite	22	0.399	0.125	0.176	0.590	0	2	3	6	5	6	0	0	0	0
	Gap Match	16	0.579	0.191	0.038	0.832	1	0	0	0	3	4	3	4	1	0
	Hot Text	2	0.416	0.189	0.283	0.550	0	0	1	0	0	1	0	0	0	0

Table B.2: Summary P-Values by Item Type: Field Test Item, cont.

5	Choice	118	0.499	0.169	0.142	0.862	0	1	17	22	20	21	18	16	3	0
	Choice Multiple	23	0.621	0.098	0.501	0.791	0	0	0	0	0	11	5	7	0	0
	Composite	20	0.386	0.149	0.185	0.709	0	1	6	5	3	3	1	1	0	0
	Gap Match	17	0.556	0.175	0.342	0.935	0	0	0	4	3	5	1	2	1	1
	Hot Text	8	0.523	0.109	0.394	0.758	0	0	0	1	1	5	0	1	0	0
6	Choice	120	0.501	0.164	0.123	0.911	0	3	11	23	22	28	18	10	4	1
	Choice Multiple	21	0.421	0.119	0.181	0.659	0	1	4	2	8	5	1	0	0	0
	Composite	15	0.416	0.143	0.203	0.624	0	0	4	3	4	1	3	0	0	0
	Gap Match	15	0.528	0.243	0.140	0.858	0	3	1	0	3	1	3	2	2	0
	Hot Text	2	0.513	0.038	0.486	0.540	0	0	0	0	1	1	0	0	0	0
7	Choice	105	0.582	0.160	0.241	0.925	0	0	2	13	20	27	16	14	12	1
	Choice Multiple	19	0.463	0.147	0.122	0.670	0	1	3	1	3	8	3	0	0	0
	Composite	25	0.469	0.132	0.235	0.782	0	0	2	6	7	6	2	2	0	0
	Gap Match	15	0.427	0.203	0.099	0.865	1	0	4	4	1	3	0	1	1	0
	Hot Text	16	0.589	0.102	0.414	0.786	0	0	0	0	3	6	5	2	0	0
8	Choice	152	0.583	0.183	0.169	0.957	0	3	11	14	19	31	32	22	16	4
	Choice Multiple	26	0.503	0.169	0.168	0.792	0	2	1	2	10	2	5	4	0	0
	Composite	26	0.521	0.119	0.238	0.737	0	0	1	3	6	11	3	2	0	0
	Gap Match	11	0.523	0.204	0.184	0.762	0	1	0	3	1	1	2	3	0	0
	Hot Text	12	0.674	0.138	0.466	0.868	0	0	0	0	2	1	3	3	3	0
Mathematics																
3	Choice	138	0.555	0.189	0.193	0.961	0	2	11	24	21	25	16	23	13	3
	Choice Multiple	14	0.410	0.186	0.146	0.809	0	3	1	3	3	2	1	0	1	0
	Composite	18	0.504	0.164	0.245	0.787	0	0	3	2	4	3	5	1	0	0
	Gap Match	20	0.445	0.264	0.046	0.814	3	2	2	3	0	3	2	4	1	0
	Graphic Gap Match	15	0.341	0.174	0.012	0.761	1	1	5	2	5	0	0	1	0	0
	Hot Text	4	0.379	0.273	0.127	0.618	0	2	0	0	0	0	2	0	0	0
	Text Entry	22	0.495	0.180	0.144	0.868	0	1	2	4	4	5	3	1	2	0
4	Choice	68	0.550	0.166	0.159	0.855	0	1	5	6	15	14	12	10	5	0
	Choice Multiple	19	0.483	0.144	0.211	0.745	0	0	2	3	5	4	4	1	0	0
	Composite	15	0.548	0.156	0.285	0.804	0	0	1	2	3	4	2	2	1	0
	Gap Match	13	0.435	0.178	0.261	0.790	0	0	4	3	1	2	2	1	0	0
	Graphic Gap Match	9	0.570	0.124	0.294	0.703	0	0	1	0	1	3	3	1	0	0
	Hot Text	4	0.450	0.173	0.280	0.618	0	0	1	1	0	1	1	0	0	0
	Text Entry	22	0.536	0.173	0.290	0.858	0	0	2	3	5	5	3	2	2	0
5	Choice	86	0.608	0.166	0.207	0.972	0	0	3	4	13	23	16	18	4	5
	Choice Multiple	19	0.507	0.187	0.212	0.800	0	0	4	1	5	2	3	4	0	0
	Composite	24	0.514	0.156	0.153	0.793	0	1	1	2	8	7	2	3	0	0
	Gap Match	14	0.498	0.211	0.154	0.834	0	1	2	1	3	2	3	1	1	0
	Graphic Gap Match	4	0.393	0.107	0.307	0.549	0	0	0	3	0	1	0	0	0	0
	Hot Text	6	0.516	0.225	0.209	0.736	0	0	2	0	0	1	1	2	0	0
	Text Entry	29	0.513	0.174	0.137	0.870	0	2	1	4	6	7	4	4	1	0
6	Choice	184	0.566	0.176	0.111	0.914	0	3	10	22	30	46	27	26	17	3
	Choice Multiple	15	0.194	0.095	0.054	0.311	3	5	6	1	0	0	0	0	0	0
	Composite	16	0.410	0.198	0.201	0.852	0	0	6	3	2	1	3	0	1	0
	Gap Match	4	0.264	0.044	0.214	0.317	0	0	3	1	0	0	0	0	0	0

Table B.2: Summary P-Values by Item Type: Field Test Item, cont.

	Graphic Gap Match	2	0.246	0.237	0.078	0.413	1	0	0	0	1	0	0	0	0	0
	Hot Text	9	0.289	0.133	0.100	0.515	0	2	3	3	0	1	0	0	0	0
	Text Entry	1	0.322	.	0.322	0.322	0	0	0	1	0	0	0	0	0	0
7	Choice	149	0.531	0.186	0.160	0.926	0	3	13	25	25	27	23	19	13	1
	Choice Multiple	13	0.192	0.144	0.032	0.558	4	4	2	2	0	1	0	0	0	0
	Composite	23	0.364	0.142	0.047	0.687	1	1	4	10	2	4	1	0	0	0
	Gap Match	5	0.321	0.164	0.095	0.511	1	0	1	1	1	1	0	0	0	0
	Graphic Gap Match	3	0.322	0.182	0.175	0.525	0	1	1	0	0	1	0	0	0	0
	Hot Text	8	0.361	0.107	0.149	0.502	0	1	0	5	1	1	0	0	0	0
	Text Entry	25	0.211	0.157	0.022	0.655	6	9	5	1	2	1	1	0	0	0
8	Choice	68	0.523	0.183	0.168	0.860	0	4	4	11	13	12	11	9	4	0
	Choice Multiple	23	0.194	0.090	0.030	0.365	3	8	7	5	0	0	0	0	0	0
	Composite	28	0.351	0.118	0.133	0.647	0	3	6	10	8	0	1	0	0	0
	Gap Match	10	0.322	0.231	0.030	0.785	1	3	1	2	1	1	0	1	0	0
	Graphic Gap Match	2	0.522	0.148	0.417	0.626	0	0	0	0	1	0	1	0	0	0
	Hot Text	10	0.433	0.158	0.215	0.652	0	0	3	2	0	4	1	0	0	0
	Text Entry	16	0.275	0.153	0.070	0.572	1	6	3	3	1	2	0	0	0	0

table note test test test

Appendix C. Summary Item-Total Correlations by Item Types

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

							#Items by Item-Total Correlations Range						
Grade	**Item Type**	**#Items**	**Mean**	**SD**	**Min**	**Max**	**≤0.1**	**≤0.2**	**≤0.3**	**≤0.4**	**≤0.5**	**≤0.6**	**>0.6**
ELA													
3	Choice	515	0.384	0.083	0.010	0.906	2	5	56	241	178	28	5
	Choice Multiple	23	0.490	0.109	0.219	0.618	0	0	1	3	8	9	2
	Composite	23	0.472	0.117	0.269	0.657	0	0	1	7	4	8	3
	Gap Match	28	0.393	0.091	0.163	0.548	0	2	1	12	11	2	0
	Hot Text	1	0.278	.	0.278	0.278	0	0	1	0	0	0	0
4	Choice	504	0.377	0.076	0.000	0.781	2	2	60	265	159	12	4
	Choice Multiple	32	0.441	0.126	0.000	0.621	1	0	0	10	12	4	5
	Composite	18	0.505	0.113	0.208	0.612	0	0	1	2	4	8	3
	Gap Match	23	0.382	0.079	0.220	0.565	0	0	5	10	7	1	0
	Hot Text	2	0.514	0.056	0.474	0.553	0	0	0	0	1	1	0
5	Choice	450	0.377	0.072	0.057	0.648	1	5	49	235	144	15	1
	Choice Multiple	18	0.385	0.209	-0.188	0.605	2	0	2	2	7	4	1
	Composite	16	0.479	0.084	0.212	0.557	0	0	1	1	5	9	0
	Gap Match	24	0.372	0.094	0.194	0.544	0	1	4	8	10	1	0
6	Choice	453	0.381	0.077	0.087	0.609	1	5	58	212	145	31	1
	Choice Multiple	31	0.453	0.091	0.295	0.730	0	0	1	7	15	6	2
	Composite	16	0.514	0.100	0.232	0.663	0	0	1	0	4	8	3
	Gap Match	16	0.400	0.095	0.262	0.618	0	0	3	6	5	1	1
	Hot Text	2	0.365	0.199	0.224	0.506	0	0	1	0	0	1	0
7	Choice	427	0.376	0.075	0.000	0.658	1	1	62	214	129	19	1
	Choice Multiple	26	0.427	0.110	0.190	0.767	0	1	1	9	9	5	1
	Composite	10	0.532	0.087	0.348	0.633	0	0	0	1	3	3	3
	Gap Match	15	0.419	0.054	0.329	0.504	0	0	0	5	8	2	0
8	Choice	491	0.391	0.084	0.109	0.815	0	7	51	217	178	34	4
	Choice Multiple	34	0.416	0.123	0.051	0.663	1	0	3	13	9	6	2
	Composite	15	0.465	0.176	0.000	0.726	1	0	1	1	6	3	3
	Gap Match	13	0.424	0.110	0.231	0.584	0	0	2	4	3	4	0
Mathematics													
3	Choice	411	0.384	0.060	0.215	0.596	0	0	39	214	146	12	0
	Choice Multiple	21	0.429	0.084	0.294	0.594	0	0	2	7	8	4	0
	Composite	30	0.583	0.067	0.426	0.754	0	0	0	0	3	15	12
	Gap Match	23	0.396	0.071	0.200	0.509	0	0	2	6	14	1	0
	Graphic Gap Match	23	0.378	0.070	0.252	0.497	0	0	4	10	9	0	0
	Hot Text	5	0.493	0.139	0.337	0.634	0	0	0	2	0	2	1
	Text Entry	27	0.411	0.055	0.301	0.555	0	0	0	11	14	2	0
4	Choice	268	0.372	0.063	0.172	0.502	0	3	31	133	100	1	0
	Choice Multiple	25	0.396	0.076	0.302	0.647	0	0	0	15	8	1	1
	Composite	36	0.534	0.120	0.000	0.691	1	0	0	2	6	15	12
	Gap Match	17	0.397	0.060	0.294	0.477	0	0	2	5	10	0	0
	Graphic Gap Match	23	0.419	0.082	0.258	0.592	0	0	1	10	8	4	0

Table C.1: Summary Item-Total Correlations by Item Type: Operational Item, cont.

	Hot Text	14	0.436	0.096	0.276	0.612	0	0	1	7	3	2	1
	Text Entry	35	0.402	0.061	0.270	0.517	0	0	2	15	17	1	0
5	Choice	302	0.400	0.083	0.000	0.662	1	2	27	123	112	35	2
	Choice Multiple	25	0.434	0.101	0.296	0.723	0	0	3	6	10	5	1
	Composite	39	0.559	0.119	0.329	1.000	0	0	0	2	8	18	11
	Gap Match	18	0.415	0.083	0.279	0.533	0	0	3	5	7	3	0
	Graphic Gap Match	14	0.425	0.087	0.254	0.595	0	0	1	6	4	3	0
	Hot Text	8	0.527	0.097	0.436	0.738	0	0	0	0	3	4	1
	Text Entry	26	0.423	0.081	0.215	0.542	0	0	1	10	10	5	0
6	Choice	374	0.376	0.066	0.146	0.561	0	2	37	205	118	12	0
	Choice Multiple	40	0.430	0.099	0.252	0.609	0	0	5	15	7	12	1
	Composite	38	0.532	0.106	0.215	0.688	0	0	3	0	6	19	10
	Gap Match	27	0.384	0.061	0.215	0.476	0	0	2	13	12	0	0
	Graphic Gap Match	12	0.417	0.059	0.343	0.544	0	0	0	4	7	1	0
	Hot Text	15	0.426	0.123	0.255	0.583	0	0	3	5	1	6	0
	Text Entry	31	0.393	0.055	0.301	0.570	0	0	0	19	11	1	0
7	Choice	329	0.368	0.066	0.104	0.555	0	2	51	170	102	4	0
	Choice Multiple	23	0.444	0.097	0.221	0.619	0	0	1	7	9	5	1
	Composite	27	0.532	0.095	0.187	0.622	0	1	0	1	4	17	4
	Gap Match	20	0.405	0.052	0.277	0.512	0	0	1	7	11	1	0
	Graphic Gap Match	7	0.366	0.069	0.239	0.440	0	0	1	3	3	0	0
	Hot Text	15	0.401	0.089	0.201	0.532	0	0	2	4	7	2	0
	Text Entry	36	0.388	0.047	0.230	0.465	0	0	1	20	15	0	0
8	Choice	287	0.370	0.061	0.174	0.525	0	1	37	159	84	6	0
	Choice Multiple	16	0.441	0.102	0.309	0.647	0	0	0	6	6	2	2
	Composite	30	0.500	0.132	0.000	0.639	1	0	2	1	7	16	3
	Gap Match	33	0.385	0.064	0.205	0.501	0	0	2	15	15	1	0
	Graphic Gap Match	9	0.407	0.036	0.340	0.453	0	0	0	5	4	0	0
	Hot Text	27	0.413	0.097	0.254	0.613	0	0	4	8	11	3	1
	Text Entry	33	0.409	0.033	0.344	0.493	0	0	0	16	17	0	0

Table C.2: Summary Item-Total Correlations by Item Type: Field Test Items

							#Items by Item-Total Correlations Range						
Grade	**Item Type**	**#Items**	**Mean**	**SD**	**Min**	**Max**	**≤0.1**	**≤0.2**	**≤0.3**	**≤0.4**	**≤0.5**	**≤0.6**	**>0.6**
ELA													
3	Choice	124	0.303	0.127	-0.143	0.532	9	17	28	40	27	3	0
	Choice Multiple	17	0.341	0.136	0.022	0.516	1	2	3	5	4	2	0
	Composite	20	0.297	0.151	0.030	0.541	3	3	3	7	2	2	0
	Gap Match	16	0.482	0.085	0.316	0.614	0	0	0	3	4	7	2
	Hot Text	7	0.393	0.127	0.166	0.529	0	1	0	2	2	2	0
4	Choice	122	0.288	0.137	-0.069	0.561	15	13	32	40	21	1	0
	Choice Multiple	23	0.373	0.116	0.171	0.562	0	2	5	5	7	4	0
	Composite	22	0.338	0.130	0.103	0.554	0	3	6	5	6	2	0
	Gap Match	16	0.385	0.195	-0.149	0.537	2	0	1	2	6	5	0
	Hot Text	2	0.422	0.064	0.377	0.468	0	0	0	1	1	0	0

Table C.2: Summary Item-Total Correlations by Item Type: Field Test Item, cont.

5	Choice	118	0.266	0.130	-0.136	0.497	12	27	24	36	19	0	0
	Choice Multiple	23	0.425	0.071	0.278	0.569	0	0	2	5	14	2	0
	Composite	20	0.332	0.128	0.093	0.547	1	2	4	6	5	2	0
	Gap Match	17	0.405	0.095	0.230	0.531	0	0	3	5	6	3	0
	Hot Text	8	0.354	0.098	0.160	0.447	0	1	1	4	2	0	0
6	Choice	120	0.285	0.125	-0.159	0.498	9	21	31	32	27	0	0
	Choice Multiple	21	0.320	0.148	-0.021	0.501	2	3	1	9	5	1	0
	Composite	15	0.381	0.113	0.163	0.512	0	2	2	1	8	2	0
	Gap Match	15	0.365	0.179	-0.088	0.602	1	1	2	2	6	2	1
	Hot Text	2	0.454	0.008	0.448	0.460	0	0	0	0	2	0	0
7	Choice	105	0.286	0.105	0.047	0.521	6	16	30	42	10	1	0
	Choice Multiple	19	0.374	0.076	0.242	0.510	0	0	3	10	5	1	0
	Composite	25	0.349	0.101	0.156	0.529	0	3	5	10	4	3	0
	Gap Match	15	0.398	0.084	0.257	0.552	0	0	2	5	7	1	0
	Hot Text	16	0.347	0.133	-0.007	0.530	1	1	4	4	5	1	0
8	Choice	152	0.300	0.121	-0.109	0.508	11	18	38	50	34	1	0
	Choice Multiple	26	0.348	0.140	-0.046	0.522	2	1	3	11	8	1	0
	Composite	26	0.365	0.103	0.057	0.506	1	1	3	11	9	1	0
	Gap Match	11	0.356	0.111	0.185	0.545	0	1	2	4	3	1	0
	Hot Text	12	0.358	0.095	0.161	0.490	0	1	1	4	6	0	0
Mathematics													
3	Choice	138	0.357	0.122	-0.073	0.555	4	12	20	41	50	11	0
	Choice Multiple	14	0.412	0.118	0.106	0.544	0	1	1	3	6	3	0
	Composite	18	0.526	0.075	0.375	0.631	0	0	0	2	4	8	4
	Gap Match	20	0.400	0.104	0.209	0.527	0	0	5	2	8	5	0
	Graphic Gap Match	15	0.413	0.106	0.150	0.528	0	1	1	2	8	3	0
	Hot Text	4	0.434	0.133	0.334	0.624	0	0	0	2	1	0	1
	Text Entry	22	0.438	0.115	0.171	0.591	0	1	2	3	10	6	0
4	Choice	68	0.364	0.132	-0.188	0.595	2	4	10	22	24	6	0
	Choice Multiple	19	0.435	0.089	0.269	0.590	0	0	1	5	8	5	0
	Composite	15	0.485	0.107	0.226	0.629	0	0	1	2	4	7	1
	Gap Match	13	0.452	0.104	0.240	0.618	0	0	1	2	4	5	1
	Graphic Gap Match	9	0.468	0.035	0.402	0.522	0	0	0	0	7	2	0
	Hot Text	4	0.443	0.097	0.313	0.540	0	0	0	1	2	1	0
	Text Entry	22	0.482	0.058	0.354	0.563	0	0	0	2	11	9	0
5	Choice	86	0.352	0.113	0.094	0.610	2	7	15	32	22	7	1
	Choice Multiple	19	0.398	0.058	0.308	0.489	0	0	0	10	9	0	0
	Composite	24	0.501	0.087	0.267	0.618	0	0	1	2	7	12	2
	Gap Match	14	0.447	0.097	0.217	0.575	0	0	1	5	3	5	0
	Graphic Gap Match	4	0.524	0.087	0.411	0.604	0	0	0	0	1	2	1
	Hot Text	6	0.386	0.140	0.212	0.556	0	0	2	1	1	2	0
	Text Entry	29	0.444	0.082	0.259	0.544	0	0	2	7	11	9	0
6	Choice	184	0.363	0.107	-0.036	0.590	3	7	38	61	62	13	0
	Choice Multiple	15	0.335	0.151	0.033	0.520	2	1	2	4	5	1	0
	Composite	16	0.442	0.079	0.277	0.561	0	0	1	3	9	3	0
	Gap Match	4	0.367	0.078	0.274	0.445	0	0	1	1	2	0	0

Table C.2: Summary Item-Total Correlations by Item Type: Field Test Item, cont.

	Graphic Gap Match	2	0.338	0.202	0.195	0.480	0	1	0	0	1	0	0
	Hot Text	9	0.322	0.125	0.065	0.448	1	0	2	2	4	0	0
	Text Entry	1	0.481	.	0.481	0.481	0	0	0	0	1	0	0
7	Choice	149	0.347	0.115	-0.170	0.576	3	14	27	49	49	7	0
	Choice Multiple	13	0.381	0.129	0.144	0.588	0	2	1	3	5	2	0
	Composite	23	0.391	0.154	0.007	0.611	2	0	3	8	3	6	1
	Gap Match	5	0.352	0.066	0.284	0.457	0	0	1	3	1	0	0
	Graphic Gap Match	3	0.416	0.175	0.306	0.618	0	0	0	2	0	0	1
	Hot Text	8	0.438	0.118	0.228	0.617	0	0	1	2	3	1	1
	Text Entry	25	0.419	0.089	0.247	0.555	0	0	2	10	8	5	0
8	Choice	68	0.324	0.109	-0.015	0.530	4	3	18	23	18	2	0
	Choice Multiple	23	0.347	0.107	0.080	0.525	1	0	8	6	7	1	0
	Composite	28	0.454	0.115	0.161	0.616	0	1	3	2	10	10	2
	Gap Match	10	0.332	0.118	0.083	0.469	1	0	2	3	4	0	0
	Graphic Gap Match	2	0.426	0.028	0.407	0.446	0	0	0	0	2	0	0
	Hot Text	10	0.368	0.111	0.184	0.543	0	1	1	4	3	1	0
	Text Entry	16	0.456	0.078	0.202	0.521	0	0	1	2	8	5	0

Appendix D. Achievement Level Distributions and Scale Score Descriptive Statistics by Demographics

Table D.1: Achievement Level Distributions and Scale Score Descriptive Statistics by Demographics–ELA

				Descriptive Statistics		Percent of Students in Each Achievement Level**			
Grade	**Demographic Sub-Group***		**N**	**Mean**	**SD**	**Level 3**	**Level 2**	**Level 1**	**L2 + L1**
3		Overall	21,779	2467.04	87.35	49.9	35.8	14.3	50.1
	Gender	Female	10,623	2473.21	85.13	46.9	37.4	15.7	53.1
		Male	11,156	2461.16	89.01	52.6	34.4	13.0	47.4
	Ethnicity	AI/AN	287	2403.37	84.59	79.1	18.1	2.8	20.9
		Asian	697	2465.76	92.92	50.8	33.4	15.8	49.2
		Black	1,311	2415.52	89.75	73.0	22.2	4.8	27.0
		Hispanic	4,218	2427.92	85.73	69.3	25.2	5.5	30.7
		NH/PI	36	2435.72	86.85	66.7	25.0	8.3	33.3
		White	14,225	2485.32	80.71	41.0	41.0	18.0	59.0
		Two or More Races	1,005	2459.86	87.49	54.3	32.2	13.4	45.7
	FRL	Yes	10,808	2434.92	86.11	65.4	28.2	6.4	34.6
		No	10,971	2498.68	76.31	34.5	43.4	22.0	65.5
	LEP	Yes	3,539	2421.83	85.87	72.0	23.3	4.7	28.0
		No	18,240	2475.81	84.89	45.6	38.3	16.2	54.4
	SPED	Yes	3,595	2410.91	90.80	74.7	20.2	5.1	25.3
		No	18,184	2478.14	82.23	44.9	38.9	16.1	55.1
4		Overall	21,712	2501.11	84.01	46.3	36.5	17.2	53.7
	Gender	Female	10,572	2507.64	81.29	43.2	38.3	18.5	56.8
		Male	11,140	2494.92	86.07	49.2	34.9	15.9	50.8
	Ethnicity	AI/AN	255	2447.30	79.98	74.1	22.7	3.1	25.9
		Asian	658	2504.89	91.72	44.2	32.8	22.9	55.8
		Black	1,251	2448.68	89.70	69.8	24.3	5.9	30.2
		Hispanic	4,287	2464.80	82.55	64.8	28.2	7.0	35.2
		NH/PI	35	2466.91	85.74	68.6	20.0	11.4	31.4
		White	14,279	2518.02	77.47	37.9	40.6	21.5	62.1
		Two or More Races	947	2493.01	85.04	50.4	36.1	13.5	49.6
	FRL	Yes	10,726	2470.69	82.64	61.9	30.1	8.0	38.1
		No	10,986	2530.82	74.12	31.0	42.8	26.1	69.0
	LEP	Yes	3,376	2457.02	83.50	68.6	25.6	5.8	31.4
		No	18,336	2509.23	81.55	42.2	38.6	19.3	57.8
	SPED	Yes	3,667	2438.56	88.34	75.9	18.4	5.7	24.1
		No	18,045	2513.83	77.14	40.2	40.2	19.5	59.8
5		Overall	22,215	2514.51	81.93	54.1	31.3	14.7	45.9
	Gender	Female	10,768	2520.27	77.93	52.2	32.7	15.2	47.8
		Male	11,447	2509.08	85.18	55.9	29.9	14.2	44.1
	Ethnicity	AI/AN	281	2465.15	82.72	78.6	16.7	4.6	21.4
		Asian	635	2523.12	88.10	50.2	29.6	20.2	49.8
		Black	1,357	2462.47	85.90	78.0	16.9	5.1	22.0

Table D.1: Achievement Level Distributions and Scale Score Descriptive Statistics by Demographics–ELA, cont.

		Hispanic	4,396	2481.78	80.16	70.9	23.3	5.8	29.1
		NH/PI	34	2502.82	80.77	55.9	29.4	14.7	44.1
		White	14,542	2530.43	76.30	46.2	35.4	18.4	53.8
		Two or More Races	970	2505.97	80.66	58.2	30.0	11.8	41.8
	FRL	Yes	11,056	2484.82	80.30	69.7	23.9	6.4	30.3
		No	11,159	2543.92	72.37	38.7	38.6	22.8	61.3
	LEP	Yes	3,331	2471.12	79.20	76.4	20.0	3.6	23.6
		No	18,884	2522.16	80.00	50.2	33.2	16.6	49.8
	SPED	Yes	3,545	2448.15	82.44	83.9	12.4	3.7	16.1
		No	18,670	2527.11	75.52	48.4	34.8	16.7	51.6
6		Overall	22,295	2526.94	79.32	54.3	30.0	15.7	45.7
	Gender	Female	10,850	2533.02	76.20	51.5	31.7	16.8	48.5
		Male	11,445	2521.19	81.76	56.9	28.4	14.6	43.1
	Ethnicity	AI/AN	287	2470.32	82.67	79.4	16.4	4.2	20.6
		Asian	583	2533.85	85.78	49.4	29.7	20.9	50.6
		Black	1,304	2476.12	81.77	77.5	17.9	4.7	22.5
		Hispanic	4,509	2496.35	78.53	71.2	21.8	7.0	28.8
		NH/PI	33	2508.76	91.87	60.6	30.3	9.1	39.4
		White	14,666	2542.28	73.45	46.4	34.1	19.5	53.6
		Two or More Races	913	2518.24	81.58	58.9	27.6	13.5	41.1
	FRL	Yes	10,922	2498.59	78.78	69.6	22.8	7.6	30.4
		No	11,373	2554.18	69.73	39.6	36.9	23.5	60.4
	LEP	Yes	3,048	2479.76	77.12	79.0	16.8	4.2	21.0
		No	19,247	2534.42	77.06	50.3	32.1	17.5	49.7
	SPED	Yes	3,422	2455.21	81.55	85.7	10.6	3.7	14.3
		No	18,873	2539.95	71.59	48.6	33.6	17.9	51.4
7		Overall	22,087	2537.66	76.12	55.3	35.7	9.0	44.7
	Gender	Female	10,665	2543.77	73.66	51.8	38.3	9.8	48.2
		Male	11,422	2531.96	77.91	58.6	33.2	8.2	41.4
	Ethnicity	AI/AN	269	2490.22	76.54	81.8	16.4	1.9	18.2
		Asian	591	2549.84	83.51	46.5	38.9	14.6	53.5
		Black	1,277	2491.17	81.39	78.5	18.5	3.1	21.5
		Hispanic	4,168	2508.34	77.93	70.9	25.7	3.5	29.1
		NH/PI	35	2527.03	73.50	54.3	40.0	5.7	45.7
		White	14,827	2550.70	70.11	48.7	40.3	11.0	51.3
		Two or More Races	920	2531.46	79.81	57.9	34.0	8.0	42.1
	FRL	Yes	10,385	2510.86	77.28	70.0	26.1	3.9	30.0
		No	11,702	2561.45	66.57	42.3	44.2	13.5	57.7
	LEP	Yes	2,307	2482.09	75.72	83.4	15.7	0.9	16.6
		No	19,780	2544.14	73.48	52.0	38.0	9.9	48.0
	SPED	Yes	3,192	2471.48	80.17	85.8	11.9	2.2	14.2
		No	18,895	2548.84	69.44	50.2	39.7	10.1	49.8
8		Overall	20,689	2555.18	74.26	49.4	37.7	12.9	50.6
	Gender	Female	9,884	2562.55	71.22	45.7	39.7	14.6	54.3
		Male	10,805	2548.44	76.33	52.8	35.9	11.3	47.2

Table D.1: Achievement Level Distributions and Scale Score Descriptive Statistics by Demographics–ELA, cont.

	Ethnicity	AI/AN	279	2511.59	74.98	73.5	21.1	5.4	26.5
		Asian	498	2559.36	82.10	45.2	37.6	17.3	54.8
		Black	1,195	2510.92	77.45	72.9	23.1	4.0	27.1
		Hispanic	3,939	2525.09	75.30	66.6	28.2	5.2	33.4
		NH/PI	38	2540.16	85.93	55.3	34.2	10.5	44.7
		White	13,957	2568.61	68.68	42.1	42.0	15.9	57.9
		Two or More Races	783	2548.39	76.60	50.6	38.3	11.1	49.4
	FRL	Yes	9,561	2528.91	74.31	64.1	30.0	5.8	35.9
		No	11,128	2577.75	66.39	36.7	44.3	19.0	63.3
	LEP	Yes	1,546	2487.35	75.91	84.9	13.8	1.3	15.1
		No	19,143	2560.66	71.37	46.5	39.6	13.8	53.5
	SPED	Yes	2,749	2487.83	76.70	83.8	13.9	2.3	16.2
		No	17,940	2565.50	68.25	44.1	41.4	14.5	55.9

*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.

Table D.2: Achievement Level Distributions and Scale Score Descriptive Statistics by Demographics–Mathematics

				Descriptive Statistics		Percent of Students in Each Achievement Level**			
Grade	**Demographic Sub-Group***		**N**	**Mean**	**SD**	**Level 3**	**Level 2**	**Level 1**	**L2 + L1**
3		Overall	21,762	1183.16	78.89	52.8	37.8	9.4	47.2
	Gender	Female	10,613	1178.23	75.25	55.4	37.3	7.3	44.6
		Male	11,149	1187.87	81.92	50.3	38.2	11.4	49.7
	Ethnicity	AI/AN	284	1119.83	71.76	83.5	14.4	2.1	16.5
		Asian	696	1189.28	92.10	50.9	35.9	13.2	49.1
		Black	1,307	1128.28	71.20	79.2	19.1	1.8	20.8
		Hispanic	4,213	1146.63	70.49	73.0	24.4	2.7	27.0
		NH/PI	36	1157.64	75.37	69.4	22.2	8.3	30.6
		White	14,218	1201.11	74.17	43.3	44.6	12.1	56.7
		Two or More Races	1,008	1168.41	80.69	61.5	30.6	7.9	38.5
	FRL	Yes	10,808	1151.50	72.11	70.0	26.5	3.4	30.0
		No	10,954	1214.40	72.58	35.8	48.9	15.3	64.2
	LEP	Yes	3,535	1142.57	71.05	74.6	22.9	2.5	25.4
		No	18,227	1191.04	77.91	48.6	40.7	10.7	51.4
	SPED	Yes	3,574	1132.99	80.26	76.7	19.5	3.8	23.3
		No	18,188	1193.02	74.76	48.1	41.4	10.5	51.9
4		Overall	21,677	1212.58	74.42	54.3	37.6	8.1	45.7
	Gender	Female	10,556	1207.91	70.43	57.4	36.3	6.3	42.6
		Male	11,121	1217.02	77.77	51.4	38.8	9.8	48.6
	Ethnicity	AI/AN	254	1154.49	67.59	84.3	14.2	1.6	15.7
		Asian	655	1220.98	86.81	53.4	31.9	14.7	46.6
		Black	1,244	1156.23	66.13	84.0	14.2	1.8	16.0
		Hispanic	4,280	1180.11	67.14	72.3	25.5	2.1	27.7
		NH/PI	35	1199.80	74.31	51.4	42.9	5.7	48.6

Table D.2: Achievement Level Distributions and Scale Score Descriptive Statistics by Demographics–Mathematics, cont.

		White	14,264	1229.04	70.51	45.3	44.2	10.5	54.7
		Two or More Races	945	1195.69	70.16	62.5	32.9	4.6	37.5
	FRL	Yes	10,718	1183.88	68.73	70.6	26.3	3.1	29.4
		No	10,959	1240.65	68.87	38.4	48.6	13.1	61.6
	LEP	Yes	3,375	1174.65	68.41	75.9	21.6	2.5	24.1
		No	18,302	1219.57	73.37	50.3	40.5	9.2	49.7
	SPED	Yes	3,640	1164.09	72.13	79.6	17.9	2.5	20.4
		No	18,037	1222.37	70.97	49.2	41.5	9.3	50.8
5		Overall	22,191	1228.94	72.13	54.4	38.1	7.5	45.6
	Gender	Female	10,750	1226.32	68.25	56.6	37.2	6.2	43.4
		Male	11,441	1231.39	75.52	52.3	38.9	8.8	47.7
	Ethnicity	AI/AN	279	1180.97	64.66	84.2	14.0	1.8	15.8
		Asian	635	1246.17	89.90	49.8	34.0	16.2	50.2
		Black	1,353	1173.45	68.06	82.7	15.4	1.8	17.3
		Hispanic	4,389	1197.84	63.99	73.9	23.6	2.5	26.1
		NH/PI	34	1226.38	75.64	52.9	38.2	8.8	47.1
		White	14,534	1244.52	68.00	45.0	45.5	9.5	55.0
		Two or More Races	967	1216.09	69.17	62.2	32.4	5.5	37.8
	FRL	Yes	11,064	1200.65	65.12	71.5	26.0	2.6	28.5
		No	11,127	1257.07	67.62	37.4	50.1	12.5	62.6
	LEP	Yes	3,331	1190.10	63.86	77.7	20.3	2.0	22.3
		No	18,860	1235.79	71.33	50.3	41.2	8.5	49.7
	SPED	Yes	3,529	1176.10	67.66	83.2	15.1	1.8	16.8
		No	18,662	1238.93	68.51	49.0	42.4	8.6	51.0
6		Overall	22,276	1237.61	73.73	52.9	39.0	8.1	47.1
	Gender	Female	10,845	1237.25	71.34	53.1	39.3	7.5	46.9
		Male	11,431	1237.95	75.92	52.7	38.7	8.6	47.3
	Ethnicity	AI/AN	284	1180.76	71.74	79.9	19.7	0.4	20.1
		Asian	585	1252.66	87.00	46.3	38.1	15.6	53.7
		Black	1,306	1181.28	71.19	80.6	17.5	1.8	19.4
		Hispanic	4,505	1207.57	67.85	70.6	26.9	2.5	29.4
		NH/PI	33	1212.97	83.67	66.7	30.3	3.0	33.3
		White	14,652	1253.09	69.05	44.3	45.4	10.3	55.7
		Two or More Races	911	1226.76	73.20	59.8	33.5	6.7	40.2
	FRL	Yes	10,930	1209.42	69.44	69.2	27.6	3.2	30.8
		No	11,346	1264.76	67.27	37.2	49.9	12.8	62.8
	LEP	Yes	3,046	1195.22	67.05	77.6	20.7	1.7	22.4
		No	19,230	1244.32	72.49	49.0	41.9	9.1	51.0
	SPED	Yes	3,416	1175.79	68.63	84.5	13.9	1.6	15.5
		No	18,860	1248.80	68.92	47.2	43.5	9.3	52.8
7		Overall	22,050	1245.76	68.34	53.7	38.4	7.9	46.3
	Gender	Female	10,646	1243.91	64.94	54.9	38.1	7.0	45.1
		Male	11,404	1247.48	71.33	52.6	38.6	8.8	47.4
	Ethnicity	AI/AN	268	1199.50	55.88	79.5	19.4	1.1	20.5
		Asian	593	1273.50	88.66	43.3	35.4	21.2	56.7

Table D.2: Achievement Level Distributions and Scale Score Descriptive Statistics by Demographics–Mathematics, cont.

		Black	1,274	1198.15	62.44	80.5	18.0	1.5	19.5
		Hispanic	4,159	1219.31	61.07	70.4	26.8	2.7	29.6
		NH/PI	35	1247.26	49.25	60.0	34.3	5.7	40.0
		White	14,808	1257.46	65.56	46.4	44.0	9.6	53.6
		Two or More Races	913	1238.46	69.43	58.2	35.3	6.6	41.8
	FRL	Yes	10,376	1220.05	60.77	69.8	27.3	2.9	30.2
		No	11,674	1268.61	66.55	39.4	48.2	12.4	60.6
	LEP	Yes	2,307	1201.82	56.96	80.5	18.3	1.2	19.5
		No	19,743	1250.89	67.71	50.6	40.7	8.7	49.4
	SPED	Yes	3,175	1193.64	60.07	83.8	14.8	1.4	16.2
		No	18,875	1254.53	65.69	48.6	42.3	9.0	51.4
8		Overall	20,659	1259.13	71.82	54.6	37.7	7.7	45.4
	Gender	Female	9,878	1260.36	68.14	53.7	39.4	7.0	46.3
		Male	10,781	1257.99	75.02	55.5	36.1	8.3	44.5
	Ethnicity	AI/AN	276	1211.84	66.57	80.8	17.4	1.8	19.2
		Asian	497	1279.90	89.02	43.9	39.0	17.1	56.1
		Black	1,195	1209.69	65.83	81.8	16.1	2.1	18.2
		Hispanic	3,937	1229.47	65.57	71.9	25.3	2.9	28.1
		NH/PI	39	1249.69	64.02	59.0	35.9	5.1	41.0
		White	13,939	1272.38	68.48	47.1	43.5	9.4	52.9
		Two or More Races	776	1251.55	72.62	58.8	35.6	5.7	41.2
	FRL	Yes	9,568	1230.78	64.97	71.7	25.5	2.8	28.3
		No	11,091	1283.57	68.38	40.0	48.2	11.9	60.0
	LEP	Yes	1,552	1203.31	59.03	86.2	13.1	0.7	13.8
		No	19,107	1263.66	70.86	52.1	39.7	8.2	47.9
	SPED	Yes	2,730	1197.24	62.41	86.8	12.0	1.2	13.2
		No	17,929	1268.55	68.40	49.7	41.6	8.7	50.3

*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.

Table D.3: Raw Score Descriptive Statistics by Demographics–Science

					Descriptive Statistics	
Grade	Form	Demographic Sub-Group*		N	Mean	SD
5	A		Overall	4,233	11.70	4.29
		Gender	Female	2,057	11.59	4.17
			Male	2,176	11.80	4.41
		Ethnicity	AI/AN	57	8.47	3.42
			Asian	113	11.92	4.37
			Black	236	7.89	3.63
			Hispanic	802	9.83	4.05
			NH/PI	3	13.67	4.04
			White	2,840	12.64	4.04
			Two or More Races	182	11.02	4.28
		FRL	Yes	2,088	10.17	4.20
			No	2,145	13.19	3.84

Table D.3: Raw Score Descriptive Statistics by Demographics–Science, cont.

		LEP	Yes	590	9.08	3.94
			No	3,643	12.12	4.20
		SPED	Yes	666	8.60	4.04
			No	3,567	12.28	4.09
5	B		Overall	3,056	9.15	4.14
		Gender	Female	1,491	8.96	3.96
			Male	1,565	9.33	4.29
		Ethnicity	AI/AN	32	8.09	3.44
			Asian	88	10.08	4.49
			Black	196	6.63	3.78
			Hispanic	640	7.48	3.52
			NH/PI	3	7.33	6.11
			White	1,961	9.98	4.08
			Two or More Races	136	8.27	4.00
		FRL	Yes	1,561	7.84	3.81
			No	1,495	10.51	4.02
		LEP	Yes	496	7.06	3.56
			No	2,560	9.55	4.12
		SPED	Yes	485	6.56	3.65
			No	2,571	9.64	4.04
5	C		Overall	3,580	11.59	4.50
		Gender	Female	1,721	11.43	4.32
			Male	1,859	11.73	4.65
		Ethnicity	AI/AN	37	9.11	3.69
			Asian	109	11.93	4.75
			Black	239	8.48	4.30
			Hispanic	712	9.60	4.26
			NH/PI	5	13.40	4.88
			White	2,331	12.56	4.24
			Two or More Races	146	11.16	4.17
		FRL	Yes	1,788	9.93	4.27
			No	1,791	13.24	4.09
		LEP	Yes	557	9.09	4.12
			No	3,022	12.05	4.41
		SPED	Yes	607	8.46	4.09
			No	2,973	12.23	4.31
5	D		Overall	4,280	10.95	3.55
		Gender	Female	2,068	10.79	3.43
			Male	2,212	11.09	3.64
		Ethnicity	AI/AN	50	8.68	3.57
			Asian	121	11.17	3.64
			Black	252	8.13	3.50
			Hispanic	846	9.67	3.57
			NH/PI	6	9.83	3.82
			White	2,829	11.63	3.29
			Two or More Races	176	10.70	3.51

Table D.3: Raw Score Descriptive Statistics by Demographics–Science, cont.

		FRL	Yes	2,149	9.76	3.51
			No	2,131	12.14	3.15
		LEP	Yes	610	9.03	3.59
			No	3,670	11.26	3.44
		SPED	Yes	676	8.74	3.48
			No	3,604	11.36	3.40
5	E		Overall	3,001	12.88	3.92
		Gender	Female	1,456	12.84	3.86
			Male	1,545	12.91	3.98
		Ethnicity	AI/AN	42	10.90	3.16
			Asian	84	12.40	4.29
			Black	195	10.41	3.97
			Hispanic	609	11.32	3.92
			NH/PI	10	10.70	4.55
			White	1,914	13.77	3.62
			Two or More Races	147	11.93	3.68
		FRL	Yes	1,473	11.51	3.97
			No	1,527	14.20	3.39
		LEP	Yes	480	10.84	3.95
			No	2,521	13.26	3.80
		SPED	Yes	475	10.11	4.15
			No	2,526	13.40	3.65
5	F		Overall	4,051	10.02	4.61
		Gender	Female	1,958	10.06	4.48
			Male	2,093	9.98	4.73
		Ethnicity	AI/AN	62	6.71	3.77
			Asian	120	9.88	4.80
			Black	237	7.27	4.06
			Hispanic	776	7.96	4.24
			NH/PI	6	8.67	3.56
			White	2,673	10.98	4.44
			Two or More Races	175	9.47	4.66
		FRL	Yes	1,992	8.33	4.28
			No	2,057	11.65	4.33
		LEP	Yes	590	7.52	4.08
			No	3,459	10.44	4.57
		SPED	Yes	648	7.17	4.18
			No	3,403	10.56	4.49
8	A		Overall	3,067	7.93	2.97
		Gender	Female	1,475	7.87	2.91
			Male	1,592	7.98	3.02
		Ethnicity	AI/AN	53	6.98	2.87
			Asian	74	8.01	3.16
			Black	198	5.78	2.80
			Hispanic	601	6.69	2.80
			NH/PI	7	7.00	3.21

Table D.3: Raw Score Descriptive Statistics by Demographics–Science, cont.

			White	2,005	8.54	2.82
			Two or More Races	129	7.74	2.65
		FRL	Yes	1,427	6.89	2.85
			No	1,640	8.82	2.77
		LEP	Yes	241	5.37	2.63
			No	2,826	8.14	2.89
		SPED	Yes	386	5.39	2.78
			No	2,681	8.29	2.81
8	B		Overall	4,242	7.78	4.16
		Gender	Female	1,993	7.80	4.14
			Male	2,249	7.76	4.18
		Ethnicity	AI/AN	49	4.88	4.13
			Asian	93	7.75	4.45
			Black	196	4.56	3.56
			Hispanic	780	5.91	3.51
			NH/PI	11	7.27	4.58
			White	2,978	8.56	4.07
			Two or More Races	133	7.15	3.86
		FRL	Yes	1,993	6.35	3.81
			No	2,247	9.05	4.04
		LEP	Yes	329	4.39	3.02
			No	3,911	8.07	4.12
		SPED	Yes	602	4.87	3.52
			No	3,640	8.26	4.06
8	C		Overall	3,900	10.42	4.49
		Gender	Female	1,866	10.44	4.47
			Male	2,034	10.40	4.51
		Ethnicity	AI/AN	48	8.67	4.95
			Asian	99	11.66	4.84
			Black	198	7.36	3.86
			Hispanic	717	8.11	4.13
			NH/PI	7	11.43	1.13
			White	2,692	11.29	4.29
			Two or More Races	139	9.53	4.24
		FRL	Yes	1,809	8.81	4.23
			No	2,091	11.81	4.24
		LEP	Yes	291	6.62	3.48
			No	3,609	10.73	4.42
		SPED	Yes	506	6.79	3.89
			No	3,394	10.96	4.32
8	D		Overall	3,352	6.91	3.11
		Gender	Female	1,605	6.99	3.03
			Male	1,747	6.84	3.18
		Ethnicity	AI/AN	40	5.03	2.81
			Asian	75	6.60	3.13
			Black	203	4.99	2.88

Table D.3: Raw Score Descriptive Statistics by Demographics–Science, cont.

			Hispanic	637	5.88	2.98
			NH/PI	5	6.40	2.30
			White	2,260	7.44	3.00
			Two or More Races	132	6.70	3.25
		FRL	Yes	1,552	5.86	2.94
			No	1,800	7.83	2.95
		LEP	Yes	233	4.36	2.51
			No	3,119	7.10	3.06
		SPED	Yes	451	4.63	2.67
			No	2,901	7.27	3.02
8	E		Overall	3,069	4.88	2.86
		Gender	Female	1,479	4.68	2.71
			Male	1,590	5.06	2.99
		Ethnicity	AI/AN	47	3.19	1.91
			Asian	74	5.01	2.94
			Black	203	3.37	2.13
			Hispanic	601	3.79	2.50
			NH/PI	7	5.29	2.36
			White	2,014	5.40	2.88
			Two or More Races	123	4.62	3.03
		FRL	Yes	1,391	3.86	2.44
			No	1,678	5.72	2.91
		LEP	Yes	234	3.00	2.13
			No	2,835	5.03	2.86
		SPED	Yes	411	3.08	2.06
			No	2,658	5.16	2.87
8	F		Overall	3,063	8.22	3.90
		Gender	Female	1,475	8.05	3.80
			Male	1,588	8.39	4.00
		Ethnicity	AI/AN	43	5.16	2.99
			Asian	85	9.13	4.58
			Black	197	5.66	3.37
			Hispanic	598	6.89	3.56
			NH/PI	1	5.00	.
			White	2,012	8.95	3.80
			Two or More Races	127	7.37	3.62
		FRL	Yes	1,397	6.84	3.63
			No	1,666	9.39	3.75
		LEP	Yes	217	5.57	3.48
			No	2,846	8.43	3.86
		SPED	Yes	414	5.27	3.35
			No	2,649	8.69	3.78

*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 E. Marginal Reliability by Demographics

Table E.1: Marginal Reliability by Demographics - ELA

Grade 3						
Grade	**Demographic**	**Sub-Group**	**N**	**Variance**	**MSE**	**Marginal Reliability**
3		Overall	21784	7629.2	896.0	0.88
	Gender	Female	10626	7247.5	889.6	0.88
		Male	11158	7922.4	902.0	0.89
	Ethnicity	AI/AN	287	7155.6	972.1	0.86
		Asian	698	8634.6	904.8	0.90
		Black	1311	8054.3	957.9	0.88
		Hispanic	4218	7350.3	928.0	0.87
		NH/PI	36	7542.1	922.0	0.88
		White	14226	6513.9	878.5	0.87
		Two or More Races	1005	7654.6	899.6	0.88
	FRL	Yes	10810	7415.1	920.5	0.88
		No	10971	5822.7	871.9	0.85
	LEP	Yes	3540	7373.2	935.1	0.87
		No	18241	7205.7	888.4	0.88
	SPED	Yes	3596	8244.9	963.7	0.88
		No	18188	6761.8	882.6	0.87
Grade 4						
Grade	**Demographic**	**Sub-Group**	**N**	**Variance**	**MSE**	**Marginal Reliability**
4		Overall	21714	7057.9	882.7	0.88
	Gender	Female	10573	6607.4	881.6	0.87
		Male	11141	7407.3	883.8	0.88
	Ethnicity	AI/AN	255	6396.1	891.9	0.86
		Asian	658	8413.4	901.3	0.89
		Black	1251	8046.9	902.0	0.89
		Hispanic	4287	6814.7	878.9	0.87
		NH/PI	35	7350.6	856.5	0.88
		White	14280	6002.3	881.0	0.85
		Two or More Races	947	7231.2	885.4	0.88
	FRL	Yes	10727	6830.0	876.2	0.87
		No	10986	5494.1	889.1	0.84
	LEP	Yes	3376	6971.6	885.3	0.87
		No	18337	6650.2	882.2	0.87
	SPED	Yes	3667	7803.9	909.5	0.88
		No	18047	5949.9	877.2	0.85
Grade 5						
Grade	**Demographic**	**Sub-Group**	**N**	**Variance**	**MSE**	**Marginal Reliability**
5		Overall	22225	6713.1	872.8	0.87
	Gender	Female	10773	6073.1	864.4	0.86
		Male	11452	7255.0	880.6	0.88
	Ethnicity	AI/AN	281	6842.7	924.5	0.86
		Asian	636	7762.1	888.7	0.89

Table E.1: Marginal Reliability by Demographics - ELA, cont.

		Black	1357	7378.6	938.8	0.87
		Hispanic	4399	6425.3	895.7	0.86
		NH/PI	34	6523.3	862.8	0.87
		White	14544	5821.2	857.6	0.85
		Two or More Races	971	6505.8	878.5	0.86
	FRL	Yes	11062	6448.8	892.9	0.86
		No	11160	5237.4	852.8	0.84
	LEP	Yes	3334	6272.0	910.2	0.85
		No	18888	6400.6	866.2	0.86
	SPED	Yes	3549	6797.1	957.5	0.86
		No	18676	5702.6	856.7	0.85
Grade 6						
Grade	**Demographic**	**Sub-Group**	**N**	**Variance**	**MSE**	**Marginal Reliability**
6		Overall	22300	6292.3	825.4	0.87
	Gender	Female	10852	5806.7	817.2	0.86
		Male	11448	6685.0	833.2	0.88
	Ethnicity	AI/AN	287	6834.7	900.8	0.87
		Asian	584	7357.4	845.4	0.89
		Black	1305	6685.9	884.2	0.87
		Hispanic	4509	6166.7	844.6	0.86
		NH/PI	33	8440.4	873.8	0.90
		White	14669	5395.5	811.3	0.85
		Two or More Races	913	6654.7	836.7	0.87
	FRL	Yes	10927	6205.8	842.6	0.86
		No	11373	4862.0	809.0	0.83
	LEP	Yes	3050	5948.2	865.0	0.85
		No	19250	5938.7	819.2	0.86
	SPED	Yes	3424	6650.6	927.7	0.86
		No	18876	5125.3	806.9	0.84
Grade 7						
Grade	**Demographic**	**Sub-Group**	**N**	**Variance**	**MSE**	**Marginal Reliability**
7		Overall	22093	5793.7	841.7	0.86
	Gender	Female	10671	5425.8	835.3	0.85
		Male	11422	6070.2	847.7	0.86
	Ethnicity	AI/AN	269	5858.7	878.9	0.85
		Asian	591	6974.6	862.9	0.88
		Black	1277	6625.1	900.3	0.86
		Hispanic	4169	6072.5	865.2	0.86
		NH/PI	35	5403.0	834.0	0.85
		White	14829	4915.8	827.8	0.83
		Two or More Races	920	6369.3	855.0	0.87
	FRL	Yes	10388	5972.4	860.8	0.86
		No	11702	4431.6	824.8	0.81
	LEP	Yes	2308	5733.0	898.0	0.84
		No	19782	5398.8	835.2	0.85
	SPED	Yes	3193	6427.1	932.4	0.85

Table E.1: Marginal Reliability by Demographics - ELA, cont.

		No	18900	4821.9	826.4	0.83
Grade 8						
Grade	**Demographic**	**Sub-Group**	**N**	**Variance**	**MSE**	**Marginal Reliability**
8		Overall	20699	5515.3	860.6	0.84
	Gender	Female	9890	5072.3	853.5	0.83
		Male	10809	5825.9	867.1	0.85
	Ethnicity	AI/AN	279	5622.4	895.3	0.84
		Asian	500	6739.8	878.4	0.87
		Black	1195	5997.8	906.6	0.85
		Hispanic	3943	5670.3	881.3	0.84
		NH/PI	38	7383.9	879.3	0.88
		White	13961	4716.7	849.1	0.82
		Two or More Races	783	5867.9	868.3	0.85
	FRL	Yes	9570	5521.7	875.8	0.84
		No	11129	4408.2	847.6	0.81
	LEP	Yes	1548	5762.9	945.0	0.84
		No	19151	5094.0	853.8	0.83
	SPED	Yes	2749	5883.1	945.6	0.84
		No	17950	4657.7	847.6	0.82

*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

Grade 3						
Grade	**Demographic**	**Sub-Group**	**N**	**Variance**	**MSE**	**Marginal Reliability**
3		Overall	21763	6223	526.4	0.92
	Gender	Female	10613	5662.9	520	0.91
		Male	11150	6711.3	532.5	0.92
	Ethnicity	AI/AN	284	5150	535.9	0.9
		Asian	696	8481.7	577.5	0.93
		Black	1307	5069.1	531.8	0.9
		Hispanic	4213	4968.9	522	0.89
		NH/PI	36	5680.9	525.8	0.91
		White	14218	5501.3	524.5	0.9
		Two or More Races	1008	6510.4	526.2	0.92
	FRL	Yes	10808	5199.8	521	0.9
		No	10954	5268.1	531.7	0.9
	LEP	Yes	3535	5048.6	525.1	0.9
		No	18227	6069.4	526.6	0.91
	SPED	Yes	3575	6442.1	540.6	0.92
		No	18188	5588.4	523.6	0.91
Grade 4						
Grade	**Demographic**	**Sub-Group**	**N**	**Variance**	**MSE**	**Marginal Reliability**
4		Overall	21680	5539.0	522.3	0.91
	Gender	Female	10557	4959.8	519.2	0.90
		Male	11123	6048.9	525.2	0.91

Table E.2: Marginal Reliability by Demographics - Mathematics, cont.

	Ethnicity	AI/AN	254	4567.9	569.0	0.88
		Asian	655	7535.6	536.3	0.93
		Black	1244	4373.8	569.6	0.87
		Hispanic	4280	4507.4	539.4	0.88
		NH/PI	36	5521.6	517.3	0.91
		White	14265	4972.1	511.2	0.90
		Two or More Races	945	4923.0	528.1	0.89
	FRL	Yes	10719	4724.3	537.0	0.89
		No	10960	4743.0	507.9	0.89
	LEP	Yes	3376	4679.5	547.7	0.88
		No	18303	5383.6	517.6	0.90
	SPED	Yes	3640	5202.6	566.3	0.89
		No	18040	5036.9	513.4	0.90
Grade 5						
Grade	**Demographic**	**Sub-Group**	**N**	**Variance**	**MSE**	**Marginal Reliability**
5		Overall	22198	5202.9	525.4	0.90
	Gender	Female	10752	4657.4	516.9	0.89
		Male	11446	5703.4	533.3	0.91
	Ethnicity	AI/AN	279	4181.4	534.2	0.87
		Asian	635	8082.5	608.3	0.92
		Black	1353	4632.8	537.9	0.88
		Hispanic	4392	4094.6	516.1	0.87
		NH/PI	34	5721.4	521.0	0.91
		White	14535	4623.4	523.6	0.89
		Two or More Races	968	4784.1	519.0	0.89
	FRL	Yes	11068	4240.5	515.3	0.88
		No	11128	4573.1	535.3	0.88
	LEP	Yes	3332	4078.4	520.9	0.87
		No	18864	5088.3	526.1	0.90
	SPED	Yes	3531	4578.4	536.7	0.88
		No	18667	4693.4	523.2	0.89
Grade 6						
Grade	**Demographic**	**Sub-Group**	**N**	**Variance**	**MSE**	**Marginal Reliability**
6		Overall	22280	5435.9	517.0	0.91
	Gender	Female	10847	5089.8	514.0	0.90
		Male	11433	5764.6	519.8	0.91
	Ethnicity	AI/AN	284	5146.0	566.3	0.89
		Asian	586	7568.5	534.4	0.93
		Black	1306	5067.9	562.4	0.89
		Hispanic	4507	4603.9	530.6	0.88
		NH/PI	33	7001.4	544.5	0.92
		White	14653	4768.3	506.5	0.89
		Two or More Races	911	5357.9	525.4	0.90
	FRL	Yes	10932	4821.3	531.8	0.89
		No	11348	4525.8	502.6	0.89
	LEP	Yes	3049	4495.6	542.4	0.88

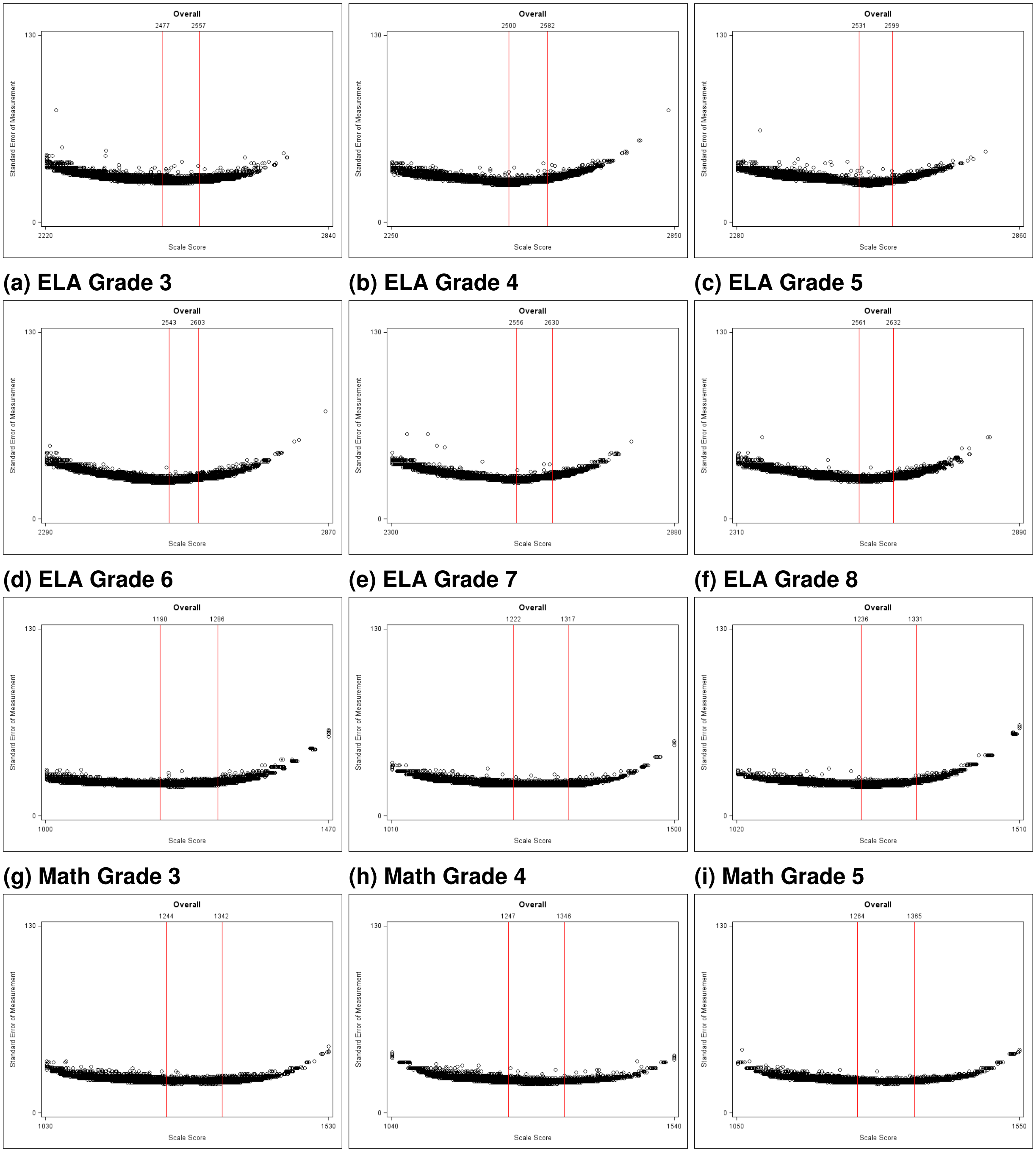
Table E.2: Marginal Reliability by Demographics - Mathematics, cont.

		No	19231	5255.5	512.9	0.90
	SPED	Yes	3416	4710.6	565.1	0.88
		No	18864	4750.1	508.2	0.89
Grade 7						
Grade	**Demographic**	**Sub-Group**	**N**	**Variance**	**MSE**	**Marginal Reliability**
7		Overall	22058	4670.3	519.9	0.89
	Gender	Female	10651	4217.0	516.8	0.88
		Male	11407	5087.8	522.9	0.90
	Ethnicity	AI/AN	268	3122.3	566.9	0.82
		Asian	593	7861.1	536.6	0.93
		Black	1274	3899.3	573.9	0.85
		Hispanic	4160	3730.0	540.7	0.86
		NH/PI	35	2425.5	492.5	0.80
		White	14812	4297.6	507.6	0.88
		Two or More Races	914	4821.1	527.2	0.89
	FRL	Yes	10381	3693.3	539.7	0.85
		No	11675	4429.5	502.4	0.89
	LEP	Yes	2308	3244.4	563.8	0.83
		No	19748	4585.1	514.8	0.89
	SPED	Yes	3178	3608.9	580.1	0.84
		No	18880	4315.2	509.8	0.88
Grade 8						
Grade	**Demographic**	**Sub-Group**	**N**	**Variance**	**MSE**	**Marginal Reliability**
8		Overall	20666	5158.1	516.2	0.90
	Gender	Female	9882	4642.5	510.9	0.89
		Male	10784	5628.3	521.0	0.91
	Ethnicity	AI/AN	276	4432.1	554.5	0.87
		Asian	497	7925.0	532.2	0.93
		Black	1195	4333.8	556.9	0.87
		Hispanic	3939	4298.8	534.4	0.88
		NH/PI	39	4098.4	510.3	0.88
		White	13943	4689.1	505.9	0.89
		2 or More Races	776	5274.0	521.4	0.90
	FRL	Yes	9571	4221.6	532.6	0.87
		No	11094	4675.7	502.0	0.89
	LEP	Yes	1552	3484.9	563.5	0.84
		No	19113	5020.6	512.3	0.90
	SPED	Yes	2731	3895.4	572.5	0.85
		No	17935	4678.6	507.6	0.89

*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



(a) ELA Grade 3

(b) ELA Grade 4

(c) ELA Grade 5

(d) ELA Grade 6

(e) ELA Grade 7

(f) ELA Grade 8

(g) Math Grade 3

(h) Math Grade 4

(i) Math Grade 5

(j) Math Grade 6

(k) Math Grade 7

(l) Math Grade 8