NSCAS NEBRASKA STUDENT-CENTERED ASSESSMENT SYSTEM

Spring 2018 NSCAS Summative ELA, Mathematics, and Science Technical Report

nwea

© 2018 Nebraska Department of Education. No part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of the publisher.

Table of Contents

List of Appendices

List of Tables

List of Figures

List of Abbreviations

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

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

SGL	School Group List
STARS	School-based Teacher-led Assessment and Reporting System
TAC	Technical Advisory Committee
TAM	Test Administration Manual
TCC	test characteristic curve
TEI	technology-enhanced item
TOA	theory of action
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

The 2018 Nebraska Student-Centered Assessment System (NSCAS) Summative technical report documents the processes and procedures implemented to support the Spring 2018 NSCAS Summative English Language Arts (ELA), Mathematics, and Science assessments by NWEA under the supervision of the Nebraska Department of Education (NDE). The technical report shows how the processes, methods applied, and results relate to the issues of validity and reliability and to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). Some principal information presented in this technical report is summarized below for each section in the technical report.

Section 1: Introduction

Students taking the ELA and Mathematics tests were placed into one of the following achievement levels: Developing, On Track, or College and Career Readiness (CCR) Benchmark. Students taking the Science tests were placed into one of the following achievement levels: Below the Standards, Meets the Standards, or Exceeds the Standards.

The purposes of the 2018 NSCAS Summative assessments are to measure and report Nebraska students' depth of achievement regarding Nebraska's College and Career Ready Standards for ELA and Mathematics in Grades 3–8 and Nebraska's Science standards for Grades 5 and 8; to report if student achievement is sufficient academic proficiency in ELA and Mathematics to be on track for achieving college readiness; to measure students' annual progress toward college and career readiness in ELA and Mathematics; to inform teachers how student thinking differs along different areas of the scale as represented by the ALDs as information to support instructional planning; and to assess students' construct relevant achievement in ELA, Mathematics, and Science for all students and subgroups of students.

Section 2: Test Design and Development

Nebraska's College and Career Ready Standards have been adopted by the Nebraska State Board of Education for ELA, Mathematics, and Science in 2014, 2015, and 2017, respectively. The Spring 2018 NSCAS assessments were aligned to the Nebraska's College and Career Ready Standards for ELA and Mathematics in Grades 3 to 8. To fully represent the constructs being assessed by the NSCAS to determine if students are ready for college and careers, the adherence to specifications, common interpretations of the standards, and an agreed-upon approach for cognitive complexity across all item types were closely monitored during item, passage, and test development.

Section 3: Test Administration and Security

The Spring 2018 NSCAS Summative testing window was from March 19 to April 27, 2018. 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.

Section 4: Scoring and Reporting

The online ELA and Mathematics assessments were administered adaptively via NWEA's constraint-based engine, whereas the Science assessments, all paper-pencil tests, and all Spanish versions were administered as fixed-form. All tests were scored with maximum likelihood estimation (MLE) scoring. All steps of scoring for online and paper-pencil went through a quality control process. Score reports were produced and delivered at the individual student level, and aggregated reports were delivered at the school, district, and state levels.

Section 5: Psychometric Analyses

After the testing window was closed, a series of post-administration analyses were conducted to calibrate the items for ELA, Mathematics, and Science, including engine evaluation, classical item analyses, differential item functioning (DIF), item response theory (IRT) calibration, vertical scaling for ELA and Mathematics, and post-equating checks for Science.

Section 6: Standard Setting

The NDE held a standard setting for the NSCAS Mathematics assessments and a cut score review for ELA from July 26–28, 2018, using the Item-Descriptor (ID) Matching method to determine the cut scores delineating the Developing, On Track, and CCR Benchmark achievement levels. Standard setting panelists went through multiple rounds of ratings and vertical articulation to recommend cut scores. The recommended cut scores were presented to the Nebraska State Board of Education on August 2, 2018 for final approval.

Section 7: Test Results

The number of students who attempted at least one item are reported by demographics. Regarding achievement level distributions, 43–52% of students are at Developing and 48–57% of students are at On Track or CCR Benchmark for ELA. For Mathematics, 45–50% of students are at Developing and 50–55% of students are at On Track or CCR Benchmark. For Science, 30–33% of students are at Below the Standards and 67–70% are at Meets or Exceeds the Standards.

Section 8: Reliability

The reliability/precision of the 2018 NSCAS assessments was examined through analysis of measurement error in simulated and operational conditions, including constraint 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.

Section 9: Validity

As the technical report progresses, it covers the different phases of the testing cycle and the procedures and processes applied in the NSCAS, as well as the results. The last section revisits phases and summarizes relevant evidence and a rationale in support of any test score interpretations and indented uses based on the *Standards.*

Section 1: Introduction

The purpose of this technical report is to summarize the design, development, administration, technical processes, and results of the Spring 2018 Nebraska Student-Centered Assessment System (NSCAS) Summative assessments in English Language Arts (ELA) and Mathematics for Grades 3–8 and in Science for Grades 5 and 8 to support test users in evaluating the intended purposes, uses, and interpretations of the test scores. NSCAS was designed by the state of Nebraska with support from its vendor NWEA to meet the requirements of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) and the federal peer review requirements (U.S. Department of Education, 2015) with an emphasis on using a principled assessment design process.

1.1. NSCAS Overview

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

The Spring 2018 NSCAS Summative assessments were administered online with paper-pencil versions available as an accommodation. They included a variety of item types, including multiple-choice items and technology-enhanced items (TEIs). Student scores were reported as composite scale scores, reporting category scale scores, and achievement levels. The ELA and Mathematics assessments were administered using a multi-stage adaptive design, whereas the Science assessments were administered in fixed form online. For each grade and content area, there were two cut scores that distinguished among three achievement levels. The first cut score demarked the minimum level of performance considered to be proficient for accountability purposes. Students taking the ELA and Mathematics tests were placed into one of the following achievement levels:

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

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

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

Items for the ELA and Mathematics tests came from the item bank that the Nebraska Department of Education (NDE) and Nebraska educators built over the previous years with its previous vendor DRC. Items were aligned to the 2014 and 2015 College and Career Ready Standards, respectively. The tests also included newly developed field test items that will be added to the operational pool for the future depending on the field test data and data review. Items for the Science test came from the operational pool that the NDE had built over the previous years and were aligned to the 2010 Nebraska Legacy Standards in Science.

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. The 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, the NDE administered one writing assessment annually in Grades 4, 8, and 11. The NDE also provided an alternate assessment for students severely challenged by cognitive disabilities.

Legislative Bill 1157 passed by the 2008 Nebraska Legislature[1] required a single 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 2018 was the first administration of the NSCAS Summative ELA and Mathematics assessments and they were administered via computer adaptive testing (CAT), whereas Science continued to be administered as a fixed-form assessment.

1.3. Schedule of Major Events

Table 1.1 presents the major events regarding the development, administration, and reporting of the 2017–2018 NSCAS Summative assessment, including passage review, item writer workshops (IWWs), administration training, Technical Advisory Committee (TAC) meetings, testing windows, standard setting, data review, and delivery of reports.

Table 1.1. Schedule of Major Events

Event	Date(s)
ELA passage review	August 1–2, 2017
IWW in ELA and Mathematics	August 29–31, 2017
Content and bias review in ELA and Mathematics	September 19–21, 2017
Fall 2017 regional workshop	October 10–16, 2017
TAC meeting	October 31, 2017
TAC meeting	February 12–13, 2018
Summative test administration training	February 19–26, 2018

[1] http://www.legislature.ne.gov/laws/statutes.php?statute=79-760.03

Event	Date(s)
Spring 2018 testing window	March 19 – April 27, 2018
Mathematics ALD workshop	April 25–26, 2018
Make-up testing window	April 30 – May 4, 2018
Mathematics standard setting	July 26–28, 2018
ELA cut score review & ALD workshop	July 26–28, 2018
NDE and State Board Approval of standard setting cut scores	August 2, 2018
Data review with NDE	September 11–12, 2018
NDE and districts review preliminary data and submit updates	September 6–21, 2018
Delivery of online reports	October 23, 2018
Delivery of printed Individual Student Reports (ISRs)	October 29 – November 6, 2018

1.4. Theory of Action (TOA)

A theory of action (TOA) outlines the educational policy claims, goals, and actions for which the NDE designs test scores. TOA claims, goals, and intended uses describe the network of inferences that would be validated through policy-based studies such as evidence of consequential validity. The following are purposes of the 2018 NSCAS Summative assessments:

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

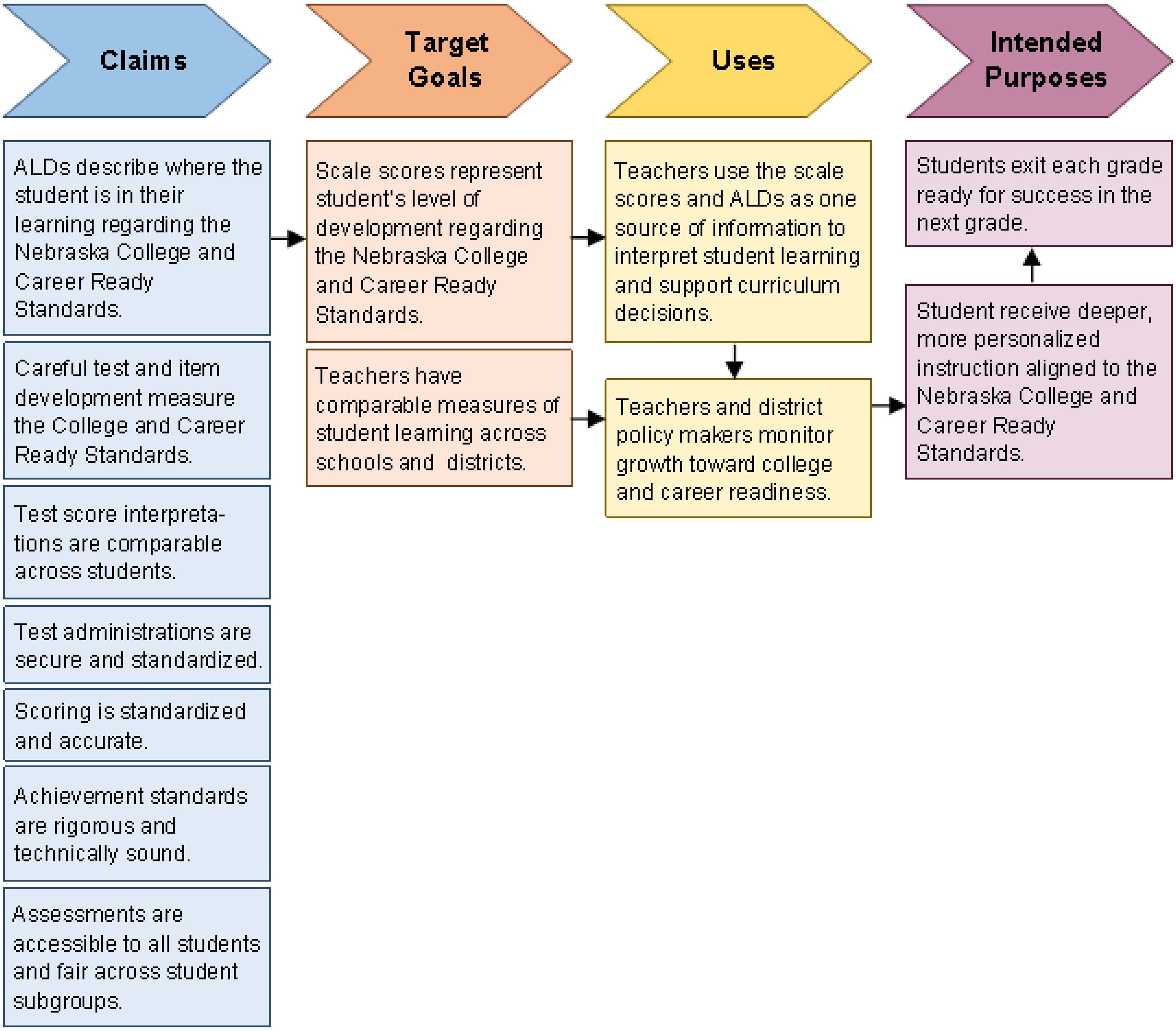
Intended uses of the NSCAS test results include the following:

- 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

Figure 1.1 presents the TOA for the NSCAS comprehensive assessment 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. NSCAS Theory of Action (TOA)



Claims

ALDs describe where the student is in their learning regarding the Nebraska College and Career Ready Standards.

Careful test and item development measure the 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.

Target Goals

Scale scores represent student's level of development regarding the Nebraska College and Career Ready Standards.

Teachers have comparable measures of student learning across schools and districts.

Uses

Teachers use the scale scores and ALDs as one source of information to interpret student learning and support curriculum decisions.

Teachers and district policy makers monitor growth toward college and career readiness.

Intended Purposes

Students exit each grade ready for success in the next grade.

Student receive deeper, more personalized instruction aligned to the Nebraska College and Career Ready Standards.

1.5. The Validity Argument for Test Score Interpretations

The design and validation of an assessment system requires careful development processes, especially when such assessments are intended to support interpretations regarding how student learning grows more sophisticated over time (Pellegrino, DiBello, & Goldman, 2016). The development of NSCAS to support integrated content and cognitive process claims about student knowledge, skills, and abilities (KSAs) was an iterative process that balanced the tensions of ingesting an item bank from a previous vendor while working to purposefully have the assessment reflect intended test interpretations with Reporting ALDs within a single year. Reporting ALDs are intended to provide the interpretive argument regarding what test scores mean.

Under a principled approach to test design based on Range ALDs, the evidence needed to draw a conclusion about where a student is in their learning of content is made explicit in the ALDs and items are developed according to those evidence pieces (Huff, Warner, & Schweid, 2016; Egan, Schneider, & Ferrara, 2012; Schneider & Johnson, 2018). With this model, increases in cognitive processing complexity (i.e., Depth of Knowledge (DOK) levels) are intended to be embedded into evidence statements across achievement levels in a cogent way and to interact with content. In this way, the features of cognitive processing, content difficulty, and context interact to affect item difficulty.

A principled approach to test design is intended to support the validity of inferences about the student's stage of learning and the content validity of the assessment as a measure of student achievement. Under such a score interpretation model, construction of test blueprints should eventually not treat DOK as a separate blueprint constraint. Instead, DOK should be present as evidence embedded in to a descriptor for an achievement level that supports interpretations regarding the stage of thinking sophistication the student is at during the time of the test event. The items that are found within each achievement level should match the ALDs.

1.5.1. Principled Test Design Based on Range ALDs

The purposes of a test design centered in ALDs include the following:

- To show how students increase in their reasoning with specific content across achievement levels to support collecting purposeful evidence of what mastery of college and career readiness means
- To support teachers in making more accurate inferences about what students know and can do based on the student's present level of performance at year end

When test developers use a principled approach to test design, ALDs may be viewed as the score interpretation. That is, the ALDs become the construct interpretive argument described by Kane (2013). The degree of alignment of items to the assessment, a component of the evidence gathered to support a validity framework, should not simply focus on the alignment to the content and DOK of the standard. Instead, it should focus on the degree of concurrence in the DOK and content alignment of items within an achievement level to the associated ALDs. Test developers must create assessments that provide guidance under such a framework (Perie & Huff, 2016) to support educators having sufficient information to personalize instruction centered in where students are in their learning by the end of the year.

ALDs are intended to be the linchpin of the NSCAS interpretation and use argument. As such, the NDE developed ALDs for the NSCAS Summative ELA and Mathematics assessments 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

By developing ALDs this way, Nebraska communicated how standards are interpreted for assessment purposes, how tasks can align to a standard but not be of sufficient difficulty and depth to represent mastery, and what growth on the test score continuum represents.

1.5.2. Achievement Level Descriptors (ALDs)

Policy ALDs are high-level expectations of student achievement within each achievement level across grades. Range ALDs are within-standard learning progressions that describe the knowledge and skills students at each achievement level should be able to demonstrate. They describe the current stage of learning within the standard and explicate observable evidence of achievement, demonstrating how skills change and become more sophisticated across achievement levels for each standard. Reporting ALDs are finalized versions of the Range ALDs supported by evidence from the test scale that were created after the final cut scores were adopted.

1.5.2.1. Policy ALDs

The Nebraska ALDs guide the establishment of the intended policy outcomes the 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.

1.5.2.2. Range ALDs

For each expectation and indicator in the standards, Range ALDs should explicate observable evidence of achievement, demonstrating how the skill changes and becomes more sophisticated across achievement levels. For ALDs to be the foundation of test score interpretation, they should reflect more complex KSAs as the achievement levels increase (Schneider, Huff, Egan, Gaines, & Ferrara, 2012). For example, more complex KSAs should be expected for On Track than for the CCR Benchmark achievement level.

Under Nebraska's Range ALDs approach, the state defined intended content-based interpretations of what scale scores within an achievement level represent. Figure 1.2 presents the balanced practical approach the NDE took by necessity in the development process of the NSCAS Summative assessments. Content interpretations were drafted but not finalized until after the standard setting and, as the highlighted blue arrow shows, will be used to support item specifications moving forward to ensure a stable, comparable construct over time.

Figure 1.2. Principled Test Design Process to Support Test Score Interpretations and Uses

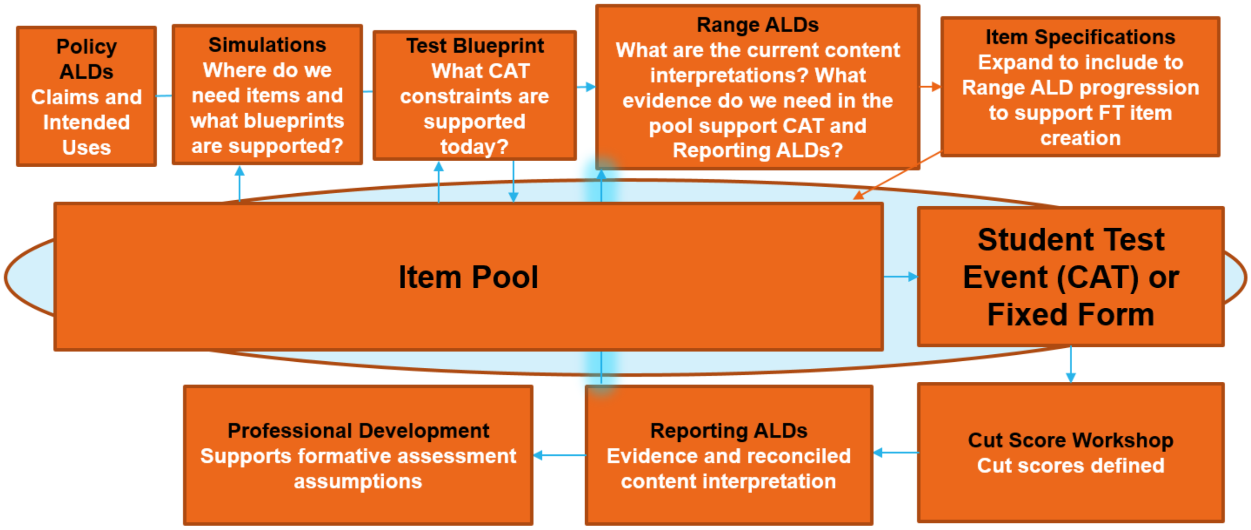

Policy ALDs Claims and Intended Uses
Simulations Where do we need items and what blueprints are supported?
Test Blueprint What CAT constraints are supported today?
Range ALDs What are the current content interpretations? What evidence do we need in the pool support CAT and Reporting ALDs?
Item Specifications Expand to include to Range ALD progression to support FT item creation
Item Pool
Student Test Event (CAT) or Fixed Form
Professional Development Supports formative assessment assumptions
Reporting ALDs Evidence and reconciled content interpretation
Cut Score Workshop Cut scores defined

Table 1.2 and Table 1.3 show how content interpretation is defined in the Range ALDs using Grade 3 as an example for both ELA and Mathematics, respectively. 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., MA 3.1.1.c) can be ranges of content- and thinking-skill difficulty that describe different stages of reasoning.

Table 1.2. Example of How Content Interpretation is Defined in Range ALDs—ELA Grade 3

Indicator	Developing	On Track	CCR Benchmark
	With a range of texts with text complexity commonly found in Grade 3, a student performing in Developing can likely	With a range of texts with text complexity commonly found in Grade 3, a student performing in On Track can likely	With a range of texts with text complexity commonly found at the intersection of Grade 3 and Grade 4, a student performing in CCR Benchmark can likely
Reading Vocabulary			
LA 3.1 **Reading:** Students will learn and apply reading skills and strategies to comprehend text.			
LA 3.1.5 **Vocabulary:** Students will build and use conversational, academic, and content-specific grade-level vocabulary			
LA 3.1.5.a Determine meaning of words through the knowledge of word structure elements, known words, and word patterns (e.g., contractions, plurals, possessives, parts of speech, syllables, affixes, base and root words, abbreviations).	Identify basic word structure elements and word patterns to determine meaning of words (e.g., plurals, parts of speech, syllables).	Apply knowledge of word structure elements, known words and word patterns to determine meaning of words (e.g., contractions, plurals, possessives, parts of speech, syllables, affixes, base and root words, abbreviations).	Analyze complex word structure elements, known words and word patterns to determine meaning of words (e.g., contractions, plurals, possessives, parts of speech, syllables, affixes, base and root words, abbreviations).

Table 1.3. Example of How Content Interpretation is Defined in Range ALDs—Mathematics Grade 3

Indicator	Developing	On Track	CCR Benchmark
	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. A developing learner…	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. An on-track learner…	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. A college-and-career ready learner…
MA 3.1 NUMBER: Students will communicate number sense concepts using multiple representations to reason, solve problems, and make connections within mathematics and across disciplines.			
MA 3.1.1.c Round a whole number to the tens or hundreds place, using place value understanding or a visual representation.	Rounds a two-digit or three-digit whole number to the tens or hundreds place with or without a visual model.	Rounds a whole number from 1,000 up to 100,000 to the tens or hundreds place given a visual model.	Uses place value understanding to round a whole number from 1,000 up to 100,000 to the tens or hundreds place without a visual model. Analyzes the rounding of a whole number up to 100,000 to the tens or hundreds place using place value understanding or a visual representation ((e.g., explain why 5,610 rounds to 6,000 when rounded to the nearest thousand).

1.5.3. Range ALD Construction Framework

The Nebraska standards are organized so that each expectation level represents a specific skill or building block for problem solving. For example, in the Grade 7 Algebra standards, creating an inequality from words is part of Algebraic Relationships, solving one-step inequalities is part of Algebraic Processes, and solving real-world problems with inequalities is part of Applications. 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. This is considered context or the conditions under which a student can show that he or she knows. 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). Finally, 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, Schneider, & Ferrara, 2012). While this approach is helpful in a fixed-form context, it is critical to item development for a computer adaptive assessment.

1.5.4. Alignment and Computer Adaptive Testing (CAT)

Within an adaptive testing context, the documentation of content blueprint features and percentages present in the item pool become one evaluation tool used to frame alignment discussions. Both item pool structure and constraints used to establish the administration of items during test events support the definition of the construct for alignment purposes. Full test blueprints must be supportable for students in each achievement level. Therefore, an ideal item pool has similar percentages of items within each indicator by achievement level cell. Thus, from a representation of the content perspective, standard alignment methodologies such as the Webb Alignment Tool are appropriate tools for assessing the item pool. However, from a student test event perspective, these approaches do not support helping to target items to optimally measure where the student is centered within the grade-level standards.

As Range ALDs were developed based on theories of how student thinking grows within the state's structure of state standards, and the evidence needed to support that conclusion, the characteristics of items depend on the student's stage of reasoning. As ALDs describe increases in student thinking and reasoning, test developers have a rationale regarding why a percentage of particular item types (e.g., technology-enhanced items) and 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, Schneider, & Ferrara, 2012) is that DOK levels follow the indicator progression. While DOK may increase across achievement levels, the DOK level should not automatically increase with the achievement level increase. What may be required from a learning theory perspective is that students have support accessing the standards, such as with visual supports demarcating a manipulation of an item context feature. They then may access the standards without the visual aids, followed by accessing the standards at a higher DOK level. Thus, if the item development is purposeful to the progression, DOK specifications are not required as a constraint conditional that items are measuring what the ALDs say they are.

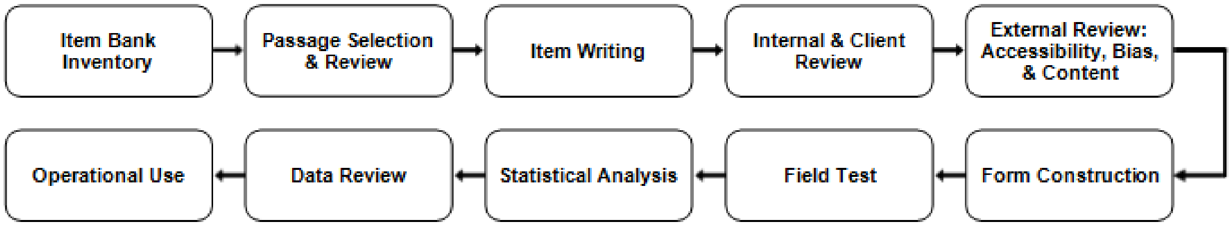
When item development is purposeful to a clearly defined construct, dictating a certain percentage of items at a particular DOK level will have an unintended consequence of routing a student to items that provide less information about their current stage of thinking and reasoning with the content. Thus,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

Section 2: Test Design and Development

This section describes the test designs for the 2018 NSCAS ELA, Mathematics, and Science assessments and the test development process for ELA and Mathematics. As mutually agreed between the NDE and NWEA, the Science content development was deferred until Summer 2018 to accommodate creation of a new three-dimensional Science test for Nebraska. In the interim, existing items were used for the Grades 5 and 8 Science assessments. A description of the Science item development process—thus, the validity evidence based on test content for Science—is included in previous NeSA technical reports (e.g., DRC, 2017).

As Nebraska transitioned to a CAT administration for ELA and Mathematics, the need to build a large, robust item bank was a key requirement, and the development of new scales had to be accomplished concurrently with thinking about the development of ALDs. To support building of a bank to sufficiently support a CAT, NWEA began passage and item development in Summer 2017 to have enough content available to populate field test slots in Spring 2018. Items were written by educators in an item writer workshop (IWW) and by independent contractors. Once initial item development was completed, all items were taken to content and bias review meetings with Nebraska educators from September 19–21, 2017. Items that survived these meetings were considered for the field test pool in Spring 2018. Figure 2.1 outlines the steps taken during the test development process in 2017–2018.

Figure 2.1. Test Development Process



2.1. Test Designs

For all three content areas, test designs had mostly been completed prior to the start of this contract. The exception was moving from a fixed-form design to a CAT in ELA and Mathematics, which was a process NWEA worked iteratively on with the NDE to agree to the modification of the Table of Specifications (TOS) that acted as both blueprint and test design. Table 2.1 summarizes the test designs for the 2018 NSCAS Summative assessments. Table 2.2 presents the number of items and points possible on each online and paper-pencil test form by content area and grade.

As shown in Table 2.1, NWEA developed a total of three online versions of the assessments that met the TOS and two paper-pencil versions. The purpose of the paper-pencils forms was to serve as accommodated versions. The online contingency version was produced as a contingency plan to address possible technology challenges with the CAT administrations because this was the first high-stakes use of the adaptive engine. The paper-pencil forms contained only operational items and were slightly longer than the CAT to support comparable levels of test score precision. Students who completed the CAT had the same number of total items as the paper forms, but the item roles differed.

Students administered the CAT took a total of 48 items (41 operational + 7 non-operational). Twenty of the operational items were selected adaptively based on student ability level, and 21 were non-adaptively pre-selected horizontal and vertical linking anchors. Thus, the test design is best classified as a multi-staged assessment in which students first received the fixed anchor set that acted as a locater with which to begin adaptive selection for the second portion of the test. Each student also saw one set of seven vertical linking or field test items. All Science items were multiple choice, so the points possible was a fixed number. Science had one form.

Table 2.1. 2018 NSCAS Summative Test Designs

Content Area	Grade(s)	Test Designs*					
		Online	PP	Spanish Online	Spanish PP	Breach	Online Contingency
ELA	3–8	Adaptive (48 total per grade, 41 OP + 7 FT/VL)	One form per grade (48 OP)	Fixed (translation of PP form)	Same form as Spanish online	N/A	2018 PP forms plus 7 VL/FT (48+7=55 items)
Mathematics	3–8	Adaptive (48 total per grade, 41 OP + 7 FT/VL)	One form per grade (48 OP)	Fixed (translation of PP form)			
Science	5	Fixed (same form as PP)	One form (50 OP)	Fixed (translation of online form)			N/A
	8	Fixed (same form as PP)	One form (60 OP)	Fixed (translation of online form)			N/A

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

Table 2.2. Number of Items and Points Per Test

		Online							
		Operational		FT/VL*		Total		Paper-Pencil	
Content Area	Grade	#Items	#Points	#Items	#Points	#Items	#Points	#Items	#Points
ELA	3	41	50–51	7	7–9	48	57–60	48	56
	4	41	48–49	7	7–9	48	55–58	48	57
	5	41	51	7	7–9	48	58–60	48	58
	6	41	53–54	7	7–9	48	60–63	48	56
	7	41	49	7	7–9	48	56–58	48	55
	8	41	52–53	7	7–9	48	59–62	48	59
Mathematics	3	41	45	7	7–9	48	52–54	48	49
	4	41	45	7	7–9	48	52–54	48	48
	5	41	45	7	7–9	48	52–54	48	48
	6	41	45	7	7–9	48	52–54	48	52
	7	41	45	7	7–9	48	52–54	48	52
	8	41	45	7	7–9	48	52–54	48	50
Science	5	50	50	–	–	50	50	50	50
	8	60	60	–	–	60	60	60	60

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

2.2. Content Standards

Nebraska Revised Statute 79-760.01[2] requires the Nebraska State Board of Education to "adopt measurable academic content standards for at least the grade levels required for statewide assessment. Those standards shall cover the subject areas of reading, writing, mathematics, science, and social studies…the State Board of Education shall develop a plan to review and update standards for those subject areas every seven years." The revised statute was effective as of August 30, 2015.[3]

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 Nebraska's College and Career Ready Standards for Science. However, these will not be implemented in the NSCAS-Summative Assessments until 2018–2019. The 2017–2018 NSCAS Science assessments continued to be aligned to the 2010 Science standards.

2.3. Content Alignment Philosophy

To fully represent the constructs being assessed by NSCAS to determine if students are ready for college and careers, solid content alignment was critical. This was covered in several ways, including adherence to specifications, common interpretations of the standards, and an agreed-upon approach for cognitive complexity across all item types. At the start of the contract, NWEA staff engaged with NDE staff to agree upon and fully understand the existing content alignment philosophy and to expand the content alignment philosophy as appropriate.

2.4. Table of Specifications (TOS)

The 2017–2018 NSCAS Summative blueprints are embedded in the NSCAS Table of Specifications (TOS) that indicate the range of test items included for each standards indicator in each content area. The adaptive test was constrained to make sure each student received items within the identified ranges. The 2017–2018 test fixed-forms and adaptive forms were not an exact match to the TOS given the attributes of available items in the item bank. Future forms will adhere more closely to the TOS as more items are available. Appendix A presents the TOS for each content area. The TOS for Science is different from ELA and Mathematics in that the total number of items is provided at the grade-level standard rather than at the indicator level. This decision was made based on input received from Science content experts from across the state. All indicators under a tested grade-level standard may be present on the Science test.

2.5. Reporting Categories

The reporting categories, shown in Table 2.3, were used for scoring and reporting. Items were mapped to a reporting category based on the indicators.

Table 2.3. Reporting Categories

Content Area	Reporting Categories
ELA	• Reading Vocabulary • Reading Comprehension • Writing Skills

[2] https://nebraskalegislature.gov/laws/statutes.php?statute=79-760.01
[3] https://www.education.ne.gov/contentareastandards/

Content Area	Reporting Categories
Mathematics	• Number • Algebra • Geometry • Data
Science	• Inquiry, Nature of Science, & Tech • Physical Science • Life Science • Earth/Space Sciences

2.6. Depth of Knowledge (DOK)

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

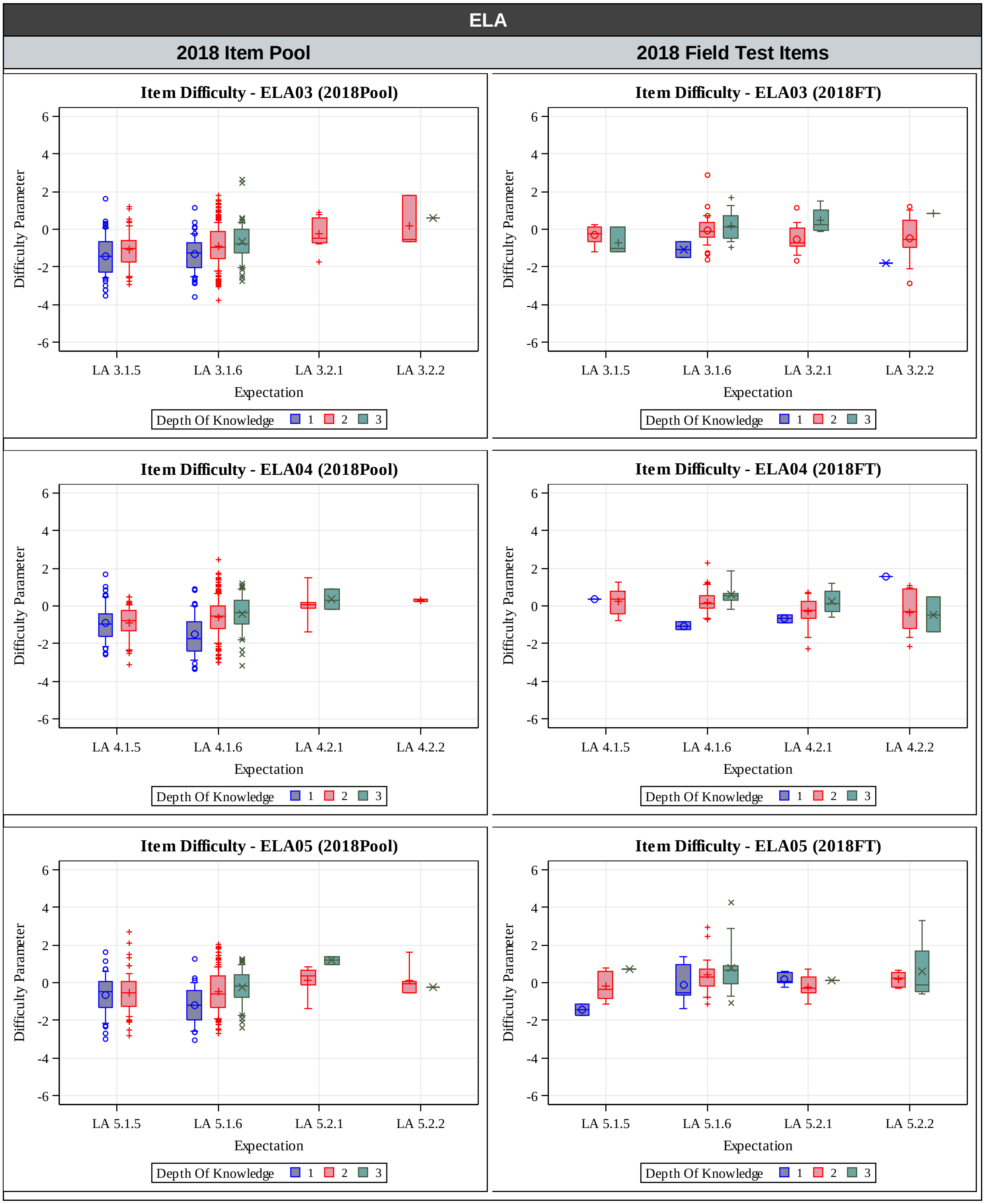
- DOK 1: Recall & Reproduction
- DOK 2: Skill & Concepts
- DOK 3: Strategic Thinking

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

Figure 2.2 and Figure 2.3 present boxplots of item DOK levels based on the state’s interpretation of DOK for the 2018 item pool and 2018 field test items. ELA items were largely successfully developed to match intended content and cognitive complexity, including higher-order thinking as shown by the trends of items increasing in difficulty on average for the same indictor as they increase in DOK with the range of item difficulties not being unreasonably restricted depending on the level of cognitive complexity.

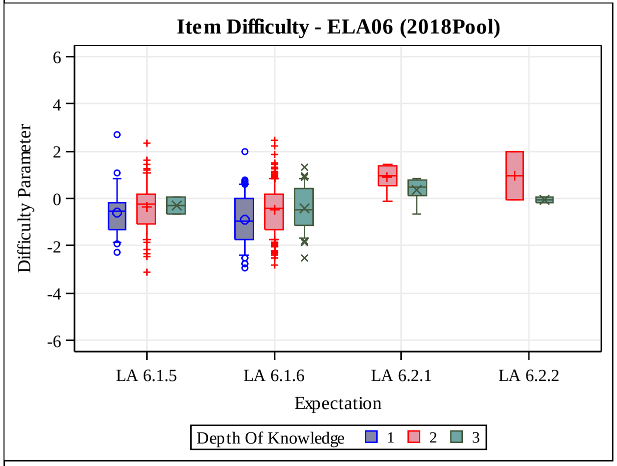
A different trend is seen Mathematics. The state considers items that measure procedural knowledge in isolation as DOK 1 and items that measure procedural knowledge in a practical real-world context as an increase in depth of knowledge (DOK 2). The data trends for these DOK levels are largely similar item difficulties not being unreasonably restricted for DOK 1 and DOK 2 items with opportunities to develop DOK 3 items in standards in the future based on the state policy decision in February 2018. As the Range ALDs will be used in the future to elicit item content, it is expected that trend data based on a priori ALD level classifications will produce expected trends.

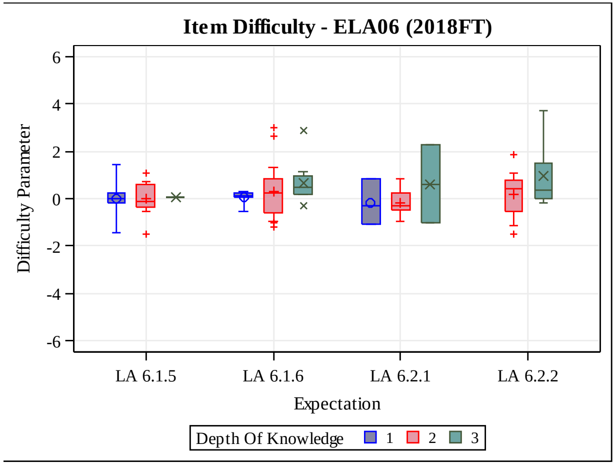
Figure 2.2. DOK Box Plots for 2018 Item Pool and Field Test Items—ELA Grades

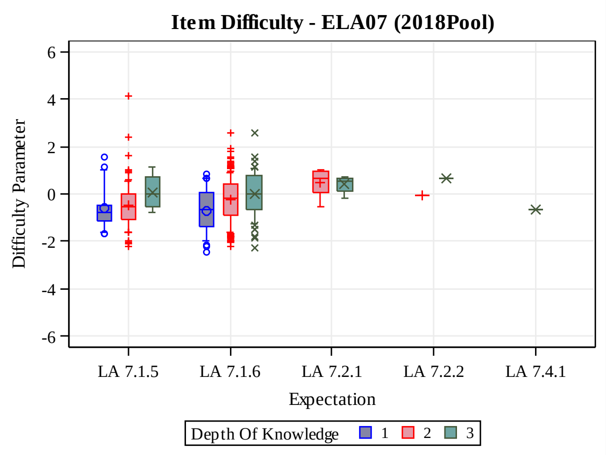


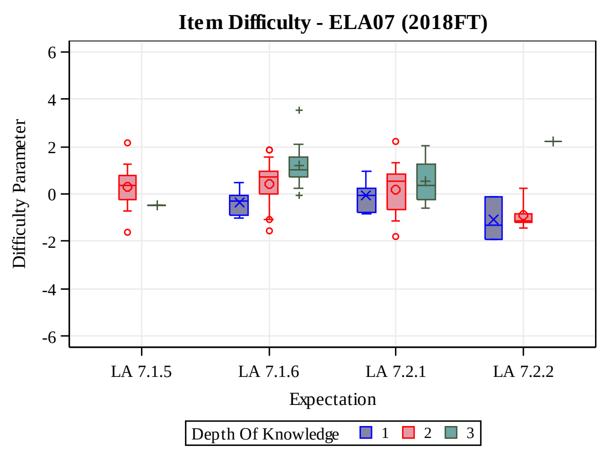
ELA

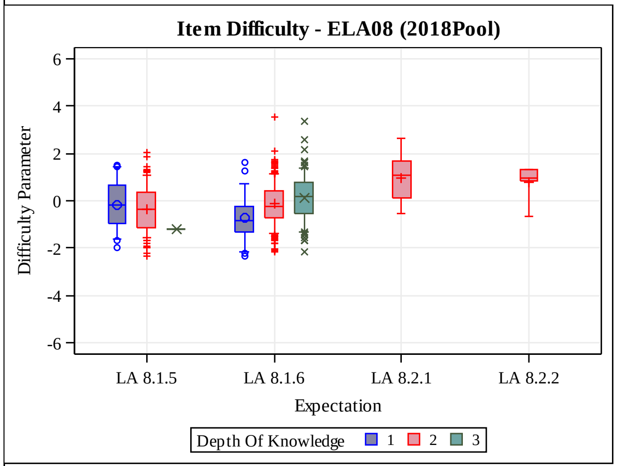
2018 Item Pool	2018 Field Test Items


Item Difficulty - ELA06 (2018Pool)


Item Difficulty - ELA06 (2018FT)


Item Difficulty - ELA07 (2018Pool)


Item Difficulty - ELA07 (2018FT)


Item Difficulty - ELA08 (2018Pool)

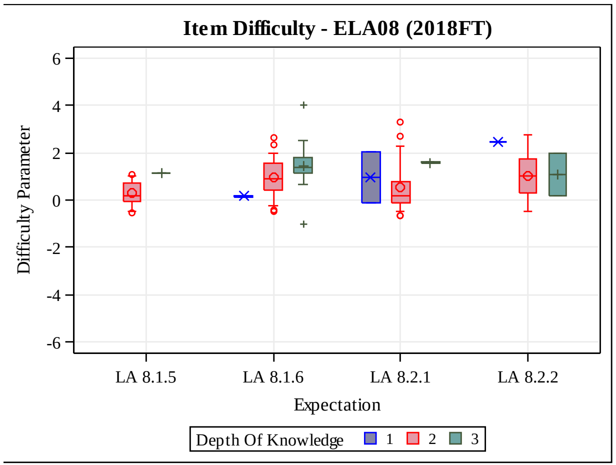
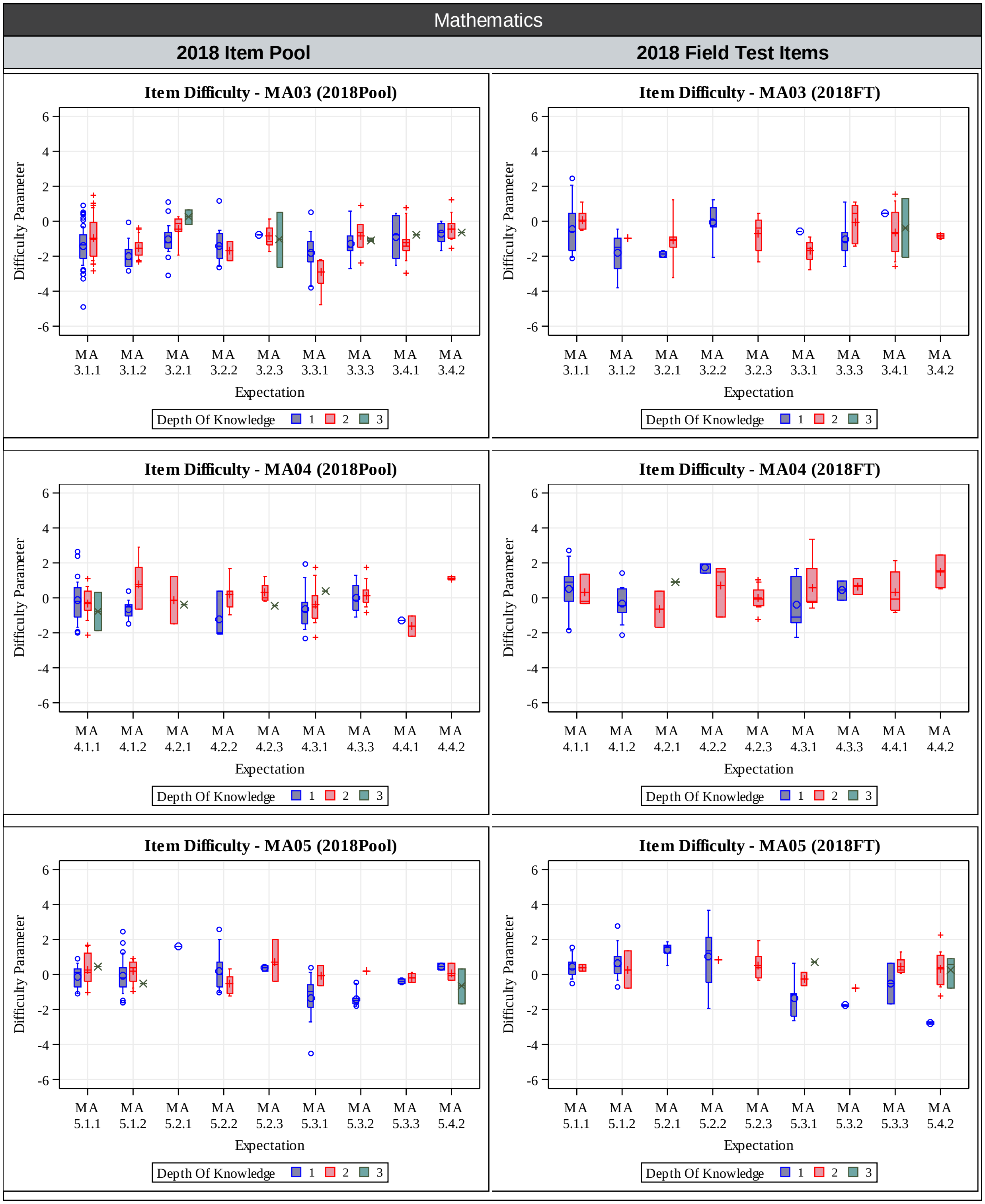
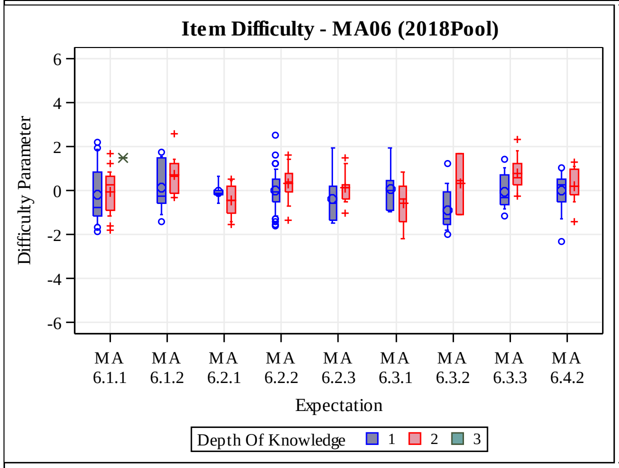

Item Difficulty - ELA08 (2018FT)

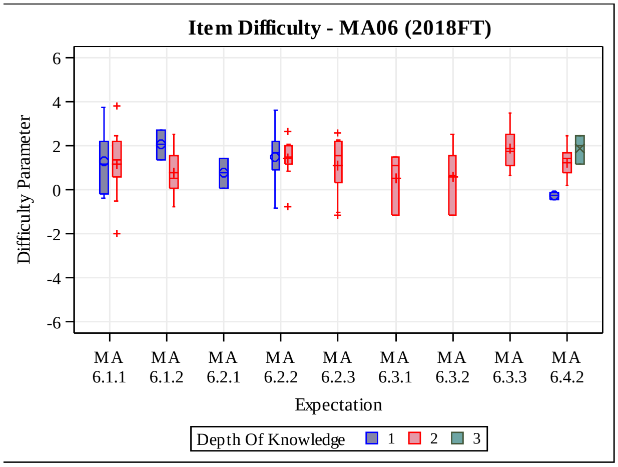
Figure 2.3. DOK Box Plots for 2018 Item Pool and Field Test Items—Mathematics

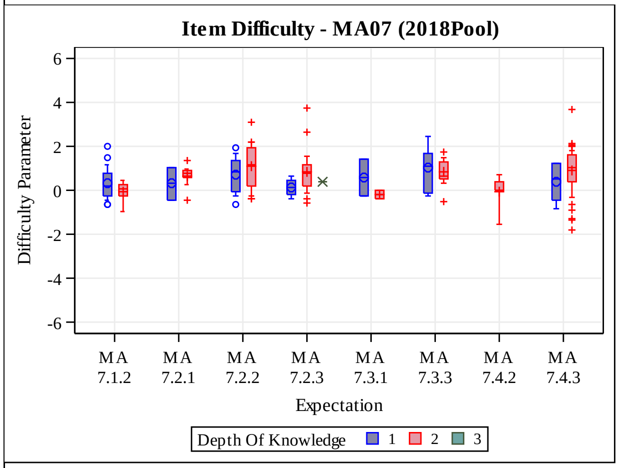


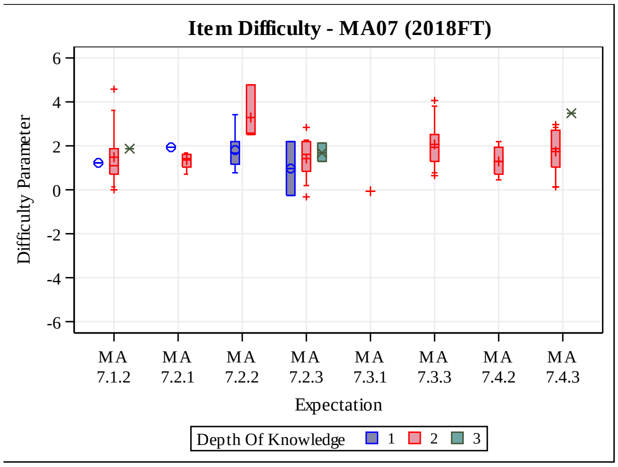
Mathematics

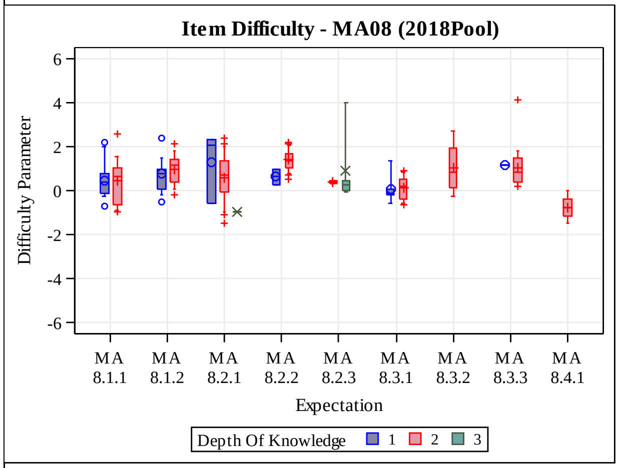
2018 Item Pool	2018 Field Test Items

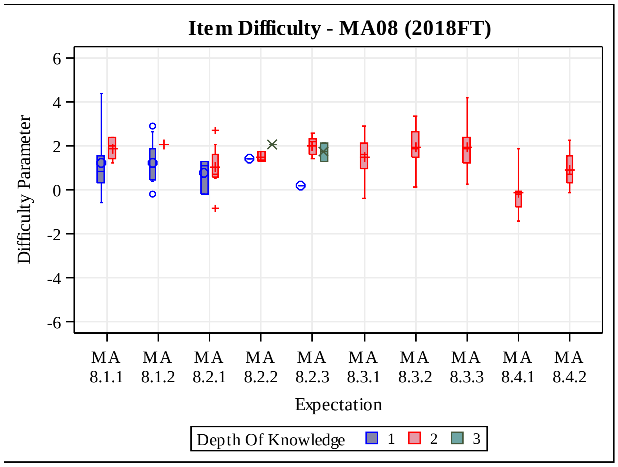

Item Difficulty - MA06 (2018Pool)


Item Difficulty - MA06 (2018FT)


Item Difficulty - MA07 (2018Pool)


Item Difficulty - MA07 (2018FT)


Item Difficulty - MA08 (2018Pool)


Item Difficulty - MA08 (2018FT)

2.7. Item Types

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

Table 2.4. Item Types for Online ELA and Mathematics

Item Type	Description
Multiple-Choice (Choice)	Students select one response from multiple options.
Multiselect (Choice Multiple)	Students select two or more responses from multiple options.
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.
Gap Match	Students select one or more answer options from the item toolbox and populate a defined area, or "gap."
Graphic Gap Match	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.
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.
Click & Pop	Students move options (e.g., words, phrases, symbols, numbers, or graphic elements) from the area called the toolbar to designated container(s) on screen by selecting an option; the option then “pops” into the container on screen.

2.8. Educator Involvement

The NDE included educators throughout the entire item development process to produce customized items and provide educators with an invaluable professional development opportunity. Educators also participated in Range ALD development meetings, IWWs, and item review meetings. They also participated in standard setting and cut score review after the first operational summative assessment results of the new assessment program were collected and analyzed.

2.9. ELA Passage Development

Table 2.5 and Table 2.6 provide the number of passages developed for the NSCAS Summative ELA assessments by NWEA and its subcontractor Data Recognition Corporation (DRC). As shown in the tables, 192 passages were either commissioned or acquired through the public domain. The sourcing of the passages was 67% commissioned and 33% from the public domain. DRC provided 10–12 passages per grade to NWEA. DRC completed the first rounds of editorial review for the items they provided, and NWEA completed the first rounds of editorial review for its own items. All passages were then reviewed during educator review meetings.

Table 2.5. Number of Passages by Passage Type: Literary vs. Informational

	Literary			Informational				
Grade	NWEA	DRC	Total	NWEA	DRC	Total	Opinion	Grand Total
3	10	6	16	10	6	16	–	**32**
4	11	6	17	9	6	15	–	**32**
5	10	7	17	10	5	15	–	**32**
6	10	6	16	10	6	16	–	**32**
7	9	5	14	11	7	18	–	**32**
8	8	5	13	12	5	17	2	**32**
Total	**58**	**35**	**93**	**62**	**35**	**97**	**2**	**192**

Table 2.6. Number of Passages by Passage Source: Commissioned vs. Public Domain

	Commissioned			Public Domain			
Grade	NWEA	DRC	Total	NWEA	DRC	Total	Grand Total
3	13	12	25	7	–	7	**32**
4	11	12	23	9	–	9	**32**
5	11	9	20	9	3	12	**32**
6	9	9	18	11	3	14	**32**
7	9	12	21	11	–	11	**32**
8	6	10	16	14	2	16	**32**
Total	**59**	**64**	**123**	**61**	**8**	**69**	**192**

2.9.1. Passage Specifications

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

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

2.9.2. Readability Measures

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

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

The quantitative measures of a passage were also considered as a factor for all passages. Lexiles where used as the readability measure for this content development work. For pieces of text such as poems that perform poorly when Lexiles are run, Flesch-Kincaid was run as a secondary measure. Table 2.7 presents the acceptable Lexile ranges for each grade, as well as the total word count per passage. The passages selected for a grade spanned a range of acceptable readabilities. The word count must be reasonable for the task and, within the acceptable word count ranges, provide enough richness to support robust item sets.

Table 2.7. Lexile and Word Count Ranges

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

2.10. Item Development

Item development for 2017–2018 occurred for ELA and Mathematics. An online item writing workshop generated 60% of the development for this cycle. Independent contractors were then used to offset gaps in the item bank (i.e., about 40% of development) to ensure that enough items were developed to fulfill the item development requirements. As mutually agreed between the NDE and NWEA, the Science content development was deferred until Summer 2018 to accommodate creation of a new three-dimensional Science test for Nebraska. In the interim, existing items were used for the NSCAS Summative Science Grades 5 and 8 assessments. A complete item bank analysis was not possible for 2017–2018 since the importing of items and metadata was not complete by the time the item development plan was assembled.

Item specifications were created for both ELA and Mathematics since new items were being developed. The documents capture aspects such as the following and will be reviewed at the start of each new development cycle to ensure accuracy.

- General item writing guidelines
- Specific guidelines for using TEIs
- Specific standard information for approaching Grades 3–8

2.10.1. Development Targets

Table 2.8 presents the passage and item development targets based on NWEA’s response to the NDE’s request for proposal (RFP). As shown in the table, the item development plan included the development of 180 passages and 2,160 items across both content areas.

Table 2.8. 2017–2018 Overall Development Targets

		#Items		
Grade	**#Passages**	**MC**	**TEI***	**Total**
ELA				
3	30	120	90	210
4	30	120	90	210
5	30	120	90	210
6	30	120	90	210
7	30	120	90	210
8	30	120	90	210
Mathematics				
3	–	60	90	150
4	–	60	90	150
5	–	60	90	150
6	–	60	90	150
7	–	60	90	150
8	–	60	90	150
Total	**180**	**1,080**	**1,080**	**2,160**

*TEIs are any item type that is not an MC item and can be worth 1 or 2 points.

2.10.1.1. ELA Development Targets

The ELA item bank had a notable shortage of writing items, most likely influenced by the Text Depend Analysis items that will not be used in this contract due to human handscoring requirements. Therefore, NWEA focused heavily on these items, which are not passage-dependent. Table 2.9 presents the ELA item development targets by MC and TEI item types.

Table 2.9. Item Development Targets—ELA

	#ELA Items Developed 2017-18								
	Reading			**Writing**			**Overall**		
Grade	**MC**	**TEI**	**Total**	**MC**	**TEI**	**Total**	**MC**	**TEI**	**Total**
3	90	58	148	30	32	62	120	90	210
4	90	58	148	30	32	62	120	90	210
5	90	58	148	30	32	62	120	90	210
6	84	65	149	36	25	61	120	90	210
7	84	65	149	36	25	61	120	90	210
8	84	65	149	36	25	61	120	90	210
Total	**522**	**369**	**891**	**198**	**171**	**369**	**720**	**540**	**1,260**

2.10.1.2. Mathematics Development Targets

The Mathematics item bank realignment from Nebraska's Academic Standards to the College and Career Readiness Standards had not been completed prior to hand off to NWEA. As a result, the item development plan relied on expertise of the Mathematics content developer rather than an item bank analysis. Table 2.10 presents the Mathematics item development targets by MC and TEI item types.

Table 2.10. Item Development Targets—Mathematics

	#Items				
		TEI			
Grade	MC	1-pt.	2-pt.	Total	Grand Total
3	60	52	38	90	**150**
4	60	45	45	90	**150**
5	60	45	45	90	**150**
6	60	46	44	90	**150**
7	60	45	45	90	**150**
8	60	48	42	90	**150**
Total	**360**	**281**	**259**	**540**	**900**

2.10.2. Item Writer Workshop (IWW)

The online IWW from August 29–31, 2017, provided a professional development opportunity to educators and allowed them to be a part of the item development process. Table 2.11 presents the number of participants in each panel who were recruited and selected by the NDE. The expertise of Nebraska teachers was critical to the item writing process. Nebraska educators wrote test items that were featured on the assessments. This ensured content that seems familiar to students as they take the tests; they will not see unfamiliar wording or approaches that might negatively impact performance.

Table 2.11. Item Writer Workshop (IWW) Panel Composition

Panel	#Panelists
ELA 3–4	18
ELA 5–6	19
ELA 7–8	20
Math 3–4	11
Math 5–6	13
Math 7–8	15
Total	**96**

During the IWW, educators were convened in-person where they were trained on how to write high-quality items aligned to the state standards. Participants in item writing met as a large group for training on the systems needed to enter items as well as an orientation to the standards and assignments. In this training, delivered collaboratively by the NDE and NWEA, participants learned to write items that met the following criteria:

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

A general session was held the first morning of the IWW to train educators on the basics of item writing and usage of NWEA technology. Participants also reviewed the standards and the assessment and practiced item writing. A second, subject-specific training was completed with each group to dive into ELA and Mathematics issues. Once trained in both general and content specific information, participants chose a standard and an item type to complete each assignment in the item management system. This process was repeated until all assignments were completed to meet the IWW targets. Throughout this process, educators partnered and shared their expertise as they wrote multiple-choice items and TEIs.

NWEA and NDE staff circulated in break-out rooms to answer questions and provide guidance to participants. After the initial draft of an item was submitted, the participants, NDE staff, and NWEA staff collaborated and engaged in brief group editing sessions that encouraged discussion and the continuing development of item-writing skills.

2.10.3. Content and Bias Review

All newly developed items underwent a rigorous internal review. All items survived internal review of content and bias/fairness. The items were then reviewed by educators during external item content and bias reviews that provided an opportunity to engage the expertise of Nebraska educators. During a three-day meeting from September 19–21, 2017, Nebraska educators gathered together for two concurrent meetings, one to review items for content validity and one to review items for any possible sources of bias and sensitivity issues. While Nebraska educators served as the originators of a significant percentage of items, educator involvement in item reviews provided another opportunity to make sure that the material was appropriate for Nebraska's assessments and to provide a valuable professional development opportunity for participants.

Stakeholders participating in the content and bias reviews received training, delivered collaboratively by the NDE and NWEA, at the beginning of each review session. Participant were provided checklists to refer to during the reviews. Participants in item content review learned to review items for qualities including, but not limited to, the following:

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

Participants in item bias review learned to review items for qualities including, but not limited to, the following:

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

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

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

Table 2.12 presents the panel compositions for each 2017 item review meeting.

Table 2.12. Item Review Meeting Panel Composition

Item Review Meeting	Panel	#Panelists
Bias Review, Sept. 19–21, 2017	ELA 3–4	4
	ELA 5–6	5
	ELA 7–8	5
	Math 3–4	4
	Math 4–5	5
	Math 7–8	5
	Total	**28**
Content Review, Sept. 19–21, 2017	ELA 3	5
	ELA 4	5
	ELA 5	5
	ELA 6	5
	ELA 7	5
	ELA 8	4
	Math 3	5
	Math 4	5
	Math 5	5
	Math 6	5
	Math 7	5
	Math 8	5
	Total	**59**
Grand Total		**87**

2.10.4. Item Development Results

Table 2.13 and Table 2.14 present the number of items taken to the external item and bias reviews. Appendix B presents the number of items by standard taken to committee for both ELA and Mathematics. Table 2.15 then provides the difference between the item development targets and the actual number of items that were fully developed. The difference was added to the Summer 2018 item development targets.

Table 2.13. Number of Items Taken to Committee Review—ELA

Grade	#Items Taken to Committee
3	187
4	169
5	153
6	169
7	180
8	192
Total	**1,050**

Table 2.14. Number of Items Taken to Committee Review—Mathematics

	#Items Taken to Committee by Item Type				
		TEI			
Grade	**MC**	**1-pt.**	**2-pt.**	**Total**	**Grand Total**
3	40	36	24	60	**100**
4	40	30	30	60	**100**
5	40	30	30	60	**100**
6	40	30	30	60	**100**
7	40	30	30	60	**100**
8	40	34	26	60	**100**
Total	**240**	**190**	**170**	**360**	**600**

Table 2.15. Item Development Targets vs. Number of Items Developed

Grade	#Items Proposed	#Items Completed	Difference to be Added to the Summer 2018 Development
ELA			
3	210	187	23
4	210	171	39
5	210	153	57
6	210	169	41
7	210	180	30
8	210	192	18
Mathematics			
3	150	100	50
4	150	100	50
5	150	100	50
6	150	100	50
7	150	100	50
8	150	100	50

Table 2.16 presents the number of items accepted, modified, or rejected results at the external content and bias review meeting. For ELA, 93.4% of items were either accepted or accepted with modifications, and 6.6% of items were rejected. For Mathematics, 96.3% of items were either accepted or accepted with modifications, and 3.7% of items were rejected.

Table 2.16. External Item Review Results

	#Items			
Grade	**Accepted**	**Modified**	**Rejected**	**Total**
ELA				
3	114	70	3	187
4	104	64	1	169
5	103	47	3	153
6	74	84	11	169
7	64	93	23	180
8	53	111	28	192
Total	**512**	**469**	**69**	**1,050**

Grade	#Items			
	Accepted	Modified	Rejected	Total
Mathematics				
3	11	88	1	100
4	23	74	3	100
5	10	90	0	100
6	25	72	3	100
7	43	50	7	100
8	11	81	8	100
Total	**123**	**455**	**22**	**600**

2.11. 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.

2.12. 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.

2.13. Test Construction

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

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

The content team selected the items based on the TOS and specifications for each grade and content area, including the following:

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

Item selection was an iterative process between the psychometrics and content teams before being sent to the NDE for review and approval. The 2017 Science forms were reused for the 2018 Science forms.

2.14. Data Review

Data review is the process of reviewing field tested items for quality and appropriateness based on the results of statistical analysis of student responses. The review of content alignment and statistics of the Spring 2018 NSCAS Summative ELA and Mathematics field tested items occurred virtually in September 2018 between the NDE and NWEA. Table 2.17 and Table 2.18 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] although items with a negative item-total correlation or polytomous items without a second step parameter were marked Do Not Use (DNU) and not included. Participants were provided a spreadsheet with the statistics for each item, as well as a data review "cheat sheet" provided in Appendix C. Only flagged items were brought to the data review meeting.

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

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

Table 2.18. Data Review Flagging Criteria—Non-Multiple-Choice Items

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

Table 2.19 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 2018 testing window, the number of field test items not accepted or labeled as DNU, and the number of accepted field test items.

[4] The details of field testing item analyses are included in Section 5 of this technical report.

Table 2.19. Data Review Results

Grade	#FT Items in the Pool	#FT Items Administered in 2018	#Rejected/DNU Items	#Accepted Items
ELA				
3	111	111	13	98
4	102	102	12	90
5	105	105	18	87
6	105	97	8	89
7	102	102	8	94
8	102	102	10	92
Mathematics				
3	97	87	3	84
4	96	94	6	88
5	96	95	3	92
6	96	93	4	89
7	94	91	8	83
8	88	86	3	83

Section 3: Test Administration and Security

The Spring 2018 NSCAS Summative testing window was from March 19 to April 27, 2018. The make-up testing window was from April 30 to May 4, 2018. The tests were administered online via NWEA's Comprehensive Assessment Platform (CAP) test management system with paper-pencil versions available as an accommodation. Each district was required to return either a paper-pencil answer sheet or online record for the 2018 NSCAS ELA and Mathematics tests for all Grade 3–8 students enrolled in the district and for the 2018 NSCAS Science test in Grades 5 and 8.

The ELA and Mathematics tests each had 48 items and were adaptive. The Science tests each had 50–60 items and were fixed form. All tests were untimed. Testing sessions were structured as a single session; however, students could complete the tests in more than one sitting by pausing the test. Students were not able to go back to previous items.

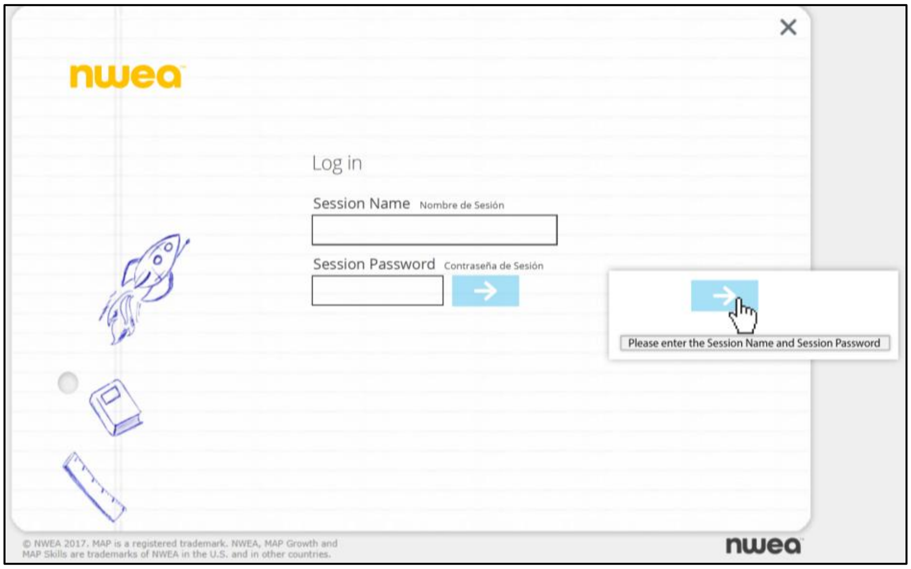
The NSCAS Summative administration supported student testing on Windows, Macintosh, iPads, and Chromebooks that met the following specifications. Chromebooks were only supported if the student was using an external keyboard.

- Windows 7, 8.1, or 10
- Mac OS® X v10.11 to 10.14
- iPad iOS 10 to 12
- Google Chrome™ OS 65 or higher

3.1. Comprehensive Assessment Platform (CAP)

The NWEA CAP is the web browser-based platform for administering assessments and viewing reports for MAP® Growth™ and the NSCAS Summative assessment. This roles-based platform used during the Spring 2018 NSCAS test administration allowed users roster students, set up test sessions, and test students. The CAP works with NWEA's secure lockdown testing browser to administer these assessments, which was required for summative testing. Figure 3.1 presents the student CAP login screen.

Figure 3.1. CAP Student Login Screen


nwea
Log in
Session Name
Nombre de Sesión
Session Password
Contraseña de Sesión
© NWEA 2017. MAP is a registered trademark. NWEA, MAP Growth and MAP Skills are trademarks of NWEA in the U.S. and in other countries.
nwea

Please enter the Session Name and Session Password

3.2. User Roles and Responsibilities

Table 3.1 summarizes the user roles and responsibilities regarding the Spring 2018 NSCAS summative administration.

Table 3.1. User Roles and Responsibilities

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

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

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

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

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

3.3. Administration Training

In addition to district- and school-held trainings, NWEA, in collaboration with the NDE, held two trainings for district leaders in advance of testing. The Fall 2017 regional workshop from October 10–16, 2017, were half-day, in-person workshops held across multiple regions of the state. Information on spring summative (including test sessions, accessibility, and student rostering was presented. The summative test administration workshop from Feb 19–26, 2018, were two-hour virtual sessions that provided important information on the NSCAS Summative assessments. Table 3.2 presents the locations and number of participants based on the registration numbers for the Fall 2017 regional workshop, and Table 3.3 presents the dates and number of participants based on the registration numbers for the summative test administration workshop. Appendix D presents the PowerPoint training presentations for each training.

Table 3.2. Fall 2017 Regional Workshop Locations and Participation

Location	#Participants
Scottsbluff – Gering Civic Center	80
Kearney – Younes Hospitality	175
West Point – Nielsen Community Center	89
Lincoln – The Cornhusker Marriott	104
Omaha – DC Centre	93

Table 3.3. Summative Test Administration Workshop Dates and Participation

Date	#Participants
February 19, 2018	66
February 21, 2018	82
February 22, 2018	59
February 23, 2018	59
February 26, 2018	44

3.4. Practice Tests

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

The practice tests were not adaptive and had the same 20 items for each grade in a content area. They were also untimed, although the estimated test-taking time for each was 40 minutes. Unlike the actual summative assessments, progress on the practice test 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.

A Practice Test Manual was provided on the NSCAS Assessment Portal with information on the Practice Test, how to access it, and recommended proctor scripts. The purpose of the practice tests was to allow students to experience the types of items, tools (e.g., calculator), and item aids (e.g., highlighter) available on the actual summative assessments. They also allowed other stakeholders such as parents and administrators to experience the summative assessment environment. For the best student experience, it was recommended that students view the Online Student Tutorial located on the NSCAS Assessment Portal to learn about the available tools and their uses before taking the practice tests. 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 in the NSCAS Summative assessment.

3.5. Accommodations and Accessibility Features

Table 4.4 presents the accessibility supports available for the Spring 2018 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, created by the NDE.

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

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

Table 3.4. Accommodations and Universal Features

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

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

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

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

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, the grade level, the construct, or the content being measured. Accommodations should have only been used if they are appropriate for the student and used during instruction throughout the year. Modifications are adjustments or changes in the test that affect test expectations, the grade level, or the construct of content being measured. Modifications were not acceptable in the NSCAS assessments.

3.5.1. Paper-Pencil Participation Criteria

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

3.5.2. Participation of English Language Learners (ELLs)

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

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

- In determining appropriate linguistic supports for students in the NSCAS system, districts should use the NSCAS Summative & Alternate Accessibility Manual.
- Districts must be aware of the difference between linguistic supports (accommodations for ELLs) and modifications.
- For ELL students, 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, the grade level, or the construct or content being measured. Modifications are not acceptable in the NSCAS assessments.

3.5.3. Participation of Recently Arrived Limited English Proficient (RAEL) Students

Recently arrived limited English proficient (RAEL) students are defined by the U.S. Department of Education as students with limited English proficiency who attended schools in the United States for fewer than 12 months. The phrase “schools in the United States” includes only schools in the 50 states and the District of Columbia. It does NOT include Puerto Rico. Districts must have provided access to all RAEL students on all NSCAS assessments each year based on the grade level of the student using linguistic supports. RAEL students are included in accountability, but the categorization has changed and is detailed below.

- In Year 1: Students are included in participation calculations, but results are excluded on the ELA and Mathematics assessments in the state accountability system.
- In Year 2: Students are included in participation calculations, and results are used in growth measures but not achievement indicators in the state accountability system.
- In Year 3: Students are included in all accountability calculations.

3.6. Test Security

In a centralized testing process, it is critical that equity of opportunity, standardization of procedures, and fairness to students is maintained. Therefore, as part of the NDE test security practices, the 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 the NDE’s Statewide Assessment Office. Districts were also 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 the NDE 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 document. The NDE encouraged all districts with questions to contact their own local school attorney for customization of such a policy.

As part of NDE's security policy, the principal of each school participating in the NSCAS Summative assessments were required to complete and sign a Building Principal Security Agreement and return it to the Statewide Assessment Office. District Administrator Contacts were required to complete and sign the District Administrator Contact Confidentiality of Information Agreement and return it to the Statewide Assessment Office. 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.

3.6.1. Online Test Security

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

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

During testing there was a 10-minute inactivity setting that engaged after 10 minutes of no activity in the testing environment. Activity consisted of any mouse movement. This was a measure to maintain the test security should the student step away from the desk.

3.6.2. Paper-Pencil Test Security

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

3.6.2.1. Physical Warehouse Security

All EDS personnel—including subcontractors, vendors, and temporary workers who have access to secure test materials—were required to agree to keep the test materials secure and sign security forms that state the understanding of the secure nature of test items and the confidentiality of student information.

Access to the document-processing warehouse was by rolling gates, which were always locked except when opened to allow pickup or receipt of test materials. A secure chain-link fence with a barbwire top surrounds the document-processing facility. A verified electronic security system monitored access to the offices and warehouse areas 24 hours a day, seven days a week. All visitors entering the facility were required to sign in at the front desk and obtain an entry badge that allowed them access to the facility. The following additional security procedures were maintained for the NSCAS Summative program:

- Test materials received from the printing subcontractors were stored in a secure warehouse facility prior to packaging and shipping to districts.
- All boxes and pallets placed in the secure warehouse for long-term storage were recorded electronically so that they could be retrieved at any time. Scanned (used) answer documents were stored in labeled "scan" boxes on labeled pallets in the same warehouse. The scan box and pallet numbers were scanned into a database for retrieval as needed. Documents are stored until the second week of January following the test administration or until the NDE provides express written consent to destroy them.

3.6.2.2. Secure Destruction of Test Materials

EDS will manage the secure destruction of test materials during the first two weeks of January 2019. Using the long-term storage database, EDS will retrieve the documents and systematically destroy them through a secure shredding process. The shredding company uses a high-capacity mobile onsite document destruction vehicle that provides the most advanced document destruction technology in the industry. The shred trucks, equipped with a 20-inch monitor so EDS staff may monitor the documents going into and being expelled in a pulverized state, provide the quickest, most complete, and most confidential destruction of sensitive documents. Every sensitive document is pulverized using a *hammermill* process that creates the smallest pieces in the document destruction industry.

After the test materials destruction process is complete, the shredding company provides a certificate of destruction that will remain on file at EDS. The long-term storage database will be updated to reflect that the materials have been destroyed. During the first two weeks of January 2019 and upon written approval, EDS will also delete the answer document images from the server hard drive and all backup drives. The deletion process will securely erase the data to ensure that the images cannot be retrieved through data restorative means. EDS will provide the NDE with archives of all data files prior to deletion, upon request.

3.6.2.3. Shipping Security

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

3.6.2.4. Electronic Security of Test Materials and Data

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

3.6.3. Caveon Test Security

3.6.3.1. Monitoring for Disclosure of Test Content

Caveon Web Patrol performed online searches with the specific goals of detecting, reporting and eliminating, where possible, online exposures and infringing content of NSCAS Summative assessments. During the administration windows, all materials and information shared with Caveon Web Patrol by either NWEA or the NDE, other than live exam items, were kept in Caveon's secure incident management platform, Caveon Core. This system was end-to-end encrypted with permissions and role-based access. Use of materials, other than live test items, were limited only to four Caveon employees specifically assigned to this project. In addition, those employees, and all Caveon Web Patrol employees, signed non-disclosure agreements before engaging in work for any client, including NWEA and the NDE. Further, they were trained 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. To ensure the highest level of security, all live test items provided to Caveon Web Patrol by the NDE or NWEA were stored on an air gapped computer and were only accessible by Caveon's Executive Web Patrol Manager for the sole purpose of threat verification. No live items, under any circumstances, were used in the Caveon Web Patrol online detection process. Once infringing content was found and verified, it was reported to the NDE through the notification tools built into Caveon Core and a concurrent email message from the Web Patrol Director of Operations or Executive Web Patrol Manager notifying NWEA and the NDE of the potential infringements.

3.6.3.2. Monitoring for Potential Test Security Violations

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

Table 3.5. Statistical Analysis and Potential Incidents

Statistical Analysis	Potential Incident
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).

Statistical Analysis	Potential Incident
Person-fit (Aberrance) Statistics	When students respond in a manner that is inconsistent with the student population, supportive evidence of pre-knowledge or test coaching may be present.
Scored Differences	Non-scored items are typically being field tested and are usually newly created. Large performance differences between the operational and non-scored items suggest that students may have access to the test content prior to the exam.
Item Performance Changes	Performance shifts, indicating the items have become easier during the test administration window, provide evidence that the item might have been disclosed to the students.
Exposed Differences	Item exposure (i.e., administrations to individual students) vary in CAT pools. When student performance is higher on frequently exposed items than on the other items, there is a possibility that some or a few students had access to the test content prior to the exam.
Linking Difference	Linking items are administered to nearly all students. Due to their greater exposure, these items have a greater risk of being compromised. This statistic compares each student's performance on these items against their performance on the non-linking items to determine if any students potentially had pre-knowledge of the linking items.
Breached Difference	This statistic compares each student's performance on the group of breached items against their performance on the remaining items to determine if any students potentially had pre-knowledge of these breached items.
Pauses and Durations	Many pauses during the exam or exams which require several days for administration increase the possibility of test content disclosure.
M4 Similarity	Exams that use fixed forms (e.g., Science) were analyzed for excessive agreement between pairs of students. These statistics can identify where answer copying by students, sharing of test responses between students, or large-scale collusion may have occurred.
Identical Test	When students receive the same items (i.e., because they were administered the same form), it is possible they may have identical responses to the items. This is more likely when they use the same disclosed answer key. When this happens, students will often have very high scores on the exam.
Perfect Test	A concentration of perfect scores at a school, which are very unusual, may indicate the presence of a test security incident.
Synchronicity	When students answered questions at or near the same time of day, there is a possibility that they were guided or paced through the exam, which should not happen. This analysis detects potential incidents when this occurred.

As provided in the data forensics report from Caveon (Mulkey, Maynes, & Scott, 2018), data for 323,457 test instances administered at 828 schools in 246 districts were analyzed. The most significant findings are as follows. Overall, the assessments appeared to have been administered securely.

- Fourteen test instances were flagged for extreme similarity for Science. These 14 test instances formed seven pairs of extremely similar tests. The observed similarity is extremely improbable under the assumption of independent test taking.
- Overall, there was not sufficient evidence to indicate than any school was involved in a security violation for the NSCAS ELA and Mathematics assessments.
- High detection rates by the M4 Similarity statistic, accompanied by increased performance for detected test instances, may be evidence of a security violation for Science Grade 8.

- High detection rates by the Synchronicity statistic, accompanied by increased performance for detected test instances, may be evidence of a security violation for a few schools.
- At the school level, three school-subject-grade groups had high detection rates by the Linking Difference statistic. For these groups, the Linking Difference anomalies were not associated with improved performance. At most, 13 individuals could have increased their scores through pre-knowledge of the linking items.
- On April 23, 2018, an educator was caught transcribing items from the NSCAS Mathematics Grade 7 assessment. This material was confiscated, and NWEA identified 72 items. At the school level, one school-subject-grade group had anomalous results related to the breached items. For this group, Breached Difference detections were associated with lower performance. At most, three individuals could have increased their scores through pre-knowledge of these items.

3.7. Partner Support

NWEA's Partner Support Services team provided implementation and technical support throughout the 2017–2018 school year for the NSCAS Summative assessments. This team provided resources to support Nebraska and its educators, assisting with generating roster files, configuration of the assessment program, accessing online reports, and general questions with the use of the online assessment system.

NWEA provided phone, email, and chat support to schools and educators from 8:00 a.m. to 5:00 p.m. Central Time (CT) Monday through Friday, and 7:00 a.m. to 5:00 p.m. CT during the testing windows. Table 3.6 presents the number of cases presented to the Partner Support team by case type for the entire 2017–2018 school year from August 2017 to June 2018 for the NSCAS Summative tests. More than half of the cases were related to testing (i.e., administration questions).

- **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. Number of NSCAS Cases to Partner Support in 2017–2018

Case Type	#Cases	% of Total Cases
Student Mobility	6	0.4
Reports	22	1.4
Navigation	89	5.6
Setup and Management	354	22.3
Rostering	69	4.4
Testing	1,044	65.9
Total	**1,584**	**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 three-week training program led by the NWEA Senior Support Specialist team and team trainers. The training program consisted of a combination of instructor-led and self-paced eLearning courses, covering all relevant team policies and procedures, including security requirements of handling student data, product expertise, and troubleshooting requirements. In addition, several days of “phone shadowing” were built into the program to ensure that each new staff member had the opportunity to participate in calls with veteran staff monitoring prior to working independently. Senior Support Specialists were responsible for continually updating training program content to ensure that all support team staff members were knowledgeable of current policies. In addition, the project managers and product training resources were dedicated to NDE’s program to train the support staff on Nebraska-specific policies. This equated to roughly 90 hours of training (80 for initial training, 10 for state-specific training and the new platform).

Section 4: Scoring and Reporting

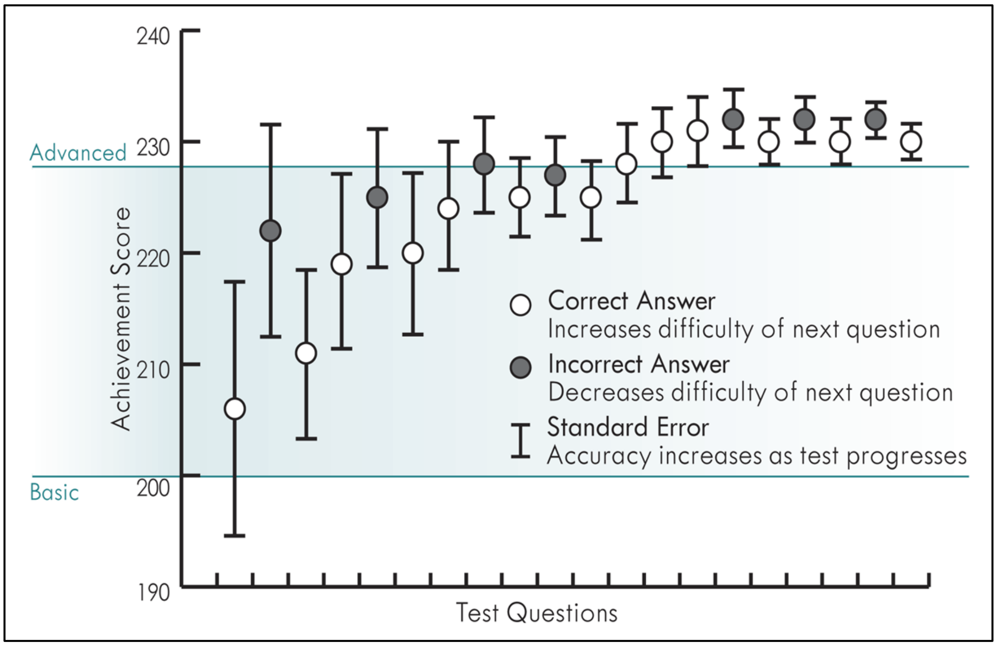
4.1. Scoring Process

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

4.1.1. Constraint-Based Adaptive Engine

A CAT adapts items to match the ability level of the student. Students receive different items based on item difficulty and their ability levels. For example, students with lower ability levels (based on their answers to previous items) receive easier items compared to students with higher ability levels who receive harder items as the test progresses. A constraint is a rule given to the engine when selecting items. For example, the engine must meet the TOS when considering the next item. The adaptive engine uses the TOS and a student's momentary theta (θ) to drive item selection, as shown in Figure 4.1. Momentary theta is the ability estimate of the student that is recalculated and updated after answering each item.

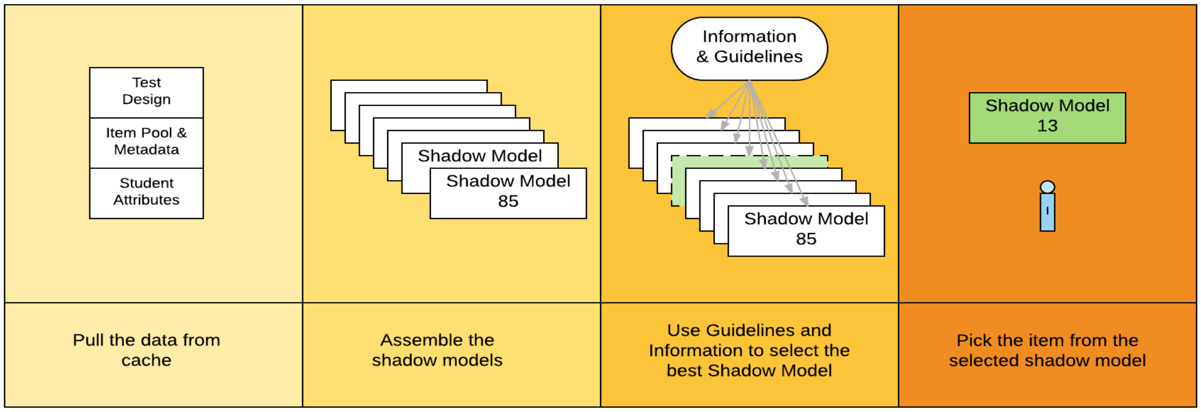
Figure 4.1. Adaptive Engine Overview



Items were selected based on the 2017 scale as a reference for item difficulty. With this context, the goal of the constraint-based engine's item selection was to provide a test that meets "must-have" (hard) constraints and "nice-to-have" (soft) constraints. Examples of hard constraints are all item selection constraints, such as all levels of standards, field test items, and operational items. Examples of soft constraints are student population exposure goals and population exposure limits by anchor items.

The adaptive engine has two stages of consideration as it selects the next item that conforms to the TOS while providing the maximum information about the student based on the student's momentary ability estimate: the shadow test approach and a variation of the weighted penalty model. As shown in Figure 4.2, the shadow test approach (Van der Linden & Reese, 1998) selects items based on the required aspects of the TOS, and a new valid shadow model is selected upon each update to the student's momentary theta. In other words, this approach uses the student's answer to the last item to create shadow models that are waiting "in the shadows" while the student answers the current item. When the student responds to the item, that answer is used to select the next correct shadow model. Because multiple shadow models can be drawn from an item pool, a variation of the weighted penalty model (Segall & Davey, 1995) then selects which shadow model is optimal based on additional content guidelines while ensuring the most representative sample for linking and field test items. The shadow model with the smallest penalty is selected when multiple shadow tests meet the required attributes of the test and have similar information.

Figure 4.2. Shadow Test Approach



4.1.2. Scoring Rules

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

Table 4.1. Attemptedness Rules for Scoring

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

The attemptedness rule was decided based on the results of the standard error of measurement (SEM) that became relatively stable after 10 operational items from the simulation data and the finding of a small number of 2017 students who attempted less than 10 items. Regarding scoring, NWEA ran analyses using a subpopulation of the 2017 students and found that the number of not-reached items increased the amount of estimation error, suggesting larger estimation error with the penalty function (i.e., to score those not-reached items as wrong).

However, scoring consistency were also considered for fixed forms (e.g. Science). Thus, the NDE made the following scoring rules in consultation with the state and district coordinators:

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

4.1.3. Paper-Pencil Scoring

4.1.3.1. Scanning of Answer Sheets by EDS

EDS scanned and imaged all paper-pencil student responses and captured student demographic information provided on the paper-pencil answer sheets. Answer sheets were scanned using the high-speed optical mark reading (OMR) NCS 5000i scanners. The scanning, editing, and scoring processes were performed after most answer sheets were returned by districts. Answer sheets were scanned and edited in accordance with the NSCAS data processing specifications, created collaboratively by EDS and NWEA. The editing processes included steps to check the spelling of the student name (i.e., that the scanner picked up all the bubbled letters and that there were no multiple marks, no embedded blanks, and no initial blanks in the name) and that the scanner picked up all the bubbled digits in the NSCAS Identifier.

Since some answer sheets contained preprinted precoded information from the roster file, the student demographics provided via the roster file was merged into the scan file so that all demographics and scan marks were included in one file. This merge was completed on the barcode ID number printed on the answer sheet. Checks were performed to eliminate duplicate barcode numbers during each step of the merging process.

Finally, EDS created a Scan Export File for each grade to merge the scan data with the NWEA response “choice” conversion data. The resulting data (student demographics, scan marks and choice conversation values) was transferred in JSON format to NWEA via an API.

4.1.3.1.1. Quality Control of Scanning and Scoring

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

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

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

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

4.1.3.1.2. Quality Control of Image Editing

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

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

4.1.3.1.3. Quality Control of Answer Document Processing and Scoring

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

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

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

4.1.3.2. Scoring by NWEA

The paper-pencil scanned documents were converted to JSON and ingested by an NWEA API. The data was then matched to existing student records and new test events were created. The test events were designated with a non-tested code (NTC) of PPA. The test events went through the constraint engine and were scored based on the test models developed for PPA. The records were treated as the online records and went through the normal scoring process. There was one exception for Grade 8, Item 26, that contained an extra bubble on the answer sheet. It was scored the following way: If the student bubbled "E" or contained a multiple mark with "E," the item was scored as 0.

4.2. Score Reporting Methods

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

Table 4.2. Scale Score Ranges

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

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

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

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

Table 4.3. Achievement Level Descriptions

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

4.3. Report Summary

The following reports were produced for the 2017–2018 NSCAS Summative test administration and made available in October 2018. Appendix E presents examples of each report. A Reports Interpretive Guide was developed to help district leaders understand, explain, and use the NSCAS results and was made available on the NSCAS Assessment Portal. A separate Individual Student Report (ISR) Reports Interpretive Guide was also created for parents. All reports were delivered online in CAP according to user role. Printed ISRs were also delivered to districts.

- Student-Level Reports
    - Individual Student Report (ISR)
    - Individual Student Report (ISR) with Non-Tested Code (NTC)

- School-Level Reports
    - School Roster
    - School Performance Level Summary

- District-Level Reports
    - District Performance Level Summary
    - State-Level Reports

- State Performance Level Summary

4.3.1. Student-Level Reports

ISRs showed a student's performance on the NSCAS summative tests. Reports were posted in PDF format online within CAP. School districts and state administrators could download them in English or Spanish from the CAP View Reports page. Two copies of each ISR in English were also printed, sorted by school, and delivered to each district. One copy was sent home with the student, and the second copy was to be filed in the student's cumulative folder.

If a non-tested code (NTC) was applied to any content area, the student's achievement level scores and proficiency by reporting category within the respective content area were reported as affected by the NTC, as defined in Table 5.3. If a student had a NTC of INV, PAR, SAE, or UTT assigned to his or her test, the automatically assigned score displayed with a score of one less than the lowest scale score for that grade and content.

Table 4.4. Non-Tested Codes (NTCs)

	Achievement Level Received		
NTC	ELA & Mathematics	Science	Description
ALT Alternate assessment	N/A	N/A	Student participated in the alternate assessment.
EMW Emergency Medical Waiver	No level	No level	Student was not tested because of an approved emergency medical waiver.
INV Score invalidated by the state	Developing	Below the Standards	Student's test was invalidated; student received the lowest scale score for that grade and content area and an achievement level of Developing/Below the Standards.
NLE No longer enrolled	No level	No level	Student was not enrolled in the district at the time of testing.
PAR Parental refusal	Developing	Below the Standards	Student was not tested because of a written request from a parent or guardian; student received the lowest scale score for that grade and content area and an achievement level of Developing/Below the Standards.

NTC	Achievement Level Received		Description
	ELA & Mathematics	Science	
SAE Student absent for entire test window	Developing	Below the Standards	Student was absent during the entire test window; student received the lowest scale score for that grade and content area and an achievement level of Developing/Below the Standards.
UTT District was unable to test student	Developing	Below the Standards	District was unable to test student; student received the lowest scale score for that grade and content area and an achievement level of Developing/Below the Standards.

4.3.2. School-Level Reports

The School Roster report listed students who were required to take the NSCAS Summative tests and presented a report of their performance. The size of this document depended on the class size. The School Performance Level Summary report presented a summary of performance and demographics for all students at a school by grade who were required to take the NSCAS Summative tests.

4.3.3. District-Level Reports

The District Performance Level Summary report was for internal district use only and was required for state and federal reporting purposes. It was available for district-level personnel to access through the Reports landing page within CAP. Information to protect small numbers of students was not suppressed.

4.3.4. State-Level Reports

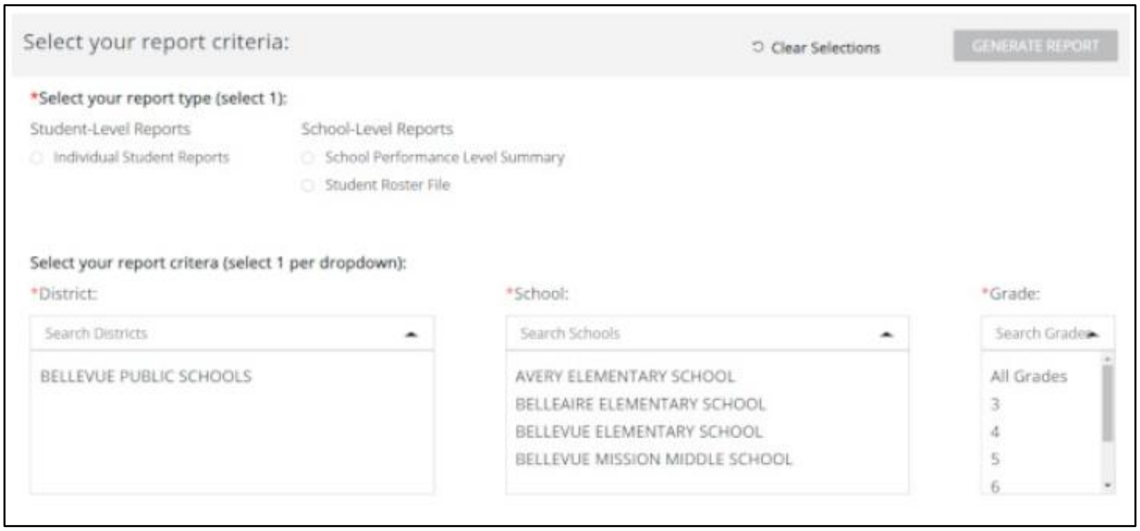
The State Performance Level Summary report presented the average state performance based on demographics for the NSCAS Summative tests. It was available for state-level personnel to access through the Reports landing page in the CAP.

4.4. Reporting Process

4.4.1. Online Reports

To access the online reports, users generated reports in the reports landing page based on their role, as shown in Figure 4.3. Users selected the report type (e.g., ISR, school performance summary, etc.) and criteria (e.g., district, school, and grade) before hitting the “Generate Report” button. The user’s role interacted properly constrained users in the reports landing page to only access reports they were authorized to see. For example, teacher-level users would only be able to access student reports for students in their own classes. The reporting page was also protected by the same security measures that applied to every aspect of CAP.

Figure 4.3. Reports Landing Page Example—District Assessment Contact



Select your report criteria:

*Select your report type (select 1):

Student-Level Reports
○ Individual Student Reports

School-Level Reports
○ School Performance Level Summary
○ Student Roster File

Select your report critera (select 1 per dropdown):

*District:
Search Districts
BELLEVUE PUBLIC SCHOOLS

*School:
Search Schools
AVERY ELEMENTARY SCHOOL
BELLEAIRE ELEMENTARY SCHOOL
BELLEVUE ELEMENTARY SCHOOL
BELLEVUE MISSION MIDDLE SCHOOL

*Grade:
Search Grades
All Grades
3
4
5
6

4.4.2. Printed ISRs

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

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

4.4.3. Report Verification

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

The objectives of report verification were to ensure that:

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

The following report sections were checked during the QA process:

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

4.5. Matrix

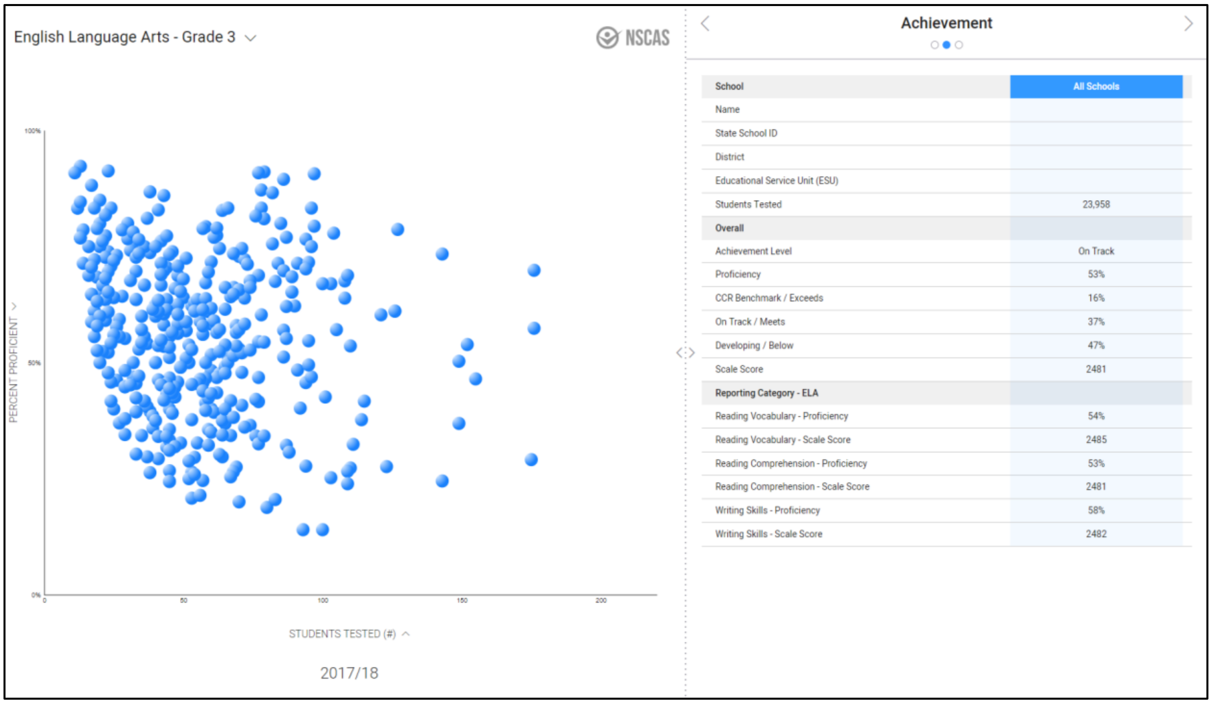
NWEA worked collaboratively with ESC to use ESC's tools to view web-based visualizations for the NSCAS Summative assessments, including combinations of aggregate and disaggregate information of results by demographics and other filtering options. This system, referred to as the Matrix, also allowed users to save and print specific plot and screen images from the interactive visualization. Through ESC's Matrix, users could interact with and explore many different levels of information to answer targeted questions about their organization (district, school, or state). The main feature of this tool was an interactive scatter chart designed to display longitudinal data, as shown in Figure 4.4, Figure 4.5, and Figure 4.6. On the Matrix, the X and Y axes were modifiable.

Districts and educational service units (ESUs) were provided direct access to the Matrix, and role-based filter conditions of the Matrix were available for state personnel and researchers who had a deep familiarity with the data. District Administrator Contacts and School Assessment Coordinators also had access. All user roles except ESUs accessed the Matrix through a hyperlink on the Reports Landing page in CAP. ESC developed videos on the navigation aspects of the Matrix to help users learn how to best use the tool. In collaboration with the NDE, ESC also developed professional development videos available on the Matrix user-based website upon initial log in for users to help them understand how to interpret and apply the data.

The state user could see data from all districts, and the School Assessment Coordinator and District Administrator Contact could only download and view data respective to their location. For example, schools saw school data, and districts saw schools within the district. Each visualization contained dropdown menus for exploring content areas and schools. This feature allowed for targeted conversations and professional development. For example, a principal could have a specific conversation with Grade 3 teachers about Grade 3 reading by simply selecting it from the dropdown menu.

Each visualization allowed the user to access and print the state-, district-, or school-level PDF report. Screen and plot images could also be saved and exported for use in other documents. The default setting of the Matrix was interactive scatterplots. Users could also change to a spreadsheet view and construct a spreadsheet from all the available variables within the visualization. This feature allowed for easy access to high-quality data that had gone through rigorous auditing. Users could then explore and sort data to meet their individual needs. No suppression rules were applied to the Matrix for the state-level use role. Suppression rules were applied to the Matrix for District Administrator Contact and School Assessment Coordinator user roles. For example, all data was suppressed for a school if the number of tested students was less than 10.

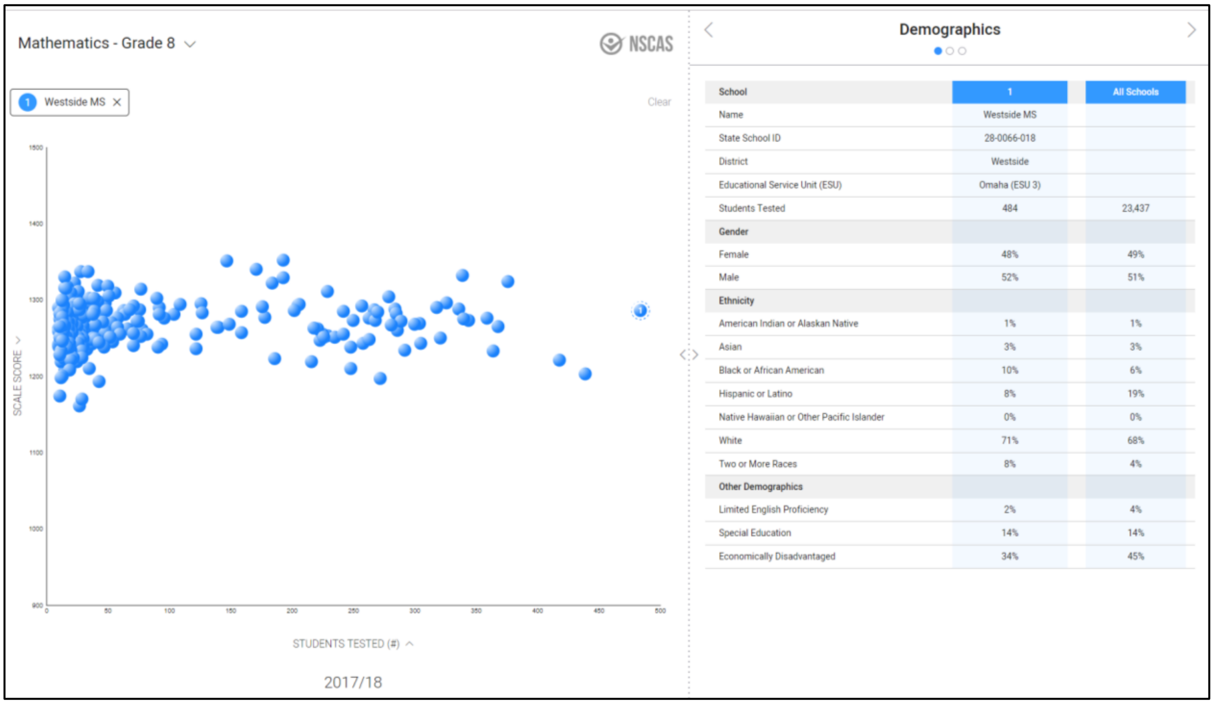
Figure 4.4. Matrix Example: Percent Proficient



Achievement

School	All Schools
Name	
State School ID	
District	
Educational Service Unit (ESU)	
Students Tested	23,958
Overall	
Achievement Level	On Track
Proficiency	53%
CCR Benchmark / Exceeds	16%
On Track / Meets	37%
Developing / Below	47%
Scale Score	2481
Reporting Category - ELA	
Reading Vocabulary - Proficiency	54%
Reading Vocabulary - Scale Score	2485
Reading Comprehension - Proficiency	53%
Reading Comprehension - Scale Score	2481
Writing Skills - Proficiency	58%
Writing Skills - Scale Score	2482

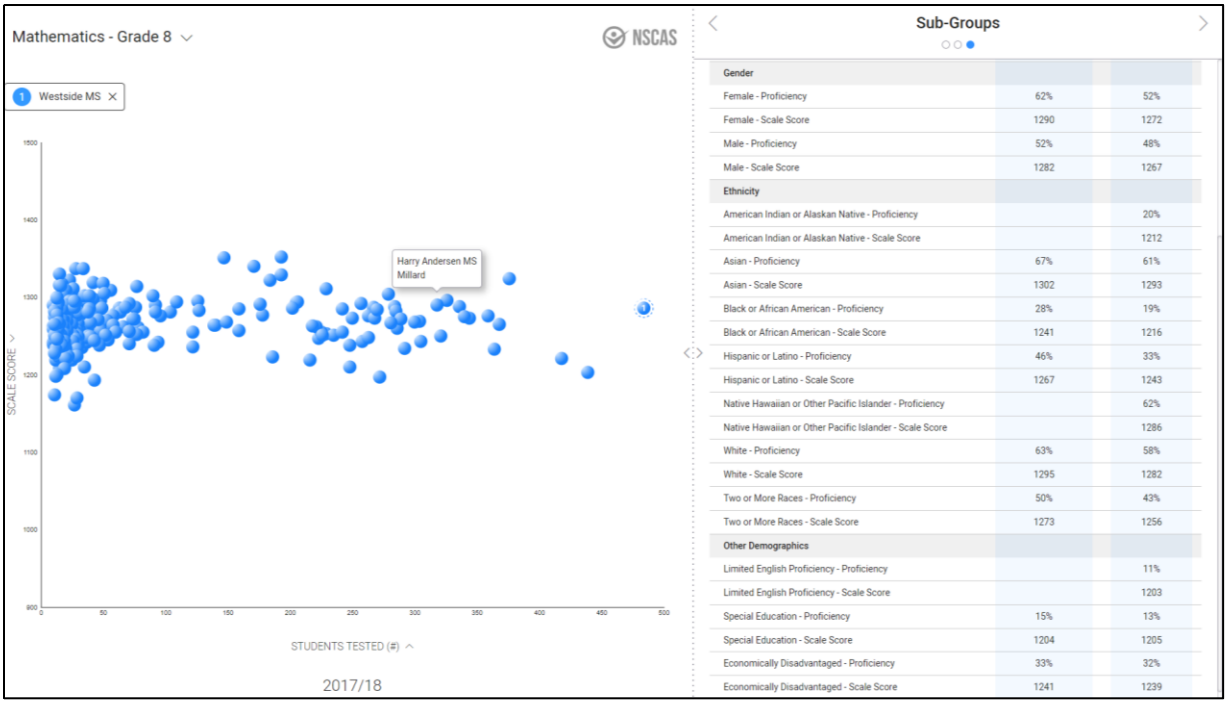
Figure 4.5. Matrix Example: Scale Score by Demographics



Demographics

School	1	All Schools
Name	Westside MS	
State School ID	28-0066-018	
District	Westside	
Educational Service Unit (ESU)	Omaha (ESU 3)	
Students Tested	484	23,437
Gender		
Female	48%	49%
Male	52%	51%
Ethnicity		
American Indian or Alaskan Native	1%	1%
Asian	3%	3%
Black or African American	10%	6%
Hispanic or Latino	8%	19%
Native Hawaiian or Other Pacific Islander	0%	0%
White	71%	68%
Two or More Races	8%	4%
Other Demographics		
Limited English Proficiency	2%	4%
Special Education	14%	14%
Economically Disadvantaged	34%	45%

Figure 4.6. Matrix Example: Scale Score by Sub-Groups



Sub-Groups

Gender		
Female - Proficiency	62%	52%
Female - Scale Score	1290	1272
Male - Proficiency	52%	48%
Male - Scale Score	1282	1267
Ethnicity		
American Indian or Alaskan Native - Proficiency		20%
American Indian or Alaskan Native - Scale Score		1212
Asian - Proficiency	67%	61%
Asian - Scale Score	1302	1293
Black or African American - Proficiency	28%	19%
Black or African American - Scale Score	1241	1216
Hispanic or Latino - Proficiency	46%	33%
Hispanic or Latino - Scale Score	1267	1243
Native Hawaiian or Other Pacific Islander - Proficiency		62%
Native Hawaiian or Other Pacific Islander - Scale Score		1286
White - Proficiency	63%	58%
White - Scale Score	1295	1282
Two or More Races - Proficiency	50%	43%
Two or More Races - Scale Score	1273	1256
Other Demographics		
Limited English Proficiency - Proficiency		11%
Limited English Proficiency - Scale Score		1203
Special Education - Proficiency	15%	13%
Special Education - Scale Score	1204	1205
Economically Disadvantaged - Proficiency	33%	32%
Economically Disadvantaged - Scale Score	1241	1239

Section 5: Psychometric Analyses

During the Spring 2018 testing window, the pre-equated item parameter estimates were used to score students and select the next items to administer for the adaptive portions of the NSCAS Summative ELA and Mathematics assessments. A constraint engine evaluation was conducted for ELA and Mathematics using these pre-equated estimates to determine whether the engine performed as expected during administration. After the testing window was closed, the following post-administration analyses were conducted to calibrate the items for ELA, Mathematics, and Science (e.g., to construct the ELA and Mathematics vertical scales). The purpose of conducting these analyses is to establish the psychometric quality of the items used in the assessments, which will bolster the arguments regarding the validity of the interpretations and uses of the test scores.

- Classical item analyses
- Differential item functioning (DIF)
- Item response theory (IRT) calibration
- Equating and scaling

5.1. Number of Student Included in the Analyses

Table 5.1 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). 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, the 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 5.1. Number of Students Included in the Psychometric Analyses

Content Area	Grade	Test ID	N
ELA	3	3220	23,875
	4	3221	23,873
	5	3222	22,290
	6	3223	23,322
	7	3224	22,965
	8	3225	23,252
Mathematics	3	3241	23,858
	4	3242	23,826
	5	3243	22,249
	6	3244	23,277
	7	3245	22,893
	8	3246	23,177
Science	5	3268	22,251
	8	3305	23,190

5.2. Constraint Engine Evaluations

Pre- and post-administration adaptive engine evaluation studies are important evidence, along with post-administration analyses, for confirming (or disconfirming) interpretation and test score use arguments regarding student proficiency with the state standards. Pre-administration simulations were conducted prior to the Spring 2018 operational testing window to evaluate the constraint engine's item selection algorithm and estimation of student ability based on the TOS to meet the state's long-term vision for the interpretation of the test scores given the item pool depth for ELA and Mathematics. Because the 2018 NSCAS ELA and Mathematics assessments were a conversion of a bank developed for fixed-form assessments that had had limited item development to the newly adopted state standards, NWEA's a priori simulation goal was that the constraint engine should function at least as well as the previous year fixed forms. The simulation tool used the operational constraint engine, thereby providing results with the same properties and functionality as what would be seen operationally. Detailed information regarding the pre-administration simulation study can be found in the full report (NWEA, 2018a).

After the Spring 2018 testing window closed, a post-administration constraint engine evaluation study was then conducted to determine whether the constraint engine performed as expected. The results included a blueprint constraint accuracy analysis to determine whether the constraint engine administered the assessments based on the TOS; item exposure rates to determine the number of items administered to students; score precision and reliability; and population explore for linking and field test items. This section provides blueprint constraint accuracy and item exposure results from the evaluation study. Score precision and reliability results are provided in Section 8.1. . Detailed information regarding all results of the post-administration evaluation study can be found in the full report (NWEA, 2018b).

Overall, the constraint engine performed as it should based on the blueprint (i.e., TOS) constraints. The strand points had a 100% match. The points at the indicator level are also matched to the blueprints. The constraint engine also showed a similar performance when estimating the students' ability in terms of SEM and reliability (see Section 8.1. for the results). Item exposure rates were also acceptable given that the constraint engine used almost half of the items to administer the test and most used items had a 0–20% exposure rate.

5.2.1. Blueprint Constraint Accuracy

Table 5.2 and Table 5.3 present the blueprint constraint results from the post-administration engine evaluation study at the strand level for ELA and Mathematics, respectively. As shown, the number of items and points at the strand level resulted in a 100% match based on the blueprint adjustment. Results were also provided at the indicator level by passage type selection, DOK level, and item range requirements (NWEA, 2018b). While most DOK levels also resulted in a 100% match, some indicators did not because the constraint engine used DOK level as a guideline or a "nice to have" given the limited number of items at a specified DOK level for some indicators. Passage type also resulted in a less than 100% match for some indicators. These findings appeared similar to those in the simulation study, as expected. Further, overall, the matching rate at the indicator level has increased compared to the simulation result, with some decreased matching rates.

Table 5.2. Blueprint Constraint by Strand—ELA

		#Items			#Points		
Grade	Strand	Min.	Max.	%Match	Min.	Max.	%Match
3	Vocabulary	10	10	100.0	10	10	100.0
	Comprehension	22	24	100.0	27	29	100.0
	Writing	8	8	100.0	12	14	100.0
4	Vocabulary	9	10	100.0	9	10	100.0
	Comprehension	24	24	100.0	28	28	100.0
	Writing	8	8	100.0	11	12	100.0
5	Vocabulary	9	9	100.0	9	9	100.0
	Comprehension	22	24	100.0	28	30	100.0
	Writing	10	10	100.0	14	14	100.0
6	Vocabulary	9	9	100.0	9	9	100.0
	Comprehension	22	23	100.0	27	29	100.0
	Writing	9	10	100.0	15	16	100.0
7	Vocabulary	9	9	100.0	9	9	100.0
	Comprehension	22	22	100.0	28	28	100.0
	Writing	10	10	100.0	12	12	100.0
8	Vocabulary	9	9	100.0	9	9	100.0
	Comprehension	21	24	100.0	28	31	100.0
	Writing	11	11	100.0	15	16	100.0

Table 5.3. Blueprint Constraint by Strand—Mathematics

		#Items			#Points		
Grade	Strand	Min.	Max.	%Match	Min.	Max.	%Match
3	Number	16	16	100.0	17	17	100.0
	Algebra	6	6	100.0	7	7	100.0
	Geometry	11	11	100.0	12	12	100.0
	Data	8	9	100.0	9	10	100.0
4	Number	17	19	100.0	18	20	100.0
	Algebra	10	12	100.0	11	12	100.0
	Geometry	8	10	100.0	9	11	100.0
	Data	6	8	100.0	7	9	100.0
5	Number	16	17	100.0	17	18	100.0
	Algebra	10	10	100.0	10	11	100.0
	Geometry	8	9	100.0	9	10	100.0
	Data	6	6	100.0	7	7	100.0
6	Number	11	12	100.0	12	13	100.0
	Algebra	14	15	100.0	15	16	100.0
	Geometry	8	9	100.0	9	10	100.0
	Data	7	8	100.0	8	9	100.0
7	Number	9	9	100.0	10	10	100.0
	Algebra	14	15	100.0	15	16	100.0
	Geometry	8	9	100.0	9	10	100.0
	Data	9	10	100.0	10	11	100.0

Grade	Strand	#Items			#Points		
		Min.	Max.	%Match	Min.	Max.	%Match
8	Number	10	12	100.0	11	12	100.0
	Algebra	13	14	100.0	14	15	100.0
	Geometry	12	13	100.0	13	14	100.0
	Data	5	5	100.0	6	6	100.0

5.2.2. Item Exposure Rates

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

All horizontal linking items were also part of the item exposure rate calculation. Horizontal Form 1 given to all students had a 100% exposure rate and is therefore included in the 81–100% exposure rate bin, and the horizontal linking Set A and Set B each had an approximately 50% exposure rate and are therefore included in the 41–60% exposure rate bin. These patterns of exposure rate are very similar to the simulation results. Most grades used half of the items from the item pool except ELA Grade 7 that used 29% of the item pool. However, most items across grades and content areas had a 0–20% exposure rate as expected.

Table 5.4. Item Exposure Rates

Grade	#Items			Exposure Rate									
				0–20%		21–40%		41–60%		61–80%		81–100%	
	Total	Administered	Unused	N	%	N	%	N	%	N	%	N	%
ELA													
3	449	209	240	154	73.7	19	9.1	10	4.8	2	1.0	24	11.5
4	448	353	95	292	82.7	24	6.8	19	5.4	4	1.1	14	4.0
5	400	188	212	126	67.0	15	8.0	23	12.2	8	4.3	16	8.5
6	427	235	192	175	74.5	17	7.2	18	7.7	5	2.1	20	8.5
7	418	123	295	64	52.0	10	8.1	26	21.1	4	3.3	19	15.5
8	429	186	243	136	73.1	14	7.5	4	2.2	6	3.2	26	14.0
Mathematics													
3	384	216	168	171	79.2	9	4.2	6	2.8	3	1.4	27	12.5
4	191	102	89	53	52.0	6	5.9	12	11.8	4	3.9	27	26.5
5	237	161	76	109	67.7	9	5.6	18	11.2	5	3.1	20	12.4
6	470	247	223	198	80.2	6	2.4	21	8.5	6	2.4	16	6.5
7	325	152	173	99	65.1	10	6.6	19	12.5	2	1.3	22	14.5
8	197	103	94	57	55.3	7	6.8	6	5.8	3	2.9	30	29.1

5.3. Classical Item Analyses

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

5.3.1. 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. For items in which a representative samples of students is obtained, a high *p*-value indicates that an item is easy (high proportion of students answered it correctly), whereas a low *p*-value indicates that an item is hard. For example, if an item has a *p*-value of 0.79, it means 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 student who responded to the item) divided by the number of possible score points on the item.

Table 5.5 and Table 5.6 present the summary statistics for the *p*-values across all operational and field test items, respectively, on the NSCAS Summative assessments, including the number of items by *p*-value range (i.e., less than or equal to a *p*-value of 0.1, 0.2, etc.). These data were calculated for items with and without a representative sample. Items without a representative sample are those administered during the adaptive stage of the assessment, and for these items, the expectation for a *p*-value is typically between 0.4 and 0.6. Appendix G provides the summary *p*-value statistics by item type. Typically, test developers target *p*-values in the range of 0.3 to 0.8. The average *p*-values range for the NSCAS assessments range from 0.4 to 0.7 across content areas and grades.

Table 5.5. 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	223	0.436	0.187	0.000	1.000	9	6	27	53	63	40	13	2	1	9
4	381	0.459	0.199	0.000	1.000	16	10	34	98	93	60	33	11	11	15
5	216	0.482	0.191	0.000	1.000	9	4	12	38	59	41	28	17	1	7
6	263	0.452	0.217	0.000	1.000	21	8	20	51	65	46	26	11	5	10
7	151	0.505	0.156	0.000	1.000	1	3	9	22	44	37	18	12	4	1
8	200	0.503	0.219	0.000	1.000	6	10	13	30	54	32	21	14	7	13
Mathematics															
3	230	0.507	0.122	0.200	0.979	–	1	6	29	84	68	22	16	3	1
4	130	0.530	0.155	0.126	0.893	–	2	2	21	39	25	20	13	8	–
5	189	0.541	0.163	0.118	1.000	–	4	6	23	48	50	25	18	12	3
6	275	0.494	0.138	0.128	0.896	–	2	13	54	88	58	32	22	6	–
7	180	0.502	0.132	0.181	0.917	–	2	6	24	78	33	21	11	4	1
8	117	0.523	0.149	0.106	0.900	–	2	4	17	30	31	17	12	4	–

						#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**
Science															
5	50	0.660	0.124	0.367	0.879	–	–	–	1	5	9	12	18	5	–
8	60	0.642	0.120	0.382	0.839	–	–	–	1	6	17	18	11	7	–

Table 5.6. 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	111	0.471	0.177	0.006	0.910	2	5	14	21	21	20	18	6	3	1
4	102	0.485	0.168	0.076	0.850	1	4	11	13	21	27	13	10	2	–
5	105	0.490	0.196	0.030	0.873	4	7	6	14	22	18	17	14	3	–
6	97	0.536	0.182	0.053	0.861	1	5	4	11	17	21	22	11	5	–
7	102	0.552	0.195	0.078	0.904	1	2	9	10	21	19	13	13	13	1
8	102	0.482	0.196	0.051	0.960	1	9	11	16	15	15	21	12	1	1
Mathematics															
3	87	0.570	0.207	0.107	0.949	–	4	6	13	7	14	10	20	11	2
4	94	0.521	0.204	0.100	0.899	1	4	11	16	11	15	19	7	10	–
5	95	0.563	0.202	0.088	0.957	2	1	4	16	11	21	15	12	8	5
6	93	0.450	0.202	0.020	0.914	1	6	16	18	19	13	6	6	7	1
7	91	0.367	0.174	0.045	0.769	8	10	14	20	18	12	7	2	–	–
8	86	0.468	0.179	0.076	0.887	2	3	9	21	17	11	13	7	3	–

5.3.2. 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 their 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. There is, however, an interaction between item discrimination and difficulty: A very difficult item (or a very easy item) would have little variance in student responses, meaning most students respond correctly (or incorrectly). The resulting item-total correlation is typically low since both groups have the same score.

Table 5.7 and Table 5.8 present the summary statistics for the item-total correlations across all operational and field items, respectively. Appendix H provides the summary item-total correlation statistics by item type. Instead of using the number-correct score, the estimated final theta score was used to compute item-total correlations for the NSCAS tests because number-correct scores would not provide much insight into student performance on an adaptive test (i.e., in theory all students get 50% correct on a CAT). The results appear out-of-bounds from traditional metrics, but this is because the 2018 NSCAS ELA and Mathematics tests were adaptive. Due to adaptive selection of items, some items were administered to small number of students. The relatively higher number of ELA items in the ≤ .2 range are mostly obtained from n-counts less than 100 (and were therefore not calibrated in 2018). Therefore, the means of the correlations are reasonable and the number of items with less than .2 are relatively small.

Table 5.7. 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	223	0.289	0.185	-0.438	1.000	31	15	60	74	26	12	5
4	381	0.315	0.191	-0.687	1.000	43	30	91	125	57	16	19
5	216	0.321	0.225	-1.000	1.000	23	10	48	64	47	15	9
6	263	0.322	0.199	-0.442	1.000	39	13	46	83	49	22	11
7	151	0.352	0.132	-0.059	0.770	7	12	31	43	40	17	1
8	200	0.297	0.180	-0.945	0.659	28	15	44	60	40	11	2
Mathematics												
3	230	0.387	0.100	0.122	1.000	–	7	21	118	62	18	4
4	130	0.374	0.099	0.080	0.622	1	4	20	58	33	12	2
5	189	0.369	0.104	-0.285	0.666	4	3	23	92	53	12	2
6	275	0.358	0.078	0.099	0.631	1	3	53	149	58	9	2
7	180	0.369	0.089	-0.044	0.614	2	2	25	93	45	11	2
8	117	0.375	0.113	-0.135	0.644	2	1	17	48	38	8	3
Science												
5	50	0.388	0.078	0.209	0.556	–	–	7	19	21	3	–
8	60	0.401	0.065	0.206	0.545	–	–	3	25	31	1	–

Table 5.8. 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	111	0.363	0.131	0.067	0.585	6	9	15	31	35	15	–
4	102	0.346	0.130	-0.097	0.575	4	8	17	41	22	10	–
5	105	0.312	0.135	-0.097	0.523	11	10	22	29	29	4	–
6	97	0.353	0.117	0.046	0.579	1	11	15	31	32	7	–
7	102	0.330	0.118	0.019	0.553	5	13	20	35	21	8	–
8	102	0.324	0.130	-0.010	0.557	6	12	24	28	25	7	–
Mathematics												
3	87	0.395	0.121	-0.036	0.603	3	4	6	26	30	17	1
4	94	0.420	0.120	-0.079	0.619	1	2	14	17	34	24	2
5	95	0.411	0.123	0.059	0.633	1	6	13	22	28	21	4
6	93	0.407	0.123	0.025	0.665	3	4	6	25	35	17	3
7	91	0.412	0.129	-0.027	0.634	2	4	7	28	22	25	3
8	86	0.427	0.109	0.006	0.623	1	2	8	17	37	19	2

5.3.3. Item Suppression

Based on the flagging criteria presented in Table 5.9 and Table 5.10 for multiple-choice and partial-credit (i.e., non-multiple-choice) items, respectively. Out of a total of 2,499 multiple-choice ELA and Mathematics administered items, 685 of them were flagged (i.e., 27% of administered items). Of those 685 flagged items, 305 of them were removed based on n-counts less than 100, leaving a total of 380 flagged items (15% of administered items). Eighteen partial-credit items were flagged out of a total of 166 items (i.e., 11% of items).

Of these flagged items, the 16 ELA items indicated in Table 5.11 were suppressed per the recommendation of NWEA content specialists. There was no suppression for Mathematics or Science. The ELA suppressed items were not included for all subsequent analyses and score reporting.

Table 5.9. Flagging Criteria for MC Items

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

*item-total = item-total correlation

Table 5.10. Flagging Criteria for Partial-Credit Items

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

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

Table 5.11. Suppressed ELA Items

Grade	Item Code	Item Role*	Item Type	Max. #Points
ELA				
3	21058840	OP	Multiple-choice	1
3	21096460	OP	Multiselect	2
4	21057660	OP	Multiple-choice	1
4	21072010	OP	Multiple-choice	1
4	21078770	OP	Multiple-choice	1
4	21107610*	VL	Composite	2
5	21074300	OP	Multiple-choice	1

Grade	Item Code	Item Role*	Item Type	Max. #Points
5	21107610*	HL	Composite	2
6	11219110	OP	Multiple-choice	1
6	21108210	OP	Composite	2
7	21077650	OP	Multiple-choice	1
7	31193000	OP	Multiple-choice	1
7	21107910	OP	Composite	2
8	11191220	OP	Multiple-choice	1
8	21048290	OP	Multiple-choice	1
8	21073270	OP	Multiple-choice	1

*Item 21107610 was also on the paper-pencil assessment (English version, Grade 5, Item 41). This item was also translated into Spanish (Item 21108380). Thus, Item 21108380 was also suppressed.

5.4. 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. When 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 are often in a good position to identify a cause of this type. DIF may also result from differential instruction closely associated with group membership.

5.4.1. DIF Methods

The Mantel-Haenszel (MH) procedure was used to detect DIF for dichotomous items (Holland & Thayer, 1988), and the standardized mean difference (SMD) analysis, developed as an extension of the MH procedure, was used to detect DIF for polytomous items (Dorans & Schmitt, 1991; Zwick, Donoghue, & Grima, 1993).

The MH method has been widely used in educational measurement due to its easy implementation in testing programs. The procedure compares the ratio of the probabilities of two groups of students (i.e., focal and reference groups) answering an item correctly across all score levels. The obtained estimate is known as the odds ratio, which is computed as follows:

$$\alpha_{MH} = \frac{\left(\sum_m \frac{R_{rm} W_{fm}}{N_m}\right)}{\left(\sum_m \frac{R_{fm} W_{rm}}{N_m}\right)} \tag{5.1}$$

where:

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

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

$$MH\ D - DIF = -2.35 \ln(\alpha_{MH}) \tag{5.2}$$

The MH chi-square statistic used to classify items into the three Educational Testing Service (ETS) DIF categories is as follows:

$$MH\ CHISQ = \frac{\left(\mid \sum_m R_{rm} - \sum_m E(R_{rm}) \mid - \frac{1}{2}\right)^2}{\sum_m Var(R_{rm})} \tag{5.3}$$

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} and are the number of students who answered the item correctly and incorrectly, respectively.

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

$$SMD = \sum p_{FK} m_{FK} - \sum p_{FK} m_{RK} \tag{5.4}$$

where:

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

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

5.4.2. Focal and Reference Groups

The focal groups for the NSCAS DIF analysis were female for gender-based DIF and minority groups for ethnicity-based DIF. The reference groups were male and white, respectively, as presented in Table 5.12. DIF was not conducted if the sample size for either the reference group or the focal group was less than 250.

Table 5.12. 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

5.4.3. DIF Categories A, B, and C

Table 5.13 and Table 5.14 present the 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 5.13. DIF Categories for Dichotomous Items

DIF Category	Level of DIF	Definition
A	Negligible	• Absolute value of the Mantel-Haenszel delta difference (MH D-DIF) is not significantly different from 0 or is less than one.
B	Moderate	• Absolute value of the MH D-DIF is significantly different from 0 but not from one, and is at least 1; or • Absolute value of the MH D-DIF is significantly different from 1, but less than 1.5. • Positive values are classified as “B+” and negative values as “B-“.
C	Strong	• Absolute value of the MH D-DIF is significantly different from 1, and is at least 1.5; and • Absolute value of the MH D-DIF is larger than 1.96 times the standard error of MH D-DIF. • Positive values are classified as “C+” and negative values are “C-“.

Table 5.14. DIF Categories for Polytomous Items

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

5.4.4. DIF Results

Table 5.15 and Table 5.16 present the number of items assigned to each DIF category for operational and field test items, respectively, for the female, black or African American, Hispanic, Asian, and two or more races focal groups. Male was the reference groups for gender, and white was the reference group for ethnicity. Field test items only included black and Hispanic ethnic groups in the data. Appendix I presents the item-level DIF statistics. The + sign next to the DIF category indicates that the item is in favor of the reference group, and the - sign indicates that the item is in favor of the focal group. As shown in the tables, most items were categorized as DIF Category A with negligible DIF.

Table 5.15. DIF Results—Operational Items

Grade	Focal Group	#Items by DIF Category					
		Total	A	B+	B-	C+	C-
ELA							
3	Female	122	119	1	2	–	–
	Black or African American	63	63	–	–	–	–
	Hispanic	107	104	2	1	–	–
	Asian	38	35	1	2	–	–
	Two or More Races	47	47	–	–	–	–
4	Female	186	178	5	3	–	–
	Black or African American	64	62	–	2	–	–
	Hispanic	146	138	1	4	–	3
	Asian	26	24	2	–	–	–
	Two or More Races	48	48	–	–	–	–
5	Female	132	131	–	1	–	–
	Black or African American	73	72	–	1	–	–
	Hispanic	92	91	–	1	–	–
	Asian	45	42	2	1	–	–
	Two or More Races	52	52	–	–	–	–
6	Female	127	120	3	2	–	2
	Black or African American	68	67	–	1	–	–
	Hispanic	85	81	–	4	–	–
	Asian	43	40	1	1	–	1
	Two or More Races	46	46	–	–	–	–
7	Female	116	113	2	1	–	–
	Black or African American	65	65	–	–	–	–
	Hispanic	109	108	1	–	–	–
	Asian	47	41	2	3	–	1
	Two or More Races	53	52	1	–	–	–
8	Female	107	103	–	3	–	1
	Black or African American	56	53	1	2	–	–
	Hispanic	80	77	–	–	–	3
	Asian	37	36	–	–	–	1
	Two or More Races	40	40	–	–	–	–

		#Items by DIF Category					
Grade	Focal Group	Total	A	B+	B-	C+	C-
Mathematics							
3	Female	151	131	6	11	–	3
	Black or African American	53	51	2	–	–	–
	Hispanic	103	99	1	3	–	–
	Asian	38	34	2	2	–	–
	Two or More Races	43	43	–	–	–	–
4	Female	121	116	–	1	–	4
	Black or African American	58	56	1	1	–	–
	Hispanic	87	86	1	–	–	–
	Asian	45	40	1	2	2	–
	Two or More Races	47	47	–	–	–	–
5	Female	147	127	2	15	–	3
	Black or African American	60	53	1	6	–	–
	Hispanic	103	101	–	2	–	–
	Asian	45	39	–	4	1	1
	Two or More Races	48	48	–	–	–	–
6	Female	193	177	9	4	–	3
	Black or African American	63	60	3	–	–	–
	Hispanic	129	125	2	2	–	–
	Asian	40	37	1	–	–	2
	Two or More Races	39	39	–	–	–	–
7	Female	164	151	8	3	–	2
	Black or African American	57	55	–	2	–	–
	Hispanic	98	97	1	–	–	–
	Asian	46	41	2	1	–	2
	Two or More Races	44	44	–	–	–	–
8	Female	94	86	1	6	–	1
	Black or African American	53	53	–	–	–	–
	Hispanic	86	82	–	3	–	1
	Asian	42	37	3	1	–	1
	Two or More Races	39	39	–	–	–	–
Science							
5	Female	50	47	–	3	–	–
	Black or African American	50	50	–	–	–	–
	Hispanic	50	49	–	1	–	–
	Asian	50	44	3	2	1	–
	Two or More Races	50	49	–	1	–	–
8	Female	60	57	2	1	–	–
	Black or African American	60	60	–	–	–	–
	Hispanic	60	60	–	–	–	–
	Asian	60	54	2	4	–	–
	Two or More Races	60	60	–	–	–	–

Table 5.16. DIF Results—Field Test Items

		#Items by DIF Category					
Grade	Focal Group	Total	A	B+	B-	C+	C-
ELA							
3	Female	111	104	3	4	–	–
	Black or African American	–	–	–	–	–	–
	Hispanic	41	36	–	4	–	1
4	Female	102	96	3	1	1	1
	Black or African American	–	–	–	–	–	–
	Hispanic	19	17	–	2	–	–
5	Female	105	97	1	5	–	2
	Black or African American	–	–	–	–	–	–
	Hispanic	11	11	–	–	–	–
6	Female	97	84	10	2	1	–
	Black or African American	6	6	–	–	–	–
	Hispanic	23	21	–	1	–	1
7	Female	102	96	3	–	3	–
	Black or African American	2	2	–	–	–	–
	Hispanic	14	12	–	2	–	–
8	Female	102	97	1	1	2	1
	Black or African American	–	–	–	–	–	–
	Hispanic	16	15	–	1	–	–
Mathematics							
3	Female	87	72	6	5	2	2
	Black or African American	–	–	–	–	–	–
	Hispanic	70	60	3	7	–	–
4	Female	94	87	4	3	–	–
	Black or African American	1	1	–	–	–	–
	Hispanic	33	32	–	1	–	–
5	Female	95	82	6	5	–	2
	Black or African American	–	–	–	–	–	–
	Hispanic	22	18	1	2	–	1
6	Female	93	82	1	7	1	2
	Black or African American	–	–	–	–	–	–
	Hispanic	29	24	2	1	–	2
7	Female	91	85	4	2	–	–
	Black or African American	–	–	–	–	–	–
	Hispanic	11	9	1	1	–	–
8	Female	86	82	3	1	–	–
	Black or African American	–	–	–	–	–	–
	Hispanic	21	21	–	–	–	–

5.5. IRT Calibration

Unidimensional item response theory (IRT) models were used to calibrate items and create the NSCAS Summative scale: the Rasch model (Rasch, 1960, 1980; Wright, 1977) for dichotomous items and the partial credit model (PCM; Masters, 1982) for polytomous items. For all content areas, item parameter estimations were implemented using WINSTEPS 3.91.0.0 (Linacre, 2015) that used joint maximum likelihood estimation (MLE) as described by Wright and Masters (1982).

The Rasch model has had a long-standing presence in applied testing programs and was the methodology used to calibrate the previous Nebraska State Accountability (NeSA) items. Under the Rasch model, the probability of a student with ability θ responding correctly to item i is:

$$P\left(u_{ij} = 1 \,\middle|\, \theta_j, b_i\right) = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}} \tag{5.5}$$

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

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

$$P\left(u_{ij} = k \middle| \theta_i\right) = \frac{e^{\left[\sum_{u=1}^{k} Da_i\left(\theta_j - b_i + d_{iu}\right)\right]}}{\sum_{v=1}^{m_i} e^{\left[\sum_{u=1}^{k} Da_i\left(\theta_j - b_i + d_{iu}\right)\right]}} \tag{5.6}$$

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

5.5.1. Checking Model Assumptions

Since the scales for the NSCAS Summative assessments were established with the IRT Rasch model for dichotomous items and the PCM for polytomous items, it is important to check the three fundamental model assumptions such as unidimensionality of the latent trait and local independence. This section evaluates the unidimensionality of the data, local independence, and item fit using the 2018 operational items. Overall, the principal component analysis (PCA) of residuals indicates one dominant dimension for all contents and grades. The median residual correlations are close to 0 and the small number of items with correlations greater than 0.20, suggesting that local item independence generally holds for all content areas and grades. A small number of items outside of 0.7 and 1.3 in terms of infit mean square statistics indicates a good fit.

5.5.1.1. Unidimensionality

Unidimensionality is the most commonly violated assumptions in the latent trait structure implied by the item response data. In most instances, it is sufficient to assume that all items in a test are sensitive to differences in examinees along a single latent trait. However, it is crucial to check if only one dominant exists among the items. PCA is the most commonly used statistical procedure to check how many dimensions exist in the data.

Table 5.17 presents the PCA of residuals results. For ELA, one dominant dimension explained most of the variance from 20.8 to 26.7 percent. Grade 5 and 8 only have four dimensions with an eigen value bigger than 1. A similar pattern was observed for Mathematics. Mathematics

Grade 7 only has two dimensions with an eigenvalue bigger than 1. Mathematics Grade 8 has only three dimensions with an eigenvalue bigger than 1. Science has 4 and 5 dimensions with an eigenvalue bigger than1 for Grades 5 and 8, respectively.

Table 5.17. Results from PCA of Residuals

Grade	Components	Eigenvalue	Explained Variance
ELA			
3	Measure	44.0	23.7
	1	1.5	0.8
	2	1.4	0.7
	3	1.4	0.7
	4	1.3	0.7
	5	1.3	0.7
4	Measure	61.0	20.8
	1	1.6	0.6
	2	1.4	0.5
	3	1.4	0.5
	4	1.4	0.5
	5	1.4	0.5
5	Measure	36.6	22.5
	1	1.4	0.9
	2	1.3	0.8
	3	1.2	0.8
	4	1.2	0.7
	5	–	–
6	Measure	43.4	24.4
	1	1.7	0.9
	2	1.4	0.8
	3	1.3	0.8
	4	1.3	0.7
	5	1.2	0.7
7	Measure	36.2	26.2
	1	1.5	1.1
	2	1.4	1.0
	3	1.3	0.9
	4	1.3	0.9
	5	1.3	0.9
8	Measure	43.4	26.7
	1	1.6	1.0
	2	1.4	0.8
	3	1.3	0.8
	4	1.2	0.7
	5	–	–
Mathematics			
3	Measure	86.6	31
	1	1.7	0.6
	2	1.6	0.6
	3	1.5	0.5
	4	1.4	0.5
	5	1.4	0.5

Grade	Components	Eigenvalue	Explained Variance
4	Measure	39.4	28.7
	1	1.9	1.3
	2	1.6	1.2
	3	1.5	1.1
	4	1.5	1.1
	5	1.4	1.0
5	Measure	55.6	27.6
	1	1.9	0.9
	2	1.8	0.9
	3	1.6	0.8
	4	1.5	0.8
	5	1.5	0.7
6	Measure	90.6	27.5
	1	1.7	0.5
	2	1.6	0.5
	3	1.6	0.5
	4	1.5	0.5
	5	–	–
7	Measure	58.8	28.2
	1	1.7	0.8
	2	1.5	0.7
	3	–	–
	4	–	–
	5	–	–
8	Measure	42.7	32.2
	1	1.8	1.4
	2	1.5	1.1
	3	1.3	1.0
	4	–	–
	5	–	–
Science			
5	Measure	13.6	21.4
	1	1.9	2.9
	2	1.4	2.2
	3	1.3	2.1
	4	1.2	1.9
	5	–	–
8	Measure	17.3	22.4
	1	1.7	2.1
	2	1.4	1.8
	3	1.3	1.7
	4	1.2	1.6
	5	1.2	1.6

5.5.1.2. Local Independence

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

- Raw
- Standardized
- Logit

The raw score residual correlation corresponds to Yen's *Q*3 index, a popular local independence statistic. The expected value for the *Q3* statistic is approximately −1/(*k*−1) when no local dependence exists, where *k* is test length (Yen, 1993). Thus, the expected *Q*3 values should be approximately −0.02 for the NSCAS tests (since most NSCAS tests had more than 50 operational items). Index values greater than 0.20 indicate a degree of local dependence that should be examined by test developers (Chen & Thissen, 1997).

Since the three residual correlations are very similar, the default "standardized residual correlation" in WINSTEPS was used for these analyses. Table 5.18 presents the summary statistics for all the residual correlations for each test, including the median, interquartile range (IQR), minimum, maximum, and several percentiles (10, 25, 50, 75, and 90). The table also presents the total number of item pairs and the number of pairs with the residual correlations greater than 0.2. The median residual correlations were slightly negative and the values were close to 0.0. Most of the correlations were very small, suggesting local item independence generally holds for NSCAS ELA, Mathematics, and Science.

Table 5.18. Summary of Item Residual Correlations

	#Item Pairs					Percentiles					
Grade	**Total**	**> 0.2**	**Median**	**IQR**	**Min.**	**P10**	**P25**	**P50**	**P75**	**P90**	**Max.**
ELA											
3	10,011	0	0.00	0.01	-0.19	-0.02	-0.01	0.00	0.00	0.00	0.15
4	26,796	0	0.00	0.00	-0.20	-0.01	0.00	0.00	0.00	0.00	0.14
5	7,875	0	0.00	0.01	-0.13	-0.02	-0.01	0.00	0.00	0.00	0.08
6	8,911	0	0.00	0.01	-0.16	-0.02	-0.01	0.00	0.00	0.00	0.14
7	5,151	0	0.00	0.01	-0.14	-0.03	-0.01	0.00	0.00	0.00	0.11
8	7,021	0	0.00	0.01	-0.16	-0.02	-0.01	0.00	0.00	0.00	0.08
Mathematics											
3	18,528	2	0.00	0.00	-0.19	-0.02	0.00	0.00	0.00	0.00	0.41
4	4,753	3	0.00	0.02	-0.10	-0.03	-0.02	0.00	0.00	0.01	0.36
5	10,585	7	0.00	0.01	-0.19	-0.02	-0.01	0.00	0.00	0.00	0.42
6	28,441	4	0.00	0.00	-0.16	-0.01	0.00	0.00	0.00	0.00	0.40
7	11,175	2	0.00	0.01	-0.10	-0.02	-0.01	0.00	0.00	0.00	0.31
8	4,005	1	0.00	0.02	-0.13	-0.03	-0.02	0.00	0.00	0.00	0.23

	#Item Pairs					Percentiles					
Grade	Total	> 0.2	Median	IQR	Min.	P10	P25	P50	P75	P90	Max.
Science											
5	1,225	1	-0.02	0.02	-0.08	-0.04	-0.03	-0.02	-0.01	0.01	0.43
8	1,770	1	-0.02	0.02	-0.07	-0.04	-0.03	-0.02	-0.01	0.01	0.23

5.5.1.3. Item Fit

Item fit refers to how well the data fit the calibration model. WINSTEPS two item fit statistics for evaluating the degree to which the Rasch model predicts the observed item responses for the NSCAS tests: infit and outfit. Each fit statistic can be expressed as a mean square (MNSQ) statistic with each statistic having an expected value of 1 and a different variance for each mean square or as a standardized statistic (ZSTD with an expected mean = 0 and expected variance = 1). Table 5.19 presents the summary MNSQ statistics, and Table 5.20 presents the summary ZSTD statistics. Overall, these results show that the data fit the model well.

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

The expected MNSQ value is 1.0 and can range from 0 to positive infinity. Values greater than 1.0 can be interpreted as indicating the presence of noise or lack of fit between the responses and the measurement model. Values less than 1.0 can be interpreted as item consistency or overfitting (i.e., too predictable and/or too much redundancy). Rules of thumb regarding "practically significant" MNSQ values vary from author to author. More conservative users might prefer items with MNSQ values that range from 0.8 to 1.2. Others believe reasonable test results can be achieved with values from 0.5 to 1.5. In Table 5.19, values outside of 0.7 to 1.3 are given practical importance.

The expected ZSTD mean value is 0.0 with an expected variance, or SD, of 1.0. It can effectively range from -9.99 to +9.99 in WINSTEPS. Values greater than 0.0 can be interpreted as indicating the presence of noise or lack of fit between the items and the model (underfitting). Values less than 0.0 can be interpreted as item redundancy or overfitting items (i.e., too predictable and/or too much redundancy). Rules of thumb regarding "practically significant" ZSTD values vary from author to author. More conservative users might prefer items with ZSTD values that range from −2 to +2. Others believe reasonable test results can be achieved with values from −3 to +3. In Table 5.20, values outside of −3 to +3 are given practical importance.

Table 5.19. Summary of Infit and Outfit MNSQ Statistics for Items

		Infit						Outfit					
Grade	#Items	Mean	SD	Min.	Max.	# [0.7,1.3]	% [0.7,1.3]	Mean	SD	Min.	Max.	# [0.7,1.3]	% [0.7,1.3]
ELA													
3	142	0.97	0.06	0.86	1.37	1	0.7	0.97	0.09	0.66	1.53	4	2.8
4	232	0.97	0.08	0.76	1.62	3	1.3	0.97	0.13	0.05	2.22	4	1.7
5	126	0.99	0.07	0.87	1.30	0	0.0	0.99	0.10	0.76	1.39	2	1.6
6	134	0.98	0.08	0.79	1.43	1	0.7	0.98	0.15	0.71	2.23	2	1.5
7	102	0.99	0.08	0.80	1.30	0	0.0	1.01	0.23	0.58	2.97	6	5.9
8	119	0.97	0.07	0.82	1.36	1	0.8	0.96	0.12	0.56	1.66	4	3.4
Mathematics													
3	193	0.96	0.05	0.82	1.29	0	0.0	0.96	0.11	0.67	1.73	6	3.1
4	98	0.98	0.08	0.75	1.39	1	1.0	0.98	0.13	0.60	1.45	5	5.1
5	146	0.97	0.07	0.79	1.31	1	0.7	0.97	0.13	0.63	1.71	6	4.1
6	239	0.96	0.06	0.81	1.55	1	0.4	0.96	0.11	0.59	2.07	5	2.1
7	150	0.97	0.08	0.84	1.44	1	0.7	0.97	0.11	0.78	1.48	4	2.7
8	90	0.98	0.07	0.85	1.32	1	1.1	1.01	0.42	0.69	4.79	5	5.6
Science													
5	50	1.00	0.09	0.85	1.22	0	0.0	0.98	0.16	0.64	1.34	3	6.0
8	60	1.00	0.08	0.86	1.18	0	0.0	0.99	0.14	0.75	1.43	1	1.7

Table 5.20. Summary of Infit and Outfit ZSTD Statistics for Items

		Infit						Outfit					
Grade	#Items	Mean	SD	Min.	Max.	# [-3.0,3.0]	% [-3.0,3.0]	Mean	SD	Min.	Max.	# [-3.0,3.0]	% [-3.0,3.0]
ELA													
3	142	-2.45	4.93	-9.9	9.9	75	52.8	-2.46	4.87	-9.9	9.9	79	55.6
4	232	-2.03	4.36	-9.9	9.9	105	45.3	-2.03	4.33	-9.9	9.9	103	44.4
5	126	-0.99	5.33	-9.9	9.9	69	54.8	-0.82	5.38	-9.9	9.9	69	54.8
6	134	-2.00	5.58	-9.9	9.9	81	60.4	-1.93	5.62	-9.9	9.9	84	62.7
7	102	-1.22	5.92	-9.9	9.9	64	62.7	-0.87	6.22	-9.9	9.9	66	64.7
8	119	-2.68	5.17	-9.9	9.9	67	56.3	-2.80	5.24	-9.9	9.9	72	60.5
Mathematics													
3	193	-3.03	4.31	-9.9	9.9	122	63.2	-2.77	4.45	-9.9	9.9	117	60.6
4	98	-2.33	6.20	-9.9	9.9	77	78.6	-2.37	6.08	-9.9	9.9	77	78.6
5	146	-2.97	4.85	-9.9	9.9	92	63.0	-2.65	5.00	-9.9	9.9	97	66.4
6	239	-2.79	4.12	-9.9	9.9	132	55.2	-2.75	4.17	-9.9	9.9	140	58.6
7	150	-3.71	5.17	-9.9	9.9	111	74.0	-3.25	5.18	-9.9	9.9	113	75.3
8	90	-2.26	5.56	-9.9	9.9	61	67.8	-2.26	5.57	-9.9	9.9	57	63.3
Science													
5	50	-0.07	8.22	-9.9	9.9	44	88.0	-0.27	8.26	-9.9	9.9	44	88.0
8	60	-0.90	7.77	-9.9	9.9	47	78.3	-0.69	7.77	-9.9	9.9	45	75.0

5.5.2. Summary IRT Item Statistics

Table 5.21 and Table 5.22 present the summary IRT item statistics across all operational and field test items, respectively. Appendix J presents the item-level IRT item statistics. Operational item parameter means increase with the grade for ELA and Mathematics, as can be expected for vertical scales.

Table 5.21. Summary IRT Item Statistics—Operational Items

Content Area	Grade	#Items	#Parameters	Mean	SD	Min.	Max.	Range (Max. – Min.)
ELA	3	142	153	-0.490	1.040	-2.601	3.431	6.032
	4	232	252	-0.134	0.899	-3.089	3.151	6.239
	5	126	145	0.311	0.991	-2.474	2.887	5.361
	6	134	151	0.336	0.905	-1.477	3.345	4.821
	7	102	115	0.396	0.970	-2.343	4.115	6.458
	8	119	139	0.537	1.099	-1.947	5.255	7.203
Mathematics	3	193	197	-1.040	0.998	-4.758	1.459	6.217
	4	98	102	-0.069	0.958	-2.172	2.887	5.059
	5	146	150	-0.086	0.933	-2.706	2.559	5.264
	6	239	244	0.054	0.990	-2.282	2.579	4.861
	7	150	156	0.755	0.943	-1.550	3.993	5.543
	8	90	94	0.787	0.914	-1.458	4.021	5.479
Science	5	50	50	-0.820	0.712	-2.146	0.609	2.755
	8	60	60	-0.703	0.698	-2.141	0.528	2.669

Table 5.22. Summary IRT Item Statistics—Field Test Items

Content Area	Grade	#Items	#Parameters	Mean	SD	Min.	Max.	Range (Max. – Min.)
ELA	3	111	137	-0.085	1.091	-2.900	5.071	7.971
	4	102	131	0.165	1.186	-3.230	2.996	6.226
	5	105	127	0.451	1.211	-2.981	4.268	7.249
	6	97	123	0.446	1.050	-1.752	3.736	5.488
	7	102	116	0.415	1.088	-1.911	3.542	5.453
	8	102	118	0.938	1.135	-2.515	4.027	6.542
Mathematics	3	87	99	-0.567	1.304	-3.820	2.589	6.409
	4	94	114	0.422	1.246	-2.229	3.908	6.137
	5	95	117	0.414	1.265	-2.822	3.695	6.517
	6	93	121	1.327	1.356	-1.965	5.740	7.704
	7	91	116	1.799	1.160	-0.531	4.755	5.286
	8	86	99	1.417	1.191	-1.396	4.719	6.115

5.6. Equating and Scaling

This section provides evidence to support the claim that test scaling and linking is accurate at the baseline and across time to support that score interpretations are constant across time and that the process supports interpretations of growth.

5.6.1. Vertical Scaling (ELA and Mathematics)

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

5.6.1.1. Linking Design

Following the 2018 test administration, the vertical scales were created based on the following decisions:

- Data collection design: Common item design
- Selection of the vertical scaling items: Above grade and below grade
- Scaling method: IRT Rasch and partial-credit models
- Calibration method: Concurrent calibration across grades
- Theta estimators and software: MLE in WINSTEPS
- Score transformation: Four-digit scales scores without anchoring cut scores

Figure 5.1 presents an in-depth illustration of the linking design from a psychometric perspective. Descriptions of the blocks are as follows:

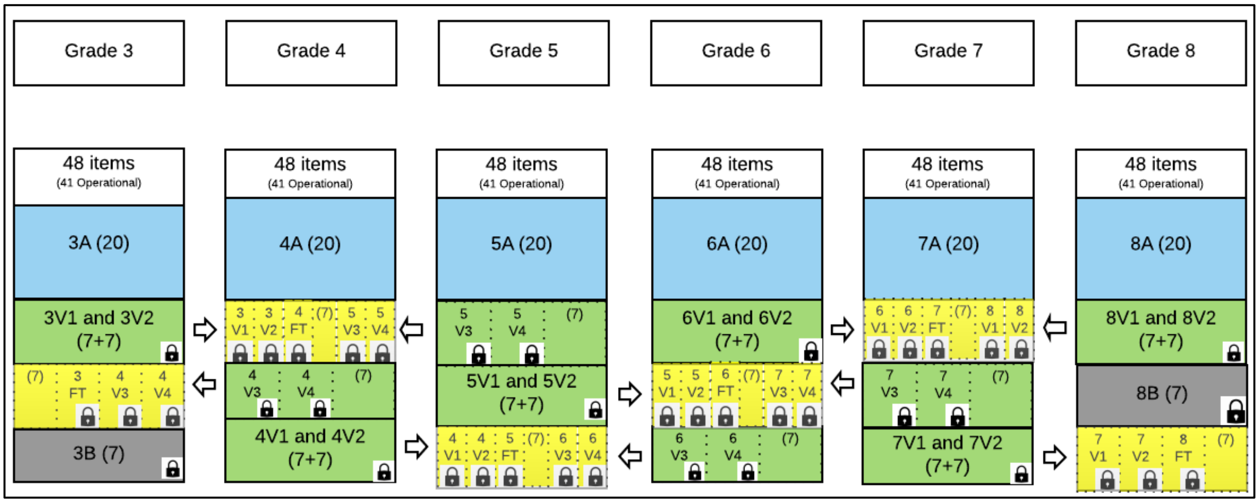
- The first number indicates the grade.
- The character in the second location represents item role:
    - A = adaptive selection
    - B = non-adaptive selection used for horizontal linking of Grades 3 and 8
    - V = vertical linking anchors where V1 and V2 sets are embedded into the grade above and V3 and V4 sets are embedded into the grade below.
- The number in parenthesis indicates the number of items for each block.
- The arrows represent item movement (e.g., the arrow from Grade 5 to Grade 6 means the Grade 5 horizontal linking items are embedded into Grade 6 as vertical linking items).
- Color representation:
    - Blue = on-grade items to be selected adaptively (20 items)
    - Gray = on-grade items pre-selected as horizontal linking items (7 items for Grades 3 and 8 only)
    - Green = on-grade items pre-selected as vertical linking items
    - Yellow = non-operational items that can be either vertical linking or field test items

The figure shows the following:

- There is a total of 28 vertical linking anchors across two adjacent grades.
- Half of the 28 vertical linking anchors are from the grade above and the other half are from the grade below.
- Four sets of 28 vertical linking anchors were placed on the field test slots. The design was intended to have 1,250 student responses for each vertical linking anchor and 750 student responses for each field test item.
- All vertical linking items selected also served as horizontal linking items.
- For Grades 3 and 8, one set of non-adaptively selected 21 items was assembled. For example, all Grade 3 students saw an item set of 3B+3V1+3V2 (two item blocks in green and gray).
- For Grades 4–7, one of two sets of non-adaptively selected 21 items was assembled. For example, half of the Grade 4 students saw 21 horizontal linking anchors of 4V1+4V2+4V3, while the other half saw 21 horizontal linking anchors of 4V1+4V2+4V4. That is, 14 items of 4V1+4V2 are common for all Grade 4 students and the other 14 items, 4V3 or 4V4, are common for half of Grade 4 students.

All student saw a total of 48 items (41 operational + 7 non-operational). Twenty of these items were selected adaptively based on student ability level, and 21 were non-adaptively pre-selected horizontal and vertical linking anchors. Each student also saw one set of seven vertical linking or field test items.

Figure 5.1. Horizontal and Vertical Linking Design for ELA and Mathematics



The first 21 operational items students saw were the anchor item set. The 22nd operational item was adaptively selected based on student responses to operational items 1–20; the 23rd item was adaptively selected based on the previous 1–21 operational items; etc. The “n-1” approach was applied, where the (n+1)th item was selected based on (n-1) items so that item selection and rendering could be quick. Off-grade vertical linking anchors or field test items were administered as non-operational items. These seven items were grouped into two or three mini-blocks and approximately located after the 10th, 20th, and 30th operational items.

5.6.1.2. Linking Item Selection

Anchor items were selected as a subset of the 2018 paper-pencil forms. The TOS was used as the reference for anchor items so that the percentages were within 10% difference for each reporting category. Anchor items were included as horizontal linking items for both the paper-pencil and online assessments and as vertical linking items in the lower/upper grades for the CAT (e.g., of the 28 horizontal anchors for ELA Grade 5, 14 of them were ELA Grade 4 vertical linking items and the other 14 were ELA Grade 6 vertical linking items). As for the statistical references, the 2017 paper-pencil forms were used to compare TCC and mean of the p-value, item-total correlation, and Rasch difficulty parameters.

5.6.1.3. Sampling

Vertical scales can function differentially if created from groups of test takers with different characteristics (Kolen, 2011). Therefore, psychometricians consider it important to use a representative sample of students from the target population to calibrate item parameters. NWEA configured the adaptive engine to select samples of students to match the proportions of students found in Nebraska's major demographic groups based on the January 2018 state roster from the NDE. Table 5.23 and
Table 5.24 present those proportions based on the January 2018 state roster.

However, the population exposure distribution for collecting a representative sample of students during the field test and vertical linking portion of the test administration was not functioning for a portion of the tests during eight days of the testing window. To ensure that the vertical linking items were properly calibrated with a representative sample of students, NWEA rebalanced the number of students who took the vertical linking forms using a stratified random sampling procedure for anchor sets that had a 5% or more difference in observed population exposures compared to the target population.

The following seven vertical anchor item sets were identified as being under-representative of the general population based on the 5% or more difference criterion:

- ELA Grade 6 Vertical Anchor Down Set B
- ELA Grade 8 Vertical Anchor Up Set A
- ELA Grade 8 Vertical Anchor Up Set B
- Mathematics Grade 3 Vertical Anchor Down Set A
- Mathematics Grade 3 Vertical Anchor Down Set B
- Mathematics Grade 7 Vertical Anchor Up Set A
- Mathematics Grade 8 Vertical Anchor Up Set B

For these seven vertical anchor item sets, a sample of 1,250 students was drawn to determine if the difference between the target and sampled students was more than 5%. For two sets of vertical anchor items (i.e., Mathematics Grade 3 Vertical Down Set B and Mathematics Grade 7 Vertical Up Set A), the number of students for those two vertical anchor item sets was decreased to 1,000. With these sample students, the demographic percentage difference between the target and sampled students were less 5% and mean theta score differences were within 10% of one standard deviation (1SD), as reported in Table 5.25 and Table 5.26. Results were reviewed and approved by an expert advisory committee comprising national experts in measurement recommended by the Nebraska Assessment team.

Table 5.23. Population Demographics based on the January Roster—Gender

		Gender			
	Total	Female		Male	
Grade	#Students	N	%	N	%
3	23,995	11,673	48.6	12,322	51.4
4	23,908	11,622	48.6	12,286	51.4
5	22,346	10,801	48.3	11,545	51.7
6	23,415	11,480	49.0	11,935	51.0
7	23,095	11,204	48.5	11,891	51.5
8	23,420	11,397	48.7	12,023	51.3

Table 5.24. Population Demographics based on the January Roster—Ethnicity

		Ethnicity*													
	Total	AI/AN		Asian		Black		Hispanic		NH/PI		White		Two or More	
Grade	#Students	N	%	N	%	N	%	N	%	N	%	N	%	N	%
3	23,995	308	1.3	642	2.7	1,606	6.7	4,695	19.6	38	0.2	15,646	65.2	1,060	4.4
4	23,908	291	1.2	671	2.8	1,672	7.0	4,586	19.2	34	0.1	15,643	65.4	1,011	4.2
5	22,346	307	1.4	588	2.6	1,520	6.8	4,242	19.0	38	0.2	14,742	66.0	909	4.1
6	23,415	298	1.3	631	2.7	1,641	7.0	4,492	19.2	30	0.1	15,428	65.9	895	3.8
7	23,095	287	1.2	608	2.6	1,570	6.8	4,279	18.5	44	0.2	15,436	66.8	871	3.8
8	23,420	300	1.3	632	2.7	1,487	6.3	4,330	18.5	29	0.1	15,833	67.6	809	3.5

*AI/AN = American Indian or Alaskan Native. NH/PI = Native Hawaiian or Pacific Islander. Black = Black or African American.

Table 5.25. Sample Demographics Comparison after Sampling

			%Target					%Sample					%Difference (Sample – Target)*				
Grade	VL	N	F	M	B	H	W	F	M	B	H	W	F	M	B	H	W
ELA																	
6	Vertical Anchor Down Set B	1,250	49.0	51.0	7.0	19.2	65.9	47.8	52.2	6.2	17.7	67.4	-1.2	1.2	-0.8	-1.5	1.5
8	Vertical Anchor Up Set A	1,250	48.7	51.3	6.3	18.5	67.6	48.7	51.3	6.2	18.2	67.4	0.0	0.0	-0.1	-0.3	-0.2
8	Vertical Anchor Up Set B	1,250	48.7	51.3	6.3	18.5	67.6	48.7	51.3	6.2	18.2	67.4	0.0	0.0	-0.1	-0.3	-0.2
Mathematics																	
3	Vertical Anchor Down Set A	1,250	48.6	51.4	6.7	19.6	65.2	48.8	51.2	6.6	19.3	64.8	0.2	-0.2	-0.1	-0.3	-0.4
3	Vertical Anchor Down Set B	1,000	48.6	51.4	6.7	19.6	65.2	50.8	49.2	6.9	20.2	63.1	2.2	-2.2	0.2	0.6	-2.1
7	Vertical Anchor Up Set A	1,000	48.5	51.5	6.8	18.5	66.8	51.2	48.8	6.7	18.7	65.8	2.7	-2.7	-0.1	0.2	-1.0
8	Vertical Anchor Up Set B	1,250	48.7	51.3	6.3	18.5	67.6	48.9	51.1	6.2	18.0	67.7	0.2	-0.2	-0.2	-0.5	0.1

*Differences of 5%+ are highlighted.

Table 5.26. Sample Theta Score Comparison after sampling

		Target					Sample					Difference (Target – Sample)	
Grade	VL	N	Mean	SD	Min.	Max.	N	Mean	SD	Min.	Max.	Mean	SD
ELA													
6	Vertical Anchor Down Set B	23,322	0.27	0.96	-3.28	3.44	1,250	0.31	0.96	-2.62	3.44	0.04	0.00
8	Vertical Anchor Up Set A	23,252	0.34	0.96	-3.65	3.45	1,250	0.36	0.98	-2.78	2.75	0.02	0.02
8	Vertical Anchor Up Set B	23,252	0.34	0.96	-3.65	3.45	1,250	0.33	0.95	-2.60	2.96	-0.01	-0.01
Mathematics													
3	Vertical Anchor Down Set A	2,3858	0.06	1.42	-4.00	4.00	1,250	0.07	1.44	-3.84	4.00	0.01	0.02
3	Vertical Anchor Down Set B	23,858	0.06	1.42	-4.00	4.00	1,000	0.09	1.47	-3.58	4.00	0.03	0.05
7	Vertical Anchor Up Set A	22,893	0.17	1.25	-3.65	4.00	1,000	0.27	1.23	-3.30	4.00	0.10	-0.02
8	Vertical Anchor Up Set B	2,3177	0.24	1.39	-4.00	4.00	1,250	0.24	1.39	-2.92	4.00	0.00	0.00

5.6.1.4. Item Suppression and Anchor Drop

As reported in Table 5.11, 16 ELA items were suppressed, including one vertical linking item for ELA Grade 4, which is also a Grade 5 horizontal linking item. This vertical linking item was excluded from vertical scaling, meaning that between Grades 4 and 5, there will be 27 vertical linking items and Grade 5 will include 27 horizontal linking items.

To determine if anchors should be dropped from the anchor set, NWEA evaluated the suitability of the vertical scaling items using item difficulty (*p*-value): If the *p*-value of an item was higher for the lower grade than for the upper grade. That is, difficulty reversal was examined to see if an item was more difficult for the higher grade than for the lower grade. An item with reversed difficulty was a candidate to drop from the anchor set. Out of the 280 vertical linking items across grades and content areas, 25 of them were flagged for reversed *p*-value, as shown in Table 5.27. Many items were observed with very small reversal. For example, ELA Item 21073590 between Grades 5 and 6 was flagged with 0.008 *p*-value reversal (*p*-value of .6591 and .6569 for Grades 5 and 6, respectively). Instead of removing all items showing the difficulty reversal, NWEA computed two indices to set the criterion to drop an anchor:

- Pseudo *t*-test
- Item-level effect size

The pseudo *t*-test was computed as follows:

$$\frac{p_{j\_lowerG} - p_{j\_upperG}}{\sqrt{\dfrac{\hat{\sigma}^2_{lowerG} + \hat{\sigma}^2_{upperG}}{k}}} \tag{5.7}$$

where:

- p_{j_lowerG} and p_{j_upperG} are the *p*-value for an item *j* of the lower and upper grades, respectively.
- $\hat{\sigma}^2_{lowerG}$ and $\hat{\sigma}^2_{upperG}$ are the variance of the *p*-value for the 28 vertical scaling items for the lower and upper grades, respectively.
- *k* is the number of vertical scaling items, which is 28 for each pair of adjacent grades.

One *p*-value difference was significant at 95% ($df = \infty$, one-tail *t*-stat > 1.645) for Item 31238420 between Grades 7 and 8 in Mathematics.

Effect size was computed for each item as:

$$\frac{\hat{\mu}(Y)_{upperG} - \hat{\mu}(Y)_{lowerG}}{SD_{pooled}} \tag{5.8}$$

$$SD_{pooled} = \sqrt{\frac{(n_{upperG} - 1)\hat{\sigma}(Y)_{upperG} + (n_{lowerG} - 1)\hat{\sigma}(Y)_{lowerG}}{n_{upperG} + n_{lowerG} - 2}}, \tag{5.9}$$

where:

- $\hat{\mu}(Y)_{upperG}$ is the item-score mean for the upper grade.
- $\hat{\mu}(Y)_{lowerG}$ is the item-score mean for the lower grade.
- $\hat{\sigma}^2(Y)_{upperG}$ is the variance for the upper grade.
- $\hat{\sigma}^2(Y)_{lowerG}$ is the variance for the lower grade.
- n_{upperG} is the number of student for the upper grade.
- n_{lowerG} number of students for the lower grade.

The effect size standardizes the mean difference between adjacent grades using the square root of the pooled variance. Considering that the conventional rule of thumb is that an effect size of .2 or higher indicates medium difference, no item is flagged for medium difference.

Based on these two indices, no any additional vertical linking items were dropped from the anchor set after suppressing the one vertical linking item from ELA Grade 4.

Table 5.27. Vertical Linking Item *P*-Value Reversal

Grades	HL Grade	VL Grade	Item Code	Max. #Pts.	N (lower grade)	N (higher grade)	P-value (lower grade)	P-value (higher grade)	*P*-value difference	*t*-stat	Prob. of *t*-stat	Item-level Effect Size
ELA												
4,5	5	4	41145360	1	1,396	10,876	0.84	0.82	-0.013	0.414	0.660	-0.03
5,6	5	6	21073590	1	22,290	1,302	0.66	0.66	-0.003	0.120	0.548	-0.01
5,6	5	6	21073600	1	22,290	1,302	0.56	0.54	-0.022	0.817	0.791	-0.04
5,6	5	6	21107270	2	22,290	1,302	0.50	0.47	-0.038	1.412	0.918	-0.11
6,7	6	7	11191960	1	23,322	1,457	0.85	0.83	-0.019	0.592	0.722	-0.05
6,7	6	7	21071250	1	23,322	1,457	0.60	0.57	-0.029	0.882	0.809	-0.06
6,7	6	7	21071280	1	23,322	1,457	0.63	0.59	-0.040	1.240	0.890	-0.08
7,8	7	8	21072330	1	22,965	3,080	0.74	0.74	-0.004	0.112	0.544	-0.01
7,8	7	8	21096440	2	22,965	2,343	0.50	0.48	-0.013	0.349	0.636	-0.03
Mathematics												
3,4	3	4	31164720	1	17,761	1,266	0.75	0.72	-0.036	0.774	0.779	-0.08
3,4	3	4	31166250	1	17,761	1,263	0.77	0.74	-0.030	0.642	0.738	-0.07
5,6	5	6	31170600	1	22,249	1,258	0.58	0.56	-0.024	0.494	0.688	-0.05
5,6	5	6	31175280	1	22,249	1,263	0.73	0.73	-0.004	0.076	0.530	-0.01
5,6	5	6	31176600	1	22,249	1,258	0.54	0.48	-0.060	1.221	0.886	-0.12
5,6	5	6	31192180	1	22,249	1,263	0.70	0.68	-0.023	0.471	0.680	-0.05
5,6	5	6	31194390	1	22,249	1,258	0.85	0.81	-0.039	0.797	0.786	-0.11
6,7	6	7	31162050	1	23,277	1,250	0.62	0.60	-0.020	0.393	0.652	-0.04
6,7	6	7	31191080	1	23,277	1,355	0.39	0.38	-0.004	0.086	0.534	-0.01
6,7	6	7	31238630	2	23,277	1,355	0.34	0.31	-0.029	0.564	0.712	-0.08
7,8	7	8	31161810	1	22,358	2,488	0.31	0.29	-0.015	0.379	0.647	-0.03
7,8	7	8	31171850	1	22,358	1,250	0.43	0.39	-0.043	1.063	0.854	-0.09
7,8	7	8	31191460	1	22,358	1,250	0.41	0.40	-0.019	0.467	0.679	-0.04
7,8	7	8	31193540	1	22,358	2,488	0.55	0.51	-0.045	1.110	0.864	-0.09
7,8	8	7	31232590	1	1,255	18,793	0.50	0.44	-0.059	1.475	0.927	-0.12
7,8	7	8	31238420	2	22,358	2,488	0.60	0.52	-0.077	1.911	0.969	-0.19

Note: Highlighted values indicate a large *p*-value difference (.5 or higher) or significant *t*-statistics.

5.6.1.5. Vertical Scaling Process

NWEA performed three steps of calibration to create the vertical scales for ELA and Mathematics:

1. Calibrate the horizontal linking, operational, and vertical linking items across grades in a single calibration run where six grades were concurrently calibrated. All items across Grades 3–8 were placed on the same scale through vertical linking.
2. Equate any remaining items in the bank that were not administered in 2018 to the 2018 scale using horizontal linking items. The mean *b* transformation constant was computed between the old and new parameter estimates of horizontal linking items. This equating was carried out separately for each grade.
3. Calibrate each grade separately while fixing horizontal linking and operational item parameter estimates from Step 1.

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

5.6.1.6. Vertical Scale Results

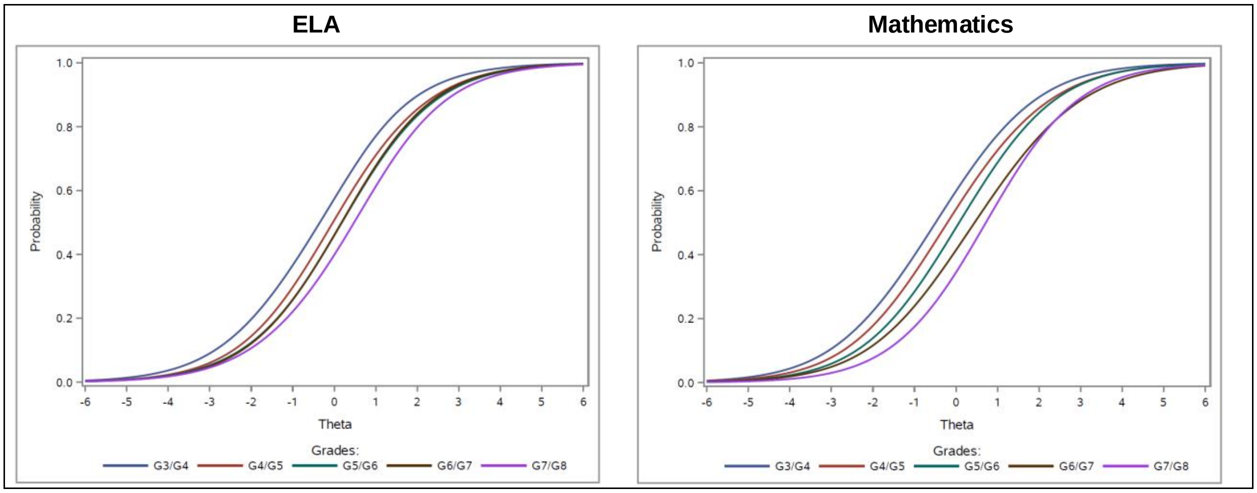
Vertical scaling results include the following:

- Test characteristic curves (TCCs)
- Grade-to-grade growth by computing the mean of student ability for each grade
- Grade-to-grade variability via the magnitude of the standard deviation (SD) of student ability for each grade
- Separation of grade distribution to show the degree of separation of performance distributions between adjacent grades

5.6.1.6.1. Test Characteristic Curves (TCCs)

Figure 5.2 presents the TCCs for ELA and Mathematics, respectively, using only vertical anchor item sets. For ELA, the TCCs reveal that the overall difficulty of the tests increases as the grade increases. However, the TCCs for Grades 5/6 and 6/7 are overlapped, indicating that the VL items between Grades 5/6 and 6/7 are very similar in terms of item difficulty. Case 3 presents the same pattern of difficulty, except the TCCs do not overlap. For Mathematics, the TCCs reveal that the overall difficulty of the tests increases as the grade increases. However, the test difficulty for Grades 6/7 folds onto the upper end of the Grades 7/8 scale at an ability of a theta 2.0.

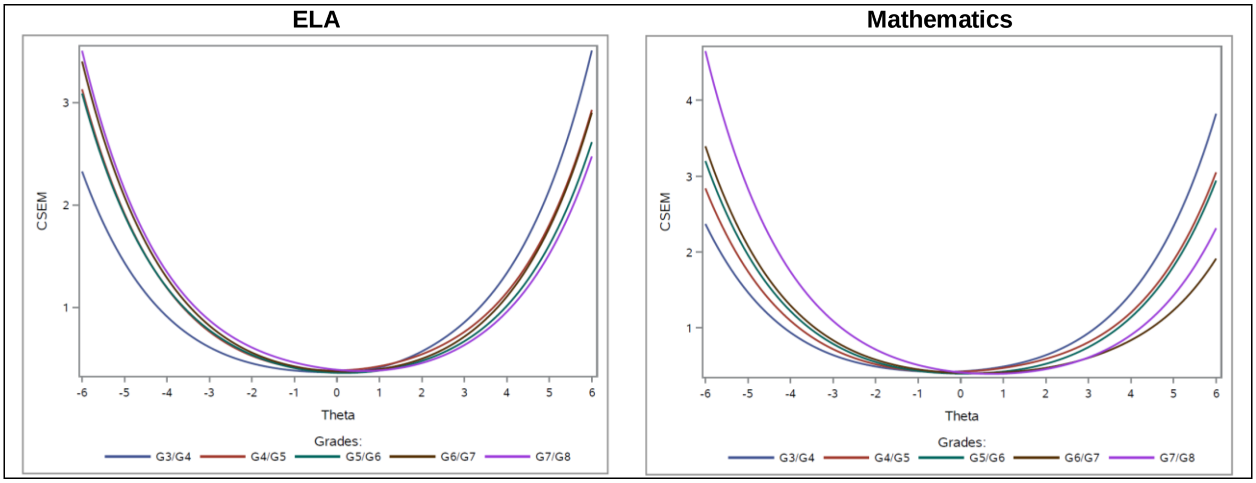
Figure 5.2. TCCs based on VL Item Sets



5.6.1.6.2. Conditional Standard Error of Measurement (CSEM)

Figure 5.3 presents the conditional standard error of measurement (CSEM) curve calculated for the vertical anchor item sets. The CSEM is high for low- and high-ability students, as expected, indicating that vertical scales measure well over the middle part of the theta distribution.

Figure 5.3. CSEM Curve based on VL Item Sets



ELA

Mathematics

5.6.1.6.3. Grade-to-Grade Growth

To examine the grade-to-grade growth, the mean of student ability (theta) was computed and compared across grades. Table 5.28 presents the mean of theta for each grade and content area. Mean difference represents the difference in mean between adjacent grades. As the grade increases, the mean of student ability increases, as expected. Figure 5.4 presents the trend of mean for each grade and content area. Table 5.29 presents student ability (theta) at the 5th, 15th, 25th, 35th, 45th, 55th, 65th, 75th, 85th, and 95th percentile ranks.

Table 5.28. Descriptive Statistics of Student Ability (Theta) and Mean Difference

Content Area	Grade	N	Mean	SD	Min.	Max.	Mean Difference
ELA	3	23,875	-0.26	1.05	-3.74	4.58	–
	4	23,873	0.16	0.99	-3.53	4.67	0.42
	5	22,290	0.43	0.92	-2.82	5.84	0.27
	6	23,322	0.53	0.92	-2.32	4.00	0.10
	7	22,965	0.69	1.02	-2.72	5.00	0.16
	8	23,252	0.84	0.91	-2.61	6.17	0.15
Mathematics	3	23,858	-0.14	1.30	-5.90	5.05	–
	4	23,826	0.49	1.22	-2.98	5.75	0.63
	5	22,249	0.76	1.21	-3.83	5.83	0.27
	6	23,277	0.98	1.32	-3.49	6.15	0.22
	7	22,893	0.99	1.23	-2.45	6.70	0.01
	8	23,177	1.28	1.30	-2.39	6.52	0.29

Figure 5.4. Mean by Grade

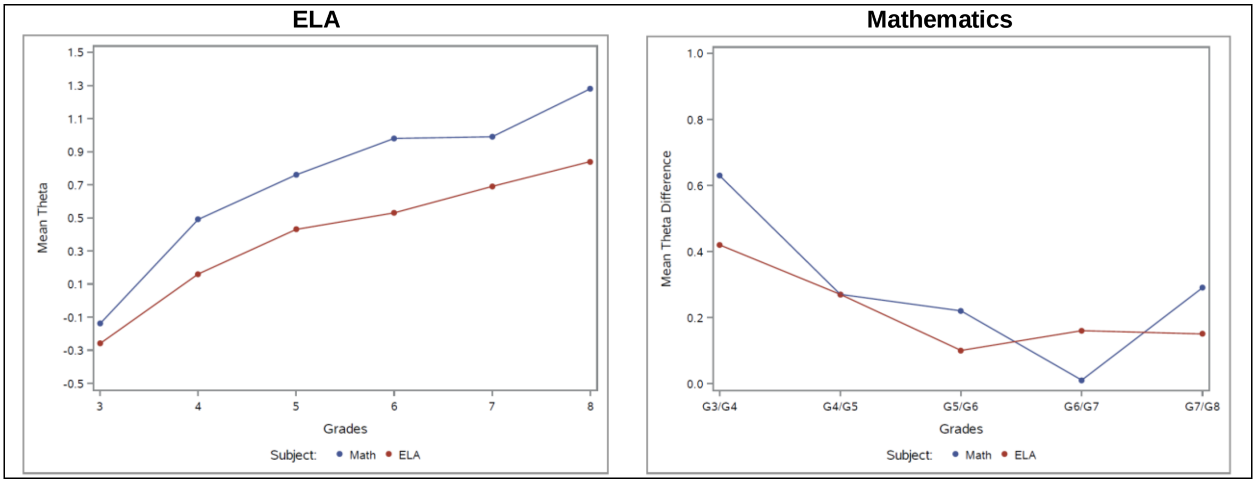


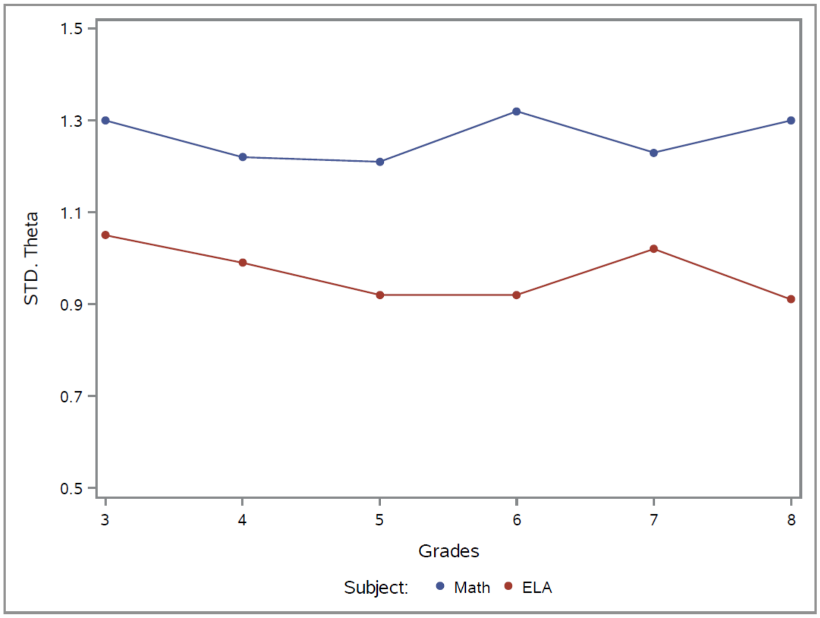
Table 5.29. Student Ability (Theta) by Percentile Rank

Content Area	Grade	Percentile Rank									
		P5	P15	P25	P35	P45	P55	P65	P75	P85	P95
ELA	3	-1.98	-1.41	-1.02	-0.68	-0.37	-0.10	0.19	0.47	0.83	1.46
	4	-1.48	-0.90	-0.52	-0.22	0.04	0.30	0.55	0.83	1.20	1.78
	5	-1.08	-0.56	-0.23	0.06	0.33	0.56	0.81	1.06	1.38	1.93
	6	-0.97	-0.46	-0.09	0.18	0.42	0.66	0.90	1.16	1.50	2.03
	7	-0.93	-0.42	-0.07	0.26	0.55	0.83	1.12	1.42	1.78	2.38
	8	-0.72	-0.13	0.22	0.52	0.74	0.97	1.19	1.46	1.81	2.27
Mathematics	3	-2.22	-1.48	-1.02	-0.67	-0.36	-0.04	0.32	0.70	1.26	2.03
	4	-1.32	-0.78	-0.42	-0.09	0.23	0.56	0.91	1.29	1.81	2.62
	5	-1.04	-0.44	-0.07	0.24	0.53	0.80	1.10	1.45	1.98	2.89
	6	-1.14	-0.37	0.06	0.44	0.78	1.12	1.47	1.83	2.26	3.15
	7	-0.75	-0.23	0.11	0.40	0.68	0.98	1.33	1.75	2.30	3.27
	8	-0.60	-0.09	0.31	0.65	1.01	1.35	1.74	2.16	2.62	3.67

5.6.1.6.4. Grade-to-Grade Variability

Figure 5.5 shows the magnitude of the standard deviation (SD) of student ability (theta) for each grade and content area. The magnitude of SD across grades is consistent, without showing any increasing or decreasing pattern. In other words, the high-achieving and low-achieving students tend to grow at similar rates as the grade increases. The SD is bigger for Mathematics compared to ELA. The sample size does not have much impact on the magnitude of the SD regardless of the content area.

Figure 5.5. Standard Deviation by Grade



5.6.1.6.5. Separation of Grade Distribution

To show the degree of separation of performance distributions between adjacent grades, effect size and horizontal distance (HD) were calculated. Table 5.30 presents the effect sizes for each grade and content area, as well as the HD at the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentile points. Figure 5.6 presents the effect sizes between adjacent grades.

Grades 3/4 have the biggest effect size, thus indicating more separation compared to the other grade bands. That pattern is the same for both ELA and Mathematics. The HD is greatest between Grades 3 and 4, which confirms the effect size results. Between Mathematics Grades 6 and 7, the HDs are all negative for P50 and higher percentile points, which indicates negative growth between these grades for the upper half of the student distribution. This may suggest that the vertical scale is not working well for Mathematics Grades 6 and 7. Consistent with the HD results, some overlaps or reversals are observed among ELA Grades 5, 6, and 7 and between Mathematics Grades 6 and 7.

Table 5.30. Effect Size and Horizontal Distance

Content Area	Grades	Effect Size	Horizontal Distance						
			P5	P10	P25	P50	P75	P90	P95
ELA	3–4	0.41	0.50	0.51	0.50	0.50	0.90	0.77	0.60
	4–5	0.28	0.40	0.36	0.29	0.22	0.28	0.33	0.35
	5–6	0.11	0.11	0.08	0.14	0.17	-0.10	0.01	0.13
	6–7	0.16	0.04	0.07	0.02	0.01	0.39	0.23	0.05
	7–8	0.16	0.20	0.28	0.29	0.23	0.15	0.13	0.19
Mathematics	3–4	0.39	0.36	0.38	0.32	0.59	0.59	0.56	0.59
	4–5	0.29	0.23	0.16	0.15	0.27	0.16	0.16	0.26
	5–6	0.07	0.10	0.11	0.10	0.30	0.38	0.40	0.27
	6–7	0.18	0.26	0.33	0.34	-0.13	-0.08	-0.03	0.11
	7–8	0.16	0.05	-0.01	-0.10	0.34	0.41	0.26	0.41

Figure 5.6. Effect Sizes between Adjacent Grades

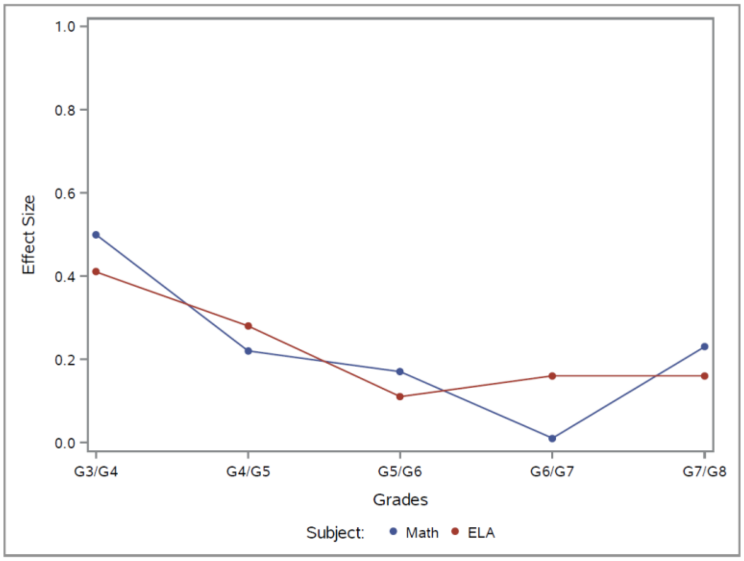
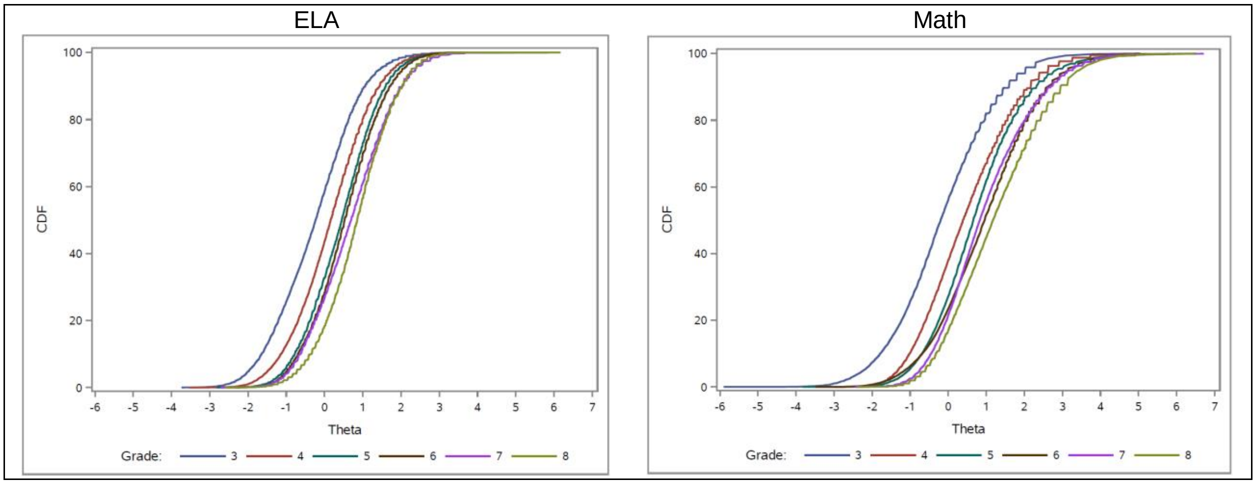


Figure 5.7. Cumulative Distribution Function (CDF)



5.6.1.7. <u>2019 Scaling Considerations</u>

To ensure the quality of vertical scales created in 2018, the same scaling process will be applied in 2019 to verify the 2018 vertical scales. Both vertical and horizontal linking items will be included for 2019. New sets of linking items will be selected and administered for 2019, but the overall design and process will be similar as for 2018. The results from 2019 vertical scaling will be compared to that from 2019 pre-equating so that the 2018 vertical scales can be verified.

Further, operational items with significant item parameter estimate changes using the Robust Z method will not be included for the 2019 item pool. A total of 237 items were flagged using the Robust Z statistics of +/-1.645 critical value, as reported in Table 5.31. Some of these items may remain in the 2019 pool to meet the TOS, including items to be used for 2019 breach forms.

Table 5.31. Number of Items with a Large Parameter Change

	#Items		
Grade	ELA	Mathematics	Total
3	12	21	33
4	28	12	40
5	25	21	46
6	12	23	35
7	21	28	49
8	20	14	34
Total	**118**	**119**	**237**

5.6.2. Equating Items to the 2018 Vertical Scale (ELA and Mathematics)

To equate the items that were calibrated prior to 2018 to the 2018 vertical scale, the scaling constant were estimated using the following steps:

1. Identify eligible items.
    a. All anchor item sets
    b. Use only horizontal anchor items, but drop those with a large *p*-value change from previous year (i.e., *p*-value change >.1)
2. Estimate the mean of difficulty item parameters from the previous year.
3. Estimate the mean of difficulty item parameters from the 2018 test administration.
4. Estimate the scaling constant as the difference between 2018 mean and previous mean.
5. Transform the 2017 cuts to the 2018 scale by applying the scaling constant.

Table 5.32 presents the resulting scaling constants.

Table 5.32. Scaling Constants

Content Area	Grade	#Items	#Parameters	Mean (2018)	Mean (Bank)	Scaling Constant
ELA	3	20	26	-0.4177	0.1524	-0.5701
	4	26	31	-0.1149	0.1849	-0.2998
	5	26	31	0.0235	0.1188	-0.0953
	6	27	32	0.2178	0.0855	0.1323
	7	26	30	0.2837	-0.0605	0.3442
	8	19	22	0.3304	-0.0065	0.3369
Mathematics	3	14	14	-1.2006	-0.8806	-0.3200
	4	23	23	-0.5805	-0.6951	0.1145
	5	20	20	-0.4995	-0.8392	0.3397
	6	22	25	0.2278	-0.4862	0.7140
	7	24	27	0.6687	-0.0740	0.7427
	8	19	20	0.5423	-0.5566	1.0989

5.6.3. Post-Equating Check (Science)

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

5.6.3.1. Post-Equating Method

Because the 2017 Science forms were reused, all the 2018 operational items were used as the linking set. This means that the raw-to-scale score (RSS) conversion tables were established prior to the operational administration. This is referred to as pre-equating because it is conducted before the operational test administration. However, it may not be appropriate to assume that the operational items maintained their relative difficulty across administrations. The same item can perform differently across administrations due to changes in the item's position or changes in the students' experiences. When the 2018 operational test data became available, the item difficulty equivalence was checked using the Robust Z post-equating check procedure to identify items that show significant difficulty changes from the bank values. If no unstable items are identified, the 2017 equating process would result in the pre-equating solution, whereas a post-equating solution would be used if items are found to be outside the normal estimation error. The subset of 2018 operational items, with the identified items excluded, was used as the set to estimate the link constant to map the 2018 test to the bank scale. This equating process is known as post-equating because it occurs after the operational administration and the RSS conversion is generated based on the operational test data.

As part of the post-equating check procedures, the item difficulty equivalence was checked by comparing the old banked item calibration (i.e., pre-calibration) with a new unanchored calibration of the 2018 data (i.e., post-calibration) using WINSTEPS 3.91.0.0 (Linacre, 2015). The evaluations were conducted for each grade using the Robust *Z* statistic (Huynh & Meyer, 2010). This method focuses on the correlations between the pre- and post-calibrated item difficulties and the ratio of standard deviations (RSD) between the two calibrations. The correlation between the two item difficulty estimates should be 0.95 or higher, and the RSD between the two sets of item difficulty estimates should range between 0.90 and 1.10 (Huynh & Meyer, 2010). To detect inconsistent item difficulty estimates, a critical value for the Robust *Z* statistic of ±1.645 was used. Items that exceeded the Robust *Z* critical value were deleted, one item at a time, until both the item difficulty correlation and SD ratio fell within the prescribed limits.

5.6.3.2. Post-Equating Results

Table 5.33 presents the 2018 Science correlation statistics and SD ratio following the process described above. Table 5.34 presents the percentage of students at each achievement level. The percentage of students at Below the Standards increases slightly (by approximately 2%).

Table 5.33. Science Pre- and Post-Equating Comparison

Grade	Iteration	SD Pre	SD Post	RSD	Correlation
5	1	0.73	0.69	1.06	0.931
5	2, excluded 41141670	0.73	0.70	1.05	0.940
5	3, excluded 41144270	0.71	0.69	1.02	0.945
5	4, excluded 41141820	0.72	0.69	1.04	0.951
8	1	0.70	0.67	1.04	0.955

Table 5.34. Science Achievement Level Distribution for 2017 and 2018

	2017*				2018*			
Grade	N	%Below	%Meets	%Exceeds	N	%Below	%Meets	%Exceeds
5	23,310	28.2	53.7	18.1	22,251	30.2	54.3	15.5
8	22,856	31.3	46.2	22.5	23,190	33.1	47.5	19.4

*The 2017 percentages are from the 2017 NeSA technical report, which includes all students. The 2018 percentages are from the 2018 psychometric analyses data, which includes students taking online tests who attempted 10 or more operational items.

Table K.1, found in Appendix K, presents the item parameter estimates for Grades 5 and 8 when all items were used. The item difficulty correlation is 0.931 for Grade 5 Science when all items were used, which did not meet the Robust Z criteria. Consequently, Grade 5 items with the highest absolute Robust Z statistic were excluded one item at a time (Item 41141670 first, followed by Items 41144270 and 41141820). With these three items excluded, the correlation is 0.951 and the RSD is 1.04, which met both criteria. The item difficulty correlation is 0.955 and the RSD is 1.04 for Grade 8 Science when all items were used, which met both criteria.

Table K.2 presents and compares the pre- and post-equated scoring tables for Grade 5 with student frequency. As shown in the "SS Diff (Pre – Post)" column, scale scores differ up to two points, but achievement levels are the same. Table K.3 presents the pre-equated scoring tables for Grade 8 with student frequency.

5.6.4. Scaling

The previously set scaling constants for Science were used again in 2018. For ELA and Mathematics, scaling constants were set 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/σ_{G5} (i.e., slope=72.47244) and intercept of 2500 for ELA
- A slope of 66.6/σ_{G5} (i.e., slope=54.92622) and intercept of 1200 for Mathematics

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

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

$$\begin{aligned} &\text{SS} = (\theta \times \text{A}) + \text{B} \\ &\text{SSCSEM} = \theta\text{-CSEM} \times \text{A}. \end{aligned} \quad (5.10)$$

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

$$\theta\text{-CSEM} = \text{CSEM}(\theta_i) = \frac{1}{\sqrt{I(\theta_i)}} \quad (5.11)$$

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

$$I(\theta_i) = \sum_j \frac{p'_{ij}(\theta_i)^2}{p_{ij}(\theta_i)\, q_{ij}(\theta_i)} \tag{5.12}$$

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

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

1. The HOSS must increase as the grade increases for tests on a vertical scale.
2. The HOSS should be high enough that it does not cause an unnecessary "pile-up" of scale scores at the HOSS, targeting less than 1%.
3. The HOSS should be low enough that SSCSEM(HOSS) < 10 x Min(SSCSEM).
4. The HOSS may be high enough that SSCSEM (Penultimate HOSS) < 5 x 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 5.35 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 5.35. Score Range (LOSS and HOSS) and Assigned Score—ELA, Mathematics, and Science

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

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
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
Science					
5	0	200	0	1	2
8	0	200	0	1	2

Table 5.36 summarizes the cut score implementation, or the conversions of student ability (theta) to scale scores that were used for the Spring 2018 final scoring. Specifically, the table presents the calculations of the slopes and intercepts for all grades of the scale score conversions for ELA, Mathematics, and Science. This conversion table was used for the Spring 2018 final scoring. Please refer to the next section, Section 6, for details on how cut scores were decided through the standard setting process.

Table 5.36. Conversion of Theta to Scale Scores

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

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

Section 6: Standard Setting

6.1. Overview

The NDE held a standard setting for the NSCAS Mathematics assessments and a cut score review for ELA from July 26–28, 2018, using the Item-Descriptor (ID) Matching method to determine the cut scores delineating the Developing, On Track, and CCR Benchmark achievement levels. The purpose of the standard setting was to set new cut scores for the NSCAS Mathematics tests, whereas the purpose of the cut score review was to validate the existing cut scores for the NSCAS ELA tests. The standard setting was conducted concurrently with the cut score review. For more in-depth information, please refer to the full standard setting and cut score review reports (EdMetric, 2018a, 2018b). No changes were made to the Science standards or assessments, and therefore a standard setting was not necessary.

Standard setting is a critical piece of evidence in establishing the validity of an assessment. As such, a standard setting must be conducted with objectivity, integrity, and attention to technical detail. 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 the 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.

6.2. Purpose of the Standard Setting and Cut Score Review

Nebraska's statewide assessment system underwent significant changes between the Spring 2016 and Spring 2017 administrations. 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, the 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.

The 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. The 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, the 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, the NDE conducted a full standard setting for the NSCAS Mathematics assessments. The 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.

6.3. ID Matching Method

The *Standards* (AERA et al., 2014) emphasize the selection of a standard setting methodology that is appropriate for the assessment being administered. Based on the technical characteristics of the NSCAS ELA and Mathematics assessments and their intended uses, NWEA and EdMetric, with the input of NDE's TAC, determined that the ID Matching methodology for standard setting would be most appropriate for the standard setting and cut score review. The ID Matching method brings together diverse panels of experts in the applicable content area (typically a wide representation of classroom educators) who complete a deep study of the content of the test items and the 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 item response theory (IRT) and assessments that include multiple item types because panelists consider the content of test items that are placed in an ordered item booklet (OIB) in ascending order of difficulty based on IRT item statistics derived from actual student performance. Using ID Matching, panelists match item demands to those described in the Range ALDs. To ensure alignment with the already-established ACT cut scores in Grade 11, panelists were provided a range of items in which they could set their cut scores.

6.4. Meeting Materials

The following materials were used during the Mathematics standard setting and ELA cut score review meetings. To review all the meeting materials, please refer to the full standard setting and cut score review reports provided by EdMetric (EdMetric, 2018a, 2018b).

- Range ALDs
- General session PowerPoint slide deck
- Training PowerPoint slide decks
- ID Matching PowerPoint slide deck
- Evaluation surveys
- Representative operational test forms
- OIB (a group of items representing the constructs measured by an assessment, in ascending order according to item difficulty)
- Item map (a table showing each item in an OIB)
- Rating sheets
- Online control panel developed by EdMetric

6.5. Meeting Process

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

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

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

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

6.6. Achievement Level Descriptors (ALDs)

At the foundation, the ID Matching method requires clear ALDs that describe the KSAs of a student at a particular achievement level. Using those ALDs to identify a cut score ensures alignment of the assessment system and allows educators to focus on the ALDs during instructional adaptations to effect change in student learning and performance. Draft ELA and Mathematics Range ALDs were first developed by NWEA and the NDE and then brought to the standard setting and cut score meetings to be reviewed and refined by educators. The final ELA and Mathematics Range ALDs, after being finalized and approved by the NDE, are provided in the standard setting and cut score review reports (EdMetric, 2018a, 2018b), as well as posted online on the NDE's website. For Science, updated ALDs will be generated before the new assessment becomes operational in Spring 2021.

Specifically, to develop the Mathematics ALDs, an educator committee was convened in April 2018 to review a first draft. After that meeting, NWEA and the NDE engaged in an extensive process of revision that involved several iterations of rework. After the Mathematics standard setting, educators reviewed the ALDs based on the cut sources that had just been set to refine the ALDs. NWEA and the NDE worked iteratively to finalize the Mathematics ALDs. For ELA, the educator committees for each grade were asked to use the ALDs used from the original standard setting during the cut score review in August 2018. After the cut score review and standard setting cut score recommendations were complete, an expert in the development of

ALDs trained ELA and Mathematics participants on the tenets of the Range ALD process. The training and presenter were the same as was given to the original set of teachers who reviewed the Mathematics ALDs during their original development process. While the training given to participants was the same regarding the framework of ALD constructional principals, the work participants engaged in to develop the Reporting ALDs differed.

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. NWEA content specialists reviewed the draft ALDs for each grade, 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. This created a third draft of ALDs. Once a coherent and cohesive third draft was created, it was sent to NDE's ELA specialists for review. When NDE returned feedback, the NWEA ELA team implemented it and sent the resulting fourth draft to the NDE for an additional review. The NDE signed off on this document, creating the current version of the ELA ALDs available on the NDE website at https://www.education.ne.gov/assessment/nscas-general-summative-assessment/nscas-english-language-arts-ela/.

Mathematics participants were asked to identify items in the OIB that they felt had not matched the ALDs during the standard setting process. Participants were trained that the order in the OIB showed how difficult items were for students. Using the content-recommended cut scores, participants could study the items that were inconsistent with the ALDs and suggest edits to the ALDs. The grade-level groups began this task at their own pace. NWEA reviewed the participants' recommendations as the ALDs were finalized along with the items in the OIB. After receiving the final approved cut scores that were higher than the content-recommended cut scores, NWEA content specialists reconciled the ALDs in line with the items in the OIB based on the content of the items, participant recommendations, and the final cut scores consistent with recommended practice (Egan, Schneider & Ferrara, 2012). Those edits were then used to inform changes throughout the ALDs for the levels or indicators that were not represented in the OIB, as edits to one level or indicator had an impact on the ALDs at another indicator given the integration of content across levels that are important for coherence. These updates were shared with the NDE for feedback. After receiving NDE's feedback, NWEA responded to their queries, making the requested edits or responding to the posted questions. The files were then formatted and submitted to the NDE to share with the field. The final Mathematics ALDs are available on the NDE website at https://www.education.ne.gov/assessment/nscas-general-summative-assessment/nscas-mathematics/.

6.7. Final Results

The recommended cut scores were presented to the Nebraska State Board of Education on August 2, 2018. Table 6.1 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 rather than to the analysis data (Section 5) or results data (Sections 7 and 8) used in this technical report.

Table 6.1. 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

Section 7: Test Results

All students who took the online, paper-pencil, and Spanish forms of the 2018 NSCAS Summative assessments were included in the test results. For results based on demographics and accommodations, all participants (i.e., student who attempted at least one item) were included. For all other results in this section, students who attempted at least 10 operational items on the online and paper-pencil forms were used (i.e., Spanish test-takers were not included). Results presented in this section are not from the state student file that the NDE received and may therefore differ slightly from the official state summary report due to ongoing resolution of test materials and slight differences in the application of exclusion rules.

7.1. Demographics and Accommodations

Table 7.1 – Table 7.6 present the number of tested students by demographics for each grade and content area, including gender, ethnicity, free and reduced lunch (FRL) status, limited English proficiency (LEP) status, special education (SPED) status, use of universal features (i.e., answer eliminator, highlighter, notepad, and zoom), and use of accommodations (text-to-speech (TTS), paper-pencil form, Spanish online or paper-pencil form, Braille, and large print). Starting in 2018, both current and former English language learner (ELL) students are considered to have LEP status, resulting in more LEP students in 2018 compared to 2017.

As shown in these tables, more than 22,000 students took the assessment in each grade and content area. Of those students across grades, half are males, half are females, over two thirds are white, and less than one fifth are Hispanic. Among the students across grades, about 44% to 49% are eligible for FRL, 14–16% are LEP/ELL, and 14–16% 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.

Table 7.1. Number of Students Tested by Demographics—Grade 3

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
	Total N-Count	**24,349**	**100.0**	**24,315**	**100.0**
Gender	Female	11,612	48.8	11,606	48.8
	Male	12,189	51.2	12,181	51.2
Ethnicity	AI/AN	287	1.2	285	1.2
	Asian	644	2.7	645	2.7
	Black or African American	1,565	6.7	1,563	6.7
	Hispanic	4,609	19.6	4,618	19.6
	NH/PI	33	0.1	33	0.1
	White	15,332	65.3	15,345	65.3
	Two or More Races	1,023	4.4	1,024	4.4
FRL	Yes	11,718	49.2	11,691	49.1
	No	12,111	50.8	12,108	50.9
LEP	Yes	3,805	16.0	3,793	15.9
	No	20,024	84.0	20,006	84.1

		ELA		Mathematics	
Demographic Sub-Group*		N	%	N	%
SPED	Yes	3,981	16.7	3,974	16.7
	No	19,848	83.3	19,825	83.3
Universal Features & Accommodations	Answer Choice Eliminator	14,746	60.6	14,161	58.2
	Highlighter	12,925	53.1	8,343	34.3
	Line Reader	15,811	64.9	7,180	29.5
	Notepad	10,497	43.1	9,068	37.3
	Text-to-Speech (TTS)	3,947	16.2	3,714	15.3
	Zoom	11,594	47.6	6,746	27.7
	Paper-Pencil (PP)	37	0.2	36	0.2
	Spanish Online	28	0.1	46	0.2
	Spanish Paper-Pencil (PP)	5	0.0	5	0.0
	Braille**	–	–	–	–
	Large Print**	7	–	8	–

*AI/AN = American Indian or Alaskan Native. NH/PI = Native Hawaiian or Other Pacific Islander. FRL = free and reduced lunch. LEP = limited English proficient. SPED = special education.
**Braille and Large Print counts are based on students who actually tested and were not included in the total n-count.

Table 7.2. Number of Students Tested by Demographics—Grade 4

		ELA		Mathematics	
Demographic Sub-Group*		N	%	N	%
	Total N-Count	**24,360**	**100.0**	**24,344**	**100.0**
Gender	Female	11,566	48.5	11,549	48.5
	Male	12,277	51.5	12,252	51.5
Ethnicity	AI/AN	282	1.2	282	1.2
	Asian	664	2.8	664	2.8
	Black or African American	1,644	7.0	1,644	7.0
	Hispanic	4,528	19.2	4,530	19.2
	NH/PI	34	0.1	34	0.1
	White	15,419	65.5	15,409	65.4
	Two or More Races	989	4.2	991	4.2
FRL	Yes	11,567	48.5	11,572	48.6
	No	12,282	51.5	12,258	51.4
LEP	Yes	3,806	16.0	3,810	16.0
	No	20,043	84.0	20,020	84.0
SPED	Yes	3,937	16.5	3,932	16.5
	No	19,912	83.5	19,898	83.5

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
Universal Features & Accommodations	Answer Choice Eliminator	16,795	68.9	17,195	70.6
	Highlighter	13,344	54.8	8,402	34.5
	Line Reader	16,031	65.8	7,034	28.9
	Notepad	11,626	47.7	10,325	42.4
	Text-to-Speech (TTS)	3,823	15.7	3,420	14.1
	Zoom	11,502	47.2	6,798	27.9
	Paper-Pencil (PP)	41	0.2	41	0.2
	Spanish Online	59	0.2	68	0.3
	Spanish Paper-Pencil (PP)	1	0.0	1	0.0
	Braille**	3	–	3	–
	Large Print**	8	–	8	–

*AI/AN = American Indian or Alaskan Native. NH/PI = Native Hawaiian or Other Pacific Islander. FRL = free and reduced lunch. LEP = limited English proficient. SPED = special education.
**Braille and Large Print counts are based on students who actually tested and were not included in the total n-count.

Table 7.3. Number of Students Tested by Demographics—Grade 5

Demographic Sub-Group*		ELA		Mathematics		Science	
		N	%	N	%	N	%
	Total N-Count	**22,751**	**100.0**	**22,736**	**100.0**	**22,683**	**100.0**
Gender	Female	10,749	48.4	10,746	48.4	10,742	48.4
	Male	11,481	51.7	11,475	51.6	11,456	51.6
Ethnicity	AI/AN	289	1.3	288	1.3	289	1.3
	Asian	588	2.7	588	2.7	588	2.7
	Black or African American	1,488	6.8	1,493	6.8	1,493	6.8
	Hispanic	4,184	19.0	4,187	19.0	4,174	19.0
	NH/PI	39	0.2	39	0.2	39	0.2
	White	14,515	66.0	14,526	66.0	14,518	66.0
	Two or More Races	892	4.1	894	4.1	893	4.1
FRL	Yes	10,521	47.2	10,490	47.1	10,482	47.2
	No	11,764	52.8	11,774	52.9	11,735	52.8
LEP	Yes	3,267	14.7	3,252	14.6	3,251	14.6
	No	19,018	85.3	19,012	85.4	18,966	85.4
SPED	Yes	3,456	15.5	3,454	15.5	3,407	15.3
	No	18,829	84.5	18,810	84.5	18,810	84.7

Demographic Sub-Group*		ELA		Mathematics		Science	
		N	%	N	%	N	%
Universal Features & Accommodations	Answer Choice Eliminator	14,281	62.8	14,827	65.2	12,474	55.0
	Highlighter	9,770	42.9	5,689	25.0	5,123	22.6
	Line Reader	13,545	59.5	5,284	23.2	5,114	22.6
	Notepad	9,327	41.0	8,187	36.0	6,293	27.7
	Text-to-Speech (TTS)	3,223	14.2	2,702	11.9	2,927	12.9
	Zoom	9,663	42.5	4,783	21.0	5,263	23.2
	Calculator (basic)	–	–	485	2.1	–	–
	Paper-Pencil (PP)	30	0.1	29	0.1	1	0.0
	Spanish Online	31	0.1	63	0.3	63	0.3
	Spanish Paper-Pencil (PP)	3	0.0	4	0.0	4	0.0
	Braille**	2	–	2	–	2	–
	Large Print**	6	–	6	–	6	–

*AI/AN = American Indian or Alaskan Native. NH/PI = Native Hawaiian or Other Pacific Islander. FRL = free and reduced lunch. LEP = limited English proficient. SPED = special education.
**Braille and Large Print counts are based on students who actually tested and were not included in the total n-count.

Table 7.4. Number of Students Tested by Demographics—Grade 6

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
	Total N-Count	**23,864**	**100.0**	**23,793**	**100.0**
Gender	Female	11,405	49.0	11,391	48.9
	Male	11,894	51.1	11,884	51.1
Ethnicity	AI/AN	291	1.3	290	1.3
	Asian	626	2.7	626	2.7
	Black or African American	1,600	6.9	1,596	6.9
	Hispanic	4,425	19.2	4,425	19.2
	NH/PI	31	0.1	32	0.1
	White	15,217	66.0	15,215	66.0
	Two or More Races	872	3.8	873	3.8
FRL	Yes	11,133	47.6	11,082	47.5
	No	12,249	52.4	12,230	52.5
LEP	Yes	3,812	16.3	3,790	16.3
	No	19,570	83.7	19,522	83.7
SPED	Yes	3,557	15.2	3,552	15.2
	No	19,825	84.8	19,760	84.8

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
Universal Features & Accommodations	Answer Choice Eliminator	13,894	58.2	17,276	72.6
	Highlighter	9,282	38.9	7,709	32.4
	Line Reader	12,883	54.0	6,998	29.4
	Notepad	8,820	37.0	10,757	45.2
	Text-to-Speech (TTS)	3,144	13.2	2,507	10.5
	Zoom	8,456	35.4	4,204	17.7
	Calculator (basic)	–	–	18,602	78.2
	Calculator (scientific)	–	–	790	3.3
	Paper-Pencil (PP)	32	0.1	32	0.1
	Spanish Online	55	0.2	85	0.4
	Spanish Paper-Pencil (PP)	5	0.0	5	0.0
	Braille**	2	–	2	–
	Large Print**	2	–	2	–

*AI/AN = American Indian or Alaskan Native. NH/PI = Native Hawaiian or Other Pacific Islander. FRL = free and reduced lunch. LEP = limited English proficient. SPED = special education.
**Braille and Large Print counts are based on students who actually tested and were not included in the total n-count.

Table 7.5. Number of Students Tested by Demographics—Grade 7

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
	Total N-Count	**23,524**	**100.0**	**23,565**	**100.0**
Gender	Female	11,148	48.6	11,137	48.6
	Male	11,814	51.5	11,790	51.4
Ethnicity	AI/AN	266	1.2	267	1.2
	Asian	602	2.7	601	2.7
	Black or African American	1,534	6.8	1,532	6.8
	Hispanic	4,215	18.6	4,215	18.6
	NH/PI	43	0.2	43	0.2
	White	15,209	67.0	15,198	67.0
	Two or More Races	845	3.7	845	3.7
FRL	Yes	10,706	46.5	10,746	46.6
	No	12,309	53.5	12,320	53.4
LEP	Yes	3,562	15.5	3,587	15.6
	No	19,453	84.5	19,479	84.5
SPED	Yes	3,384	14.7	3,394	14.7
	No	19,631	85.3	19,672	85.3

Demographic Sub-Group*		ELA		Mathematics	
		N	%	N	%
Universal Features & Accommodations	Answer Choice Eliminator	11,283	48.0	14,380	61.0
	Highlighter	6,633	28.2	4,804	20.4
	Line Reader	10,260	43.6	6,078	25.8
	Notepad	6,884	29.3	8,599	36.5
	Text-to-Speech (TTS)	2,477	10.5	1,887	8.0
	Zoom	5,840	24.8	3,681	15.6
	Calculator (basic)	–	–	1,513	6.4
	Calculator (scientific)	–	–	19,745	83.8
	Paper-Pencil (PP)	55	0.2	54	0.2
	Spanish Online	74	0.3	111	0.5
	Spanish Paper-Pencil (PP)	5	0.0	6	0.0
	Braille**	–	–	–	–
	Large Print**	6	–	5	–

*AI/AN = American Indian or Alaskan Native. NH/PI = Native Hawaiian or Other Pacific Islander. FRL = free and reduced lunch. LEP = limited English proficient. SPED = special education.
**Braille and Large Print counts are based on students who actually tested and were not included in the total n-count.

Table 7.6. Number of Students Tested by Demographics—Grade 8

Demographic Sub-Group*		ELA		Mathematics		Science	
		N	%	N	%	N	%
	Total N-Count	**23,880**	**100.0**	**23,882**	**100.0**	**23,780**	**100.0**
Gender	Female	11,333	48.8	11,314	48.7	11,301	48.8
	Male	11,913	51.3	11,904	51.3	11,874	51.2
Ethnicity	AI/AN	275	1.2	276	1.2	274	1.2
	Asian	630	2.7	629	2.7	630	2.8
	Black or African American	1,430	6.2	1,429	6.2	1,427	6.2
	Hispanic	4,239	18.5	4,227	18.4	4,231	18.4
	NH/PI	29	0.1	29	0.1	29	0.1
	White	15,589	67.9	15,579	67.9	15,571	67.9
	Two or More Races	785	3.4	786	3.4	784	3.4
FRL	Yes	10,468	44.8	10,467	44.8	10,415	44.8
	No	12,891	55.2	12,897	55.2	12,857	55.3
LEP	Yes	3,606	15.4	3,607	15.4	3,569	15.3
	No	19,753	84.6	19,757	84.6	19,703	84.7
SPED	Yes	3,254	13.9	3,263	14.0	3,203	13.8
	No	20,105	86.1	20,101	86.0	20,069	86.2

Demographic Sub-Group*		ELA		Mathematics		Science	
		N	%	N	%	N	%
Universal Features & Accommodations	Answer Choice Eliminator	9,609	40.2	14,873	62.3	7,466	31.4
	Highlighter	5,660	23.7	3,375	14.1	2,599	10.9
	Line Reader	7,966	33.4	5,455	22.8	3,064	12.9
	Notepad	5,540	23.2	7,408	31.0	3,659	15.4
	Text-to-Speech (TTS)	2,034	8.5	1,305	5.5	1,649	6.9
	Zoom	3,761	15.8	2,235	9.4	2,708	11.4
	Calculator (scientific)	–	–	20,209	84.6	–	–
	Paper-Pencil (PP)	80	0.3	70	0.3	22	0.1
	Spanish Online	74	0.3	105	0.4	120	0.5
	Spanish Paper-Pencil (PP)	2	0.0	12	0.1	7	0.0
	Braille**	1	–	1	–	1	–
	Large Print**	1	–	1	–	1	–

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

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

7.2. Students Tested and Mode Summary Data

The 2018 NSCAS assessments were administered online to the extent practical. NDE's effort of encouraging schools to take online tests worked, and very small number of students took the paper-pencil test. As shown in Table 7.7, less than 1% of students took the assessment in the paper-based version across all grades and content areas.

Table 7.7. Number of Students Tested by Administration Mode

Content Area	Grade	Total #Students	Online N	Paper-Pencil	
				N	%
ELA	3	23,769	23,734	35	0.1
	4	23,783	23,743	40	0.2
	5	22,198	22,170	28	0.1
	6	23,231	23,200	31	0.1
	7	22,870	22,826	44	0.2
	8	23,165	23,102	63	0.3
Mathematics	3	23,740	23,706	34	0.1
	4	23,734	23,694	40	0.2
	5	22,154	22,125	29	0.1
	6	23,189	23,158	31	0.1
	7	22,806	22,754	52	0.2
	8	23,096	23,029	67	0.3
Science	5	22,136	22,135	1	0.0
	8	23,043	23,026	17	0.1

7.3. Testing Time

Table 7.8, Table 7.9, and Table 7.10 present the number of minutes students took to complete the Spring 2018 NSCAS ELA, Mathematics, and Science assessments, respectively. Specifically, the tables present the number and percent of students who completed the tests in various time ranges. As shown in the tables, most students completed the ELA test in 40–120 minutes, the Mathematics test in 20–90 minutes, and the Science test in 10–60 minutes. Most students finished tests within 120 minutes (83.16–91.89% for ELA, 91.99–96.91% for Mathematics, and 99.61–99.68% for Science). The percentage of students who took more than 180 minutes is less than 3% (the highest percentage is 2.67% for ELA, 0.91% for Mathematics, and 0.07% for Science).

Table 7.8. Testing Time in Minutes—ELA

Time in Minutes	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	N	%	N	%	N	%	N	%	N	%	N	%
<10	12	0.1	10	0.0	9	0.0	20	0.1	29	0.1	33	0.1
10 – <20	107	0.5	78	0.3	40	0.2	74	0.3	113	0.5	133	0.6
20 – <30	347	1.5	233	1.0	183	0.8	218	0.9	326	1.4	311	1.3
30 – <40	934	3.9	711	3.0	503	2.3	543	2.3	864	3.8	801	3.5
40 – <50	1,905	8.0	1,465	6.2	1,182	5.3	1,244	5.4	1,863	8.2	1,687	7.3
50 – <60	2,681	11.3	2,416	10.2	1,963	8.9	2,188	9.4	3,127	13.7	2,717	11.8
60 – <70	3,108	13.1	2,961	12.5	2,739	12.4	2,902	12.5	3,666	16.0	3,516	15.2
70 – <80	3,110	13.1	3,209	13.5	3,011	13.6	3,282	14.1	3,534	15.5	3,630	15.7
80 – <90	2,860	12.0	2,974	12.5	2,871	12.9	3,074	13.2	2,906	12.7	3,149	13.6
90 – <100	2,291	9.7	2,550	10.7	2,483	11.2	2,662	11.5	2,070	9.1	2,297	9.9
100 – <110	1,725	7.3	2,056	8.7	1,960	8.8	2,059	8.9	1,482	6.5	1,633	7.1
110 – <120	1,304	5.5	1,466	6.2	1,497	6.8	1,494	6.4	1,012	4.4	1,142	4.9
120 – <130	932	3.9	1,034	4.4	1,033	4.7	1,053	4.5	645	2.8	681	2.9
130 – <140	638	2.7	745	3.1	716	3.2	671	2.9	438	1.9	450	1.9
140 – <150	441	1.9	536	2.3	527	2.4	485	2.1	238	1.0	302	1.3
150 – <160	333	1.4	388	1.6	374	1.7	343	1.5	164	0.7	214	0.9
160 – <170	265	1.1	251	1.1	295	1.3	234	1.0	116	0.5	129	0.6
170 – <180	176	0.7	172	0.7	197	0.9	171	0.7	88	0.4	85	0.4
>=180	570	2.4	493	2.1	591	2.7	494	2.1	163	0.7	204	0.9
Total	**23,739**	**100.0**	**23,748**	**100.0**	**22,174**	**100.0**	**23,211**	**100.0**	**22,844**	**100.0**	**23,114**	**100.0**

Table 7.9. Testing Time in Minutes—Mathematics

Time in Minutes	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	N	%	N	%	N	%	N	%	N	%	N	%
<10	11	0.0	7	0.0	4	0.0	17	0.1	31	0.1	44	0.2
10 – <20	131	0.6	103	0.4	59	0.3	72	0.3	125	0.5	169	0.7
20 – <30	1,014	4.3	523	2.2	470	2.1	264	1.1	443	1.9	532	2.3
30 – <40	3,298	13.9	2,156	9.1	2,226	10.1	908	3.9	1,464	6.4	1,838	8.0
40 – <50	5,033	21.2	3,940	16.6	4,045	18.3	2,181	9.4	3,174	13.9	3,597	15.6
50 – <60	4,476	18.9	4,314	18.2	4,419	20.0	3,489	15.1	4,232	18.6	4,614	20.0
60 – <70	3,263	13.8	3,687	15.6	3,469	15.7	3,854	16.6	4,085	17.9	4,060	17.6
70 – <80	2,175	9.2	2,839	12.0	2,510	11.3	3,497	15.1	3,127	13.7	3,011	13.1
80 – <90	1,463	6.2	2,021	8.5	1,685	7.6	2,662	11.5	2,148	9.4	1,870	8.1
90 – <100	918	3.9	1,423	6.0	1,082	4.9	2,004	8.7	1,394	6.1	1,252	5.4
100 – <110	583	2.5	895	3.8	736	3.3	1,424	6.1	917	4.0	742	3.2
110 – <120	373	1.6	588	2.5	479	2.2	937	4.0	578	2.5	468	2.0
120 – <130	277	1.2	373	1.6	285	1.3	595	2.6	351	1.5	298	1.3
130 – <140	174	0.7	259	1.1	207	0.9	357	1.5	248	1.1	182	0.8
140 – <150	135	0.6	178	0.8	134	0.6	262	1.1	143	0.6	129	0.6
150 – <160	92	0.4	104	0.4	88	0.4	196	0.8	105	0.5	73	0.3
160 – <170	87	0.4	85	0.4	68	0.3	160	0.7	64	0.3	50	0.2
170 – <180	52	0.2	58	0.2	37	0.2	74	0.3	44	0.2	29	0.1
>=180	153	0.6	146	0.6	132	0.6	211	0.9	94	0.4	81	0.4
Total	**23,708**	**100.0**	**23,699**	**100.0**	**22,135**	**100.0**	**23,164**	**100.0**	**22,767**	**100.0**	**23,039**	**100.0**

Table 7.10. Testing Time in Minutes—Science

Time in Minutes	Grade 5		Grade 8	
	N	%	N	%
<10	15	0.1	59	0.3
10 – <20	1,349	6.1	1,303	5.7
20 – <30	6,388	28.9	6,935	30.1
30 – <40	6,340	28.6	7,143	31.0
40 – <50	3,895	17.6	3,837	16.7
50 – <60	1,933	8.7	1,835	8.0
60 – <70	1,026	4.6	906	3.9
70 – <80	504	2.3	457	2.0
80 – <90	279	1.3	240	1.0
90 – <100	176	0.8	126	0.5
100 – <110	91	0.4	64	0.3
110 – <120	56	0.3	59	0.3
120 – <130	30	0.1	20	0.1
130 – <140	19	0.1	21	0.1
140 – <150	9	0.0	6	0.0
150 – <160	4	0.0	6	0.0
160 – <170	6	0.0	5	0.0
170 – <180	3	0.0	3	0.0
>=180	16	0.1	13	0.1
Total	**22,139**	**100.0**	**23,038**	**100.0**

7.4. Achievement Level Distributions

Table 7.11 presents the achievement level distributions for the Spring 2018 NSCAS Summative Assessments. Appendix L provides the achievement level distributions by demographic group. For ELA, 43–52% of students are at Developing and 48–57% of students are at On Track or CCR Benchmark. For Mathematics, 45–50% of students are at Developing and 50–55% of students are at On Track or CCR Benchmark. For Science, 30–33% of students are at Below the Standards and 67–70% are at Meets or Exceeds the Standards.

Table 7.11. Achievement Level Distributions

	Total N-	Level 3*		Level 2*		Level 1*		Level 2 + Level 1	
Grade	Count	N-Count	%	N-Count	%	N-Count	%	N-Count	%
ELA									
3	23,769	11,083	46.6	8,888	37.4	3,798	16.0	12,686	53.4
4	23,783	10,312	43.4	9,640	40.5	3,831	16.1	13,471	56.6
5	22,198	10,772	48.5	7,849	35.4	3,577	16.1	11,426	51.5
6	23,231	12,165	52.4	7,063	30.4	4,003	17.2	11,066	47.6
7	22,870	11,977	52.4	7,485	32.7	3,408	14.9	10,893	47.6
8	23,165	11,318	48.9	8,612	37.2	3,235	14.0	11,847	51.1
Mathematics									
3	23,740	11,891	50.1	9,395	39.6	2,454	10.3	11,849	49.9
4	23,734	11,901	50.1	9,358	39.4	2,475	10.4	11,833	49.9
5	22,154	10,937	49.4	9,119	41.2	2,098	9.5	11,217	50.6
6	23,189	10,462	45.1	10,338	44.6	2,389	10.3	12,727	54.9
7	22,806	11,513	50.5	8,956	39.3	2,337	10.2	11,293	49.5
8	23,096	11,401	49.4	9,503	41.1	2,192	9.5	11,695	50.6
Science									
5	22,136	6,661	30.1	12,042	54.4	3,433	15.5	15,475	69.9
8	23,043	7,585	32.9	10,960	47.6	4,498	19.5	15,458	67.1

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

7.5. Descriptive Statistics of Scale Scores

Table 7.12 presents the descriptive statistics for the scale scores, including the mean, standard deviation (SD), and scores at the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles. Appendix L also presents the descriptive statistics by demographic group. As expected, the mean increases with the grade for ELA and Mathematics.

Table 7.12. Scale Score Descriptive Statistics

					Percentiles						
Content Area	Grade	N-Count	Mean	SD	P5	P10	P25	P50	P75	P90	P95
ELA	3	23,769	2481.31	76.47	2357	2380	2426	2483	2535	2577	2606
	4	23,783	2511.56	71.98	2393	2417	2462	2512	2560	2604	2629
	5	22,198	2531.34	66.88	2422	2443	2483	2533	2578	2616	2640
	6	23,231	2538.47	66.68	2429	2450	2493	2538	2584	2624	2647
	7	22,870	2550.49	73.79	2433	2454	2495	2551	2603	2648	2672
	8	23,165	2560.89	66.26	2448	2475	2516	2562	2606	2648	2665

					Percentiles						
Content Area	Grade	N-Count	Mean	SD	P5	P10	P25	P50	P75	P90	P95
Mathematics	3	23,740	1192.17	71.13	1077.5	1102	1144	1189	1239	1288	1312
	4	23,734	1227.00	66.92	1127	1145	1177	1221	1271	1319	1344
	5	22,154	1241.60	66.10	1143	1163	1196	1236	1280	1328	1359
	6	23,189	1253.86	72.33	1137	1163	1203	1253	1301	1350	1373
	7	22,806	1254.73	67.33	1159	1175	1207	1246	1296	1348	1380
	8	23,096	1270.73	71.32	1167	1182	1217	1265	1319	1362	1402
Science	5	22,136	102.13	32.93	50	59	80	101	123	144	159
	8	23,043	102.59	35.99	45	56	76	102	127	150	163

7.6. Reporting Category Correlations

For each grade and content area, Pearson’s correlation coefficients between reporting categories were calculated between reporting category scores to provide information on score dimensionality, which is part of validity evidence based on the tests’ internal structure. Disattenuated correlations provide an estimate of the relationships between reporting categories if there is no measurement error. Table 7.13 – Table 7.18 provide the reporting category correlations, and Table 7.19 – Table 7.24 present the disattenuated correlations.

The correlations between reporting categories within the content areas are positive and moderate in value, ranging from .52 (between Geometry and Data for Grade 4) to .76 (between Number and Algebra for Grade 5). The correlations between reporting categories across the content areas are positive and low to moderate in value, ranging from .47 (between Writing Skills and Data for Grade 8) and .67 (between Writing Skills and Life Science for Grade 8). In general, the within-content-area reporting category correlations are higher than the across-content-area reporting category correlations.

The disattenuated correlation are higher than the correlations, which is expected given that none of the reporting categories has perfect reliabilities (see Table 8.7 – Table 8.9). The disattenuated correlations between reporting categories within the content areas are positive and high in value: .88 (between Reading Vocabulary and Writing Skills for Grade 6) or higher. The disattenuated correlations between reporting categories across the content areas are positive and moderate in value, ranging from .75 (between Number and Life Science for Grade 5, Number and Earth/Space Sciences for Grade 5, and Reading Comprehension and Geometry for Grade 7) or higher. The high disattenuated correlations within the content suggest that reporting categories might be measuring essentially the same construct, which is one evidence based on internal structure. In other words, the internal structure of the assessments is consistent with the structure of the content standards.

Table 7.13. Reporting Category Correlations—Grade 3

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00	–	–				
Reading Comprehension	0.74	1.00					
Writing Skills	0.61	0.67	1.00				

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Number	0.63	0.64	0.53	1.00			
Algebra	0.59	0.61	0.50	0.69	1.00		
Geometry	0.59	0.60	0.49	0.70	0.61	1.00	
Data	0.65	0.67	0.55	0.74	0.67	0.66	1.00

Table 7.14. Reporting Category Correlations—Grade 4

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00						
Reading Comprehension	0.71	1.00					
Writing Skills	0.56	0.66	1.00				
Number	0.55	0.63	0.53	1.00			
Algebra	0.57	0.65	0.54	0.72	1.00		
Geometry	0.52	0.57	0.47	0.66	0.61	1.00	
Data	0.48	0.54	0.45	0.61	0.60	0.52	1.00

Table 7.15. Reporting Category Correlations—Grade 5

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data	Inquiry, Nature of Science, & Tech	Physical Science	Life Science	Earth/ Space Sciences
Reading Vocabulary	1.00										
Reading Comprehension	0.66	1.00									
Writing Skills	0.61	0.69	1.00								
Number	0.52	0.58	0.56	1.00							
Algebra	0.52	0.58	0.56	0.73	1.00						
Geometry	0.50	0.55	0.51	0.61	0.59	1.00					
Data	0.52	0.57	0.54	0.61	0.60	0.55	1.00				
Inquiry, Nature of Science, & Tech	0.58	0.63	0.59	0.57	0.57	0.55	0.57	1.00			
Physical Science	0.55	0.58	0.56	0.54	0.55	0.53	0.52	0.60	1.00		
Life Science	0.57	0.61	0.57	0.51	0.52	0.52	0.52	0.62	0.63	1.00	
Earth/Space Sciences	0.55	0.58	0.55	0.52	0.53	0.52	0.52	0.60	0.63	0.64	1.00

Table 7.16. Reporting Category Correlations—Grade 6

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00						
Reading Comprehension	0.70	1.00					
Writing Skills	0.54	0.63	1.00				
Number	0.58	0.64	0.54	1.00			
Algebra	0.58	0.64	0.54	0.76	1.00		
Geometry	0.53	0.60	0.50	0.70	0.70	1.00	
Data	0.52	0.58	0.49	0.67	0.66	0.63	1.00

Table 7.17. Reporting Category Correlations—Grade 7

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00						
Reading Comprehension	0.69	1.00					
Writing Skills	0.62	0.72	1.00				
Number	0.51	0.59	0.56	1.00			
Algebra	0.56	0.65	0.61	0.72	1.00		
Geometry	0.48	0.55	0.53	0.62	0.65	1.00	
Data	0.51	0.59	0.56	0.67	0.70	0.63	1.00

Table 7.18. Reporting Category Correlations—Grade 8

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data	Inquiry, Nature of Science, & Tech	Physical Science	Life Science	Earth/ Space Sciences
Reading Vocabulary	1.00										
Reading Comprehension	0.69	1.00									
Writing Skills	0.58	0.68	1.00								
Number	0.52	0.60	0.55	1.00							
Algebra	0.56	0.65	0.58	0.74	1.00						
Geometry	0.53	0.61	0.55	0.71	0.74	1.00					
Data	0.47	0.53	0.47	0.59	0.61	0.59	1.00				
Inquiry, Nature of Science, & Tech	0.59	0.66	0.57	0.58	0.63	0.61	0.52	1.00			
Physical Science	0.57	0.62	0.55	0.58	0.62	0.60	0.52	0.67	1.00		
Life Science	0.61	0.67	0.58	0.57	0.60	0.59	0.51	0.69	0.71	1.00	
Earth/Space Sciences	0.58	0.64	0.56	0.58	0.61	0.60	0.53	0.67	0.69	0.72	1.00

Table 7.19. Reporting Category Disattenuated Correlations—Grade 3

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00						
Reading Comprehension	0.97	1.00					
Writing Skills	0.97	0.96	1.00				
Number	0.85	0.77	0.78	1.00			
Algebra	0.94	0.87	0.87	1.00	1.00		
Geometry	0.91	0.83	0.82	0.99	1.00	1.00	
Data	0.93	0.86	0.86	0.97	1.00	0.99	1.00

Table 7.20. Reporting Category Disattenuated Correlations—Grade 4

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00						
Reading Comprehension	0.99	1.00					
Writing Skills	0.92	0.92	1.00				
Number	0.79	0.76	0.76	1.00			
Algebra	0.88	0.85	0.84	0.97	1.00		
Geometry	0.85	0.79	0.77	0.94	0.94	1.00	
Data	0.87	0.83	0.82	0.96	1.00	0.94	1.00

Table 7.21. Reporting Category Disattenuated Correlations—Grade 5

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data	Inquiry, Nature of Science, & Tech	Physical Science	Life Science	Earth/ Space Sciences
Reading Vocabulary	1.00										
Reading Comprehension	0.95	1.00									
Writing Skills	0.95	0.94	1.00								
Number	0.75	0.73	0.77	1.00							
Algebra	0.80	0.77	0.81	0.98	1.00						
Geometry	0.87	0.83	0.84	0.93	0.96	1.00					
Data	1.00	1.00	1.00	1.00	1.00	1.00	1.00				
Inquiry, Nature of Science, & Tech	1.00	0.96	0.98	0.88	0.93	1.00	1.00	1.00			
Physical Science	0.90	0.83	0.87	0.78	0.84	0.92	1.00	1.00	1.00		
Life Science	0.96	0.89	0.90	0.75	0.81	0.92	1.00	1.00	1.00	1.00	
Earth/Space Sciences	0.91	0.83	0.86	0.75	0.81	0.91	1.00	1.00	1.00	1.00	1.00

Table 7.22. Reporting Category Disattenuated Correlations—Grade 6

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00						
Reading Comprehension	0.97	1.00					
Writing Skills	0.88	0.89	1.00				
Number	0.84	0.80	0.79	1.00			
Algebra	0.84	0.80	0.79	0.99	1.00		
Geometry	0.80	0.78	0.76	0.95	0.95	1.00	
Data	0.84	0.81	0.81	0.98	0.97	0.96	1.00

Table 7.23. Reporting Category Disattenuated Correlations—Grade 7

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data
Reading Vocabulary	1.00						
Reading Comprehension	0.97	1.00					
Writing Skills	0.98	0.98	1.00				
Number	0.81	0.80	0.86	1.00			
Algebra	0.81	0.81	0.85	1.00	1.00		
Geometry	0.76	0.75	0.81	0.95	0.91	1.00	
Data	0.78	0.77	0.82	0.98	0.94	0.93	1.00

Table 7.24. Reporting Category Disattenuated Correlations—Grade 8

	Reading Vocabulary	Reading Comprehension	Writing Skills	Number	Algebra	Geometry	Data	Inquiry, Nature of Science, & Tech	Physical Science	Life Science	Earth/ Space Sciences
Reading Vocabulary	1.00										
Reading Comprehension	1.00	1.00									
Writing Skills	0.95	0.94	1.00								
Number	0.82	0.79	0.82	1.00							
Algebra	0.84	0.82	0.82	1.00	1.00						
Geometry	0.80	0.78	0.79	0.98	0.97	1.00					
Data	1.00	0.95	0.95	1.00	1.00	1.00	1.00				
Inquiry, Nature of Science, & Tech	0.99	0.93	0.90	0.88	0.91	0.89	1.00	1.00			
Physical Science	0.90	0.82	0.82	0.83	0.84	0.82	1.00	1.00	1.00		
Life Science	0.97	0.89	0.87	0.82	0.82	0.82	0.99	1.00	1.00	1.00	
Earth/Space Sciences	0.94	0.87	0.86	0.85	0.85	0.85	1.00	1.00	1.00	1.00	1.00

7.7. Correlations with MAP Growth

NWEA conducted a linking study in November 2018 using Spring 2018 data that produced a set of MAP Growth Reading and Mathematics Rasch Unit (RIT) cut scores that correspond to the NSCAS Summative ELA and Mathematics scale scores (NWEA, 2018c). This linking study reported correlations of NSCAS and MAP Growth scores from the linking study sample who took both tests, as shown in Table 7.25. The correlation coefficients between MAP Growth and NSCAS scores range from 0.81 to 0.83 for ELA/Reading and 0.85 to 0.87 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 7.25. Correlation and Descriptive Statistics of NSCAS and MAP Growth Scores

Content Area	Grade	N	*r*	NSCAS*				MAP Growth*			
				Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
ELA	3	15,276	0.82	2476	76.24	2222	2832	199	15.53	142	246
	4	14,919	0.83	2508	72.16	2252	2826	207	15.29	140	250
	5	13,669	0.82	2528	67.13	2282	2833	213	15.01	140	256
	6	13,947	0.82	2537	66.52	2292	2790	217	15.17	149	264
	7	13,027	0.81	2550	73.25	2328	2862	220	15.61	139	264
	8	12,887	0.82	2559	66.55	2312	2873	223	16.18	147	268
Mathematics	3	15,182	0.87	1190	71.23	1002	1428	204	14.01	134	253
	4	14,737	0.85	1225	67.47	1040	1491	214	15.44	139	278
	5	13,673	0.86	1239	65.1	1022	1482	222	16.95	139	299
	6	14,026	0.87	1252	72.04	1038	1488	226	16.47	134	277
	7	13,356	0.85	1254	66.79	1065	1540	231	17.92	136	310
	8	13,050	0.86	1270	71.56	1069	1545	236	19.32	136	316

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

Section 8: Reliability

The *Standards* refers to reliability as the “consistency of scores across replications of a testing procedure” (AERA et al., 2014, p. 33). The level of reliability/precision of scores has implications for validity. In other words, scores must be consistent and precise enough to be useful for intended purposes. If scores are to be meaningful, tests should produce stable scores if the same group of students were to take the same test repeatedly without any fatigue or memory of the test. In addition, the range of certainty around the score should be small enough to support educational decisions. The reliability/precision of the 2018 NSCAS assessments was examined through analysis of measurement error in simulated and operational conditions, as follows:

- Constraint engine score precision and reliability
- Marginal reliability
- Conditional standard error of measurement (CSEM)
- Cronbach’s alpha and standard error of measurement (SEM) for fixed forms

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

8.1. Constraint Engine Score Precision and Reliability

The pre-administration constraint engine evaluation using simulations provided precision ability estimations that showed how well the engine recovered students’ true ability based on the item pool (NWEA, 2018a). Both the pre- and post-administration constraint engine evaluation studies included the standard deviation of estimated theta, mean SEM, and marginal reliability (NWEA, 2018a, 2018b). This section provides results from both studies for comparison purposes.

8.1.1. Pre-Administration Engine Evaluation

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

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

8.1.1.1. Evaluation Criteria

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

$$bias = N^{-1} \sum_{i=1}^{N} \left(\theta_i - \hat{\theta}_i\right) \tag{8.1}$$

$$MSE = N^{-1} \sum_{i=1}^{N} \left(\theta_i - \hat{\theta}_i\right)^2 \tag{8.2}$$

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

$$var(bias) = \sigma^2 * g'\left(\hat{\theta}_i\right)^2 = \frac{1}{N(N-1)} \sum_{i=1}^{N} \left(\theta_i - \hat{\hat{\theta}}_i\right)^2 \tag{8.3}$$

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

$$Z = bias/\sqrt{var(bias)} \tag{8.4}$$

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:

$$Mean(se) = \sqrt{N^{-1} \sum_{i=1}^{N} se\left(\hat{\theta}_i\right)^2} \tag{8.5}$$

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

$$t = \frac{\theta_i - \hat{\theta}_i}{se(\hat{\theta}_i)} \tag{8.6}$$

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

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

$$\text{Marginal Reliability} = \frac{var\,(\hat{\theta}\,) - \sigma^2}{var\,(\hat{\theta}\,)} \tag{8.7}$$

where σ is defined as:

$$\sigma = \mathrm{E}\left\{[I(\theta)]^{-1/2}\right\} \tag{8.8}$$

8.1.1.2. Student Sample

One thousand students per grade were included in the simulation study sample. The true values of student ability (theta θ) were drawn from normal distribution with a mean of 0.0 and a standard deviation (SD) of 1. Table 8.1 presents the mean of the true values of the students' ability distribution for each simulation. The student sample also had similar demographic characteristics to Nebraska's general population based on the November roster received from the NDE.

Table 8.1. Mean of True Values of Simulation Population's Ability

	ELA		Mathematics	
Grade	Mean	SD	Mean	SD
3	0.01	0.99	-0.04	1.01
4	-0.02	0.96	-0.03	1.02
5	0.04	0.99	-0.01	1.02
6	0.00	0.99	0.00	0.99
7	0.00	1.04	0.01	0.98
8	0.01	1.03	0.00	1.01

8.1.1.3. Precision Ability Estimation Results

Table 8.2 presents the results of the precision ability estimation. The mean biases across all students are small, ranging from -0.02 to 0.01 for both ELA and Mathematics. The *p*-value supports the null-hypothesis that there is not a significant difference between the simulated students' true and final estimated thetas. The MSE is also relatively small, showing that the constraint engine recovered student's true theta.

Table 8.2. Mean Bias of the Ability Estimation (True - Estimated)

	Bias		*P*-Value for		95%	99%
Grade	Mean	SE	Z-Test	MSE	Coverage	Coverage
ELA						
3	0.00	0.01	0.94	0.11	5.20	1.20
4	0.00	0.01	0.89	0.11	6.20	1.10
5	0.00	0.01	0.97	0.11	4.90	1.40
6	0.01	0.01	0.79	0.11	4.90	1.00
7	-0.01	0.01	0.88	0.12	5.80	1.50
8	0.01	0.01	0.87	0.10	6.00	1.40
Mathematics						
3	-0.02	0.01	0.51	0.16	4.70	1.00
4	-0.01	0.01	0.80	0.17	5.10	1.30
5	0.01	0.01	0.85	0.16	4.50	0.30
6	-0.01	0.01	0.65	0.13	5.20	0.70
7	0.00	0.01	0.75	0.13	5.00	1.00
8	0.01	0.01	0.83	0.16	5.40	0.80

8.1.1.4. Score Precision and Reliability Results

Table 8.5 presents the pre-administration score precision and reliability estimates, including the average number of items administered, the standard deviation (SD) of the estimated theta, the mean SEM, the RMSE, and a marginal reliability coefficient. The SD, mean SEM, and RMSE are relatively small, and the range of the marginal reliability is from 0.90 to 0.92 for ELA and 0.86 to 0.88 for Mathematics. These results indicate that, overall, the score precision is relatively good.

Table 8.3. Pre-Administration Score Precision and Reliability

Grade	Average #Items	SD of Estimated Theta	Mean SEM	RMSE	Reliability
ELA					
3	41	1.02	0.32	0.32	0.90
4	41	1.03	0.32	0.32	0.90
5	41	1.06	0.32	0.32	0.91
6	41	1.06	0.31	0.31	0.91
7	41	1.10	0.32	0.33	0.91
8	41	1.10	0.31	0.32	0.92
Mathematics					
3	41	1.13	0.40	0.41	0.87
4	41	1.12	0.38	0.39	0.88
5	41	1.06	0.40	0.40	0.86
6	41	1.05	0.36	0.36	0.88
7	41	1.08	0.36	0.37	0.89
8	41	1.11	0.39	0.39	0.87

Table 8.6 presents the average SEM by decile of the true overall proficiency score, including the overall student ability distribution. A decile is similar to a percentile rank, with 10 ranks related to the 10th, 20th…90th, 100th percentile ranks. The average SEM is similar across deciles except Decile 1 and Decile 10 that have a higher standard error compared to the other deciles. Overall, the SEM is in acceptable ranges from 0.31 to 0.40. These indexes are comparable to what the state obtained historically through its fixed-form assessments (NWEA, 2018a).

Table 8.4. Pre-Administration SEM by Deciles

	Proficiency Score Distribution										
Grade	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	Overall
ELA											
3	0.37	0.33	0.32	0.31	0.30	0.30	0.30	0.30	0.32	0.35	0.32
4	0.38	0.33	0.31	0.30	0.30	0.30	0.30	0.30	0.31	0.35	0.32
5	0.39	0.34	0.32	0.31	0.3	0.29	0.29	0.29	0.3	0.35	0.32
6	0.38	0.32	0.30	0.30	0.29	0.28	0.28	0.29	0.30	0.35	0.31
7	0.38	0.33	0.31	0.3	0.29	0.3	0.3	0.31	0.33	0.41	0.32
8	0.38	0.33	0.31	0.30	0.29	0.29	0.29	0.30	0.31	0.35	0.31
Mathematics											
3	0.38	0.36	0.36	0.37	0.38	0.39	0.40	0.41	0.45	0.54	0.40
4	0.39	0.36	0.35	0.35	0.35	0.36	0.37	0.39	0.42	0.51	0.38
5	0.40	0.37	0.37	0.37	0.38	0.38	0.39	0.41	0.42	0.49	0.40
6	0.36	0.34	0.34	0.34	0.34	0.34	0.35	0.36	0.38	0.46	0.36
7	0.41	0.36	0.35	0.34	0.34	0.34	0.34	0.35	0.37	0.43	0.36
8	0.40	0.36	0.35	0.35	0.36	0.36	0.37	0.39	0.42	0.52	0.39

8.1.2. Post-Administration Engine Evaluation

Table 8.5 presents the post-administration score precision and reliability estimates, including the average number of items administered, the SD of the estimated theta, the mean SEM, and a marginal reliability coefficient. The SD and mean SEM are relatively small, and the range of the marginal reliability is from 0.89 to 0.91 for ELA and 0.89 to 0.92 for Mathematics. Although the reliability coefficients have slightly decreased compared to the simulation results (e.g., the reliability of ELA Grade 5 decreased by 0.2 from 0.91 to 0.89). the magnitude of changes is relatively small. Overall, the score precision is still in a satisfactory range.

Table 8.5. Post-Administration Score Precision and Reliability

Grade	Average #Items	SD of Estimated Theta	Mean SEM	Reliability
ELA				
3	41	1.05	0.32	0.91
4	41	1.03	0.32	0.90
5	41	0.93	0.31	0.89
6	41	0.96	0.30	0.90
7	41	1.00	0.32	0.90
8	41	0.96	0.31	0.90
Mathematics				
3	41	1.42	0.42	0.91
4	41	1.28	0.40	0.89
5	41	1.33	0.43	0.89
6	41	1.36	0.38	0.92
7	41	1.25	0.37	0.91
8	41	1.39	0.42	0.90

Table 8.6 presents the average SEM by decile of the true overall proficiency score, including the overall student ability distribution. The average SEM is similar across deciles except Decile 1 and Decile 10 that have a higher standard error compared to the other deciles. Overall, the SEM is in acceptable ranges.

Table 8.6. Post-Administration SEM by Deciles

	Proficiency Score Distribution										
Grade	**Decile 1**	**Decile 2**	**Decile 3**	**Decile 4**	**Decile 5**	**Decile 6**	**Decile 7**	**Decile 8**	**Decile 9**	**Decile 10**	**Overall**
ELA											
3	0.37	0.33	0.31	0.30	0.30	0.30	0.30	0.31	0.32	0.36	0.32
4	0.36	0.32	0.30	0.30	0.29	0.29	0.30	0.30	0.32	0.37	0.32
5	0.36	0.32	0.31	0.30	0.29	0.29	0.29	0.29	0.31	0.36	0.31
6	0.34	0.31	0.29	0.28	0.28	0.28	0.28	0.29	0.31	0.36	0.30
7	0.36	0.31	0.30	0.29	0.29	0.29	0.30	0.32	0.34	0.41	0.32
8	0.35	0.31	0.30	0.29	0.29	0.29	0.30	0.31	0.32	0.35	0.31
Mathematics											
3	0.41	0.37	0.36	0.36	0.37	0.39	0.40	0.42	0.47	0.66	0.42
4	0.38	0.35	0.35	0.35	0.35	0.37	0.39	0.42	0.46	0.63	0.40
5	0.40	0.37	0.37	0.37	0.38	0.39	0.41	0.43	0.48	0.70	0.43
6	0.37	0.34	0.34	0.33	0.34	0.34	0.35	0.37	0.42	0.60	0.38
7	0.41	0.36	0.35	0.34	0.33	0.34	0.34	0.36	0.39	0.51	0.37
8	0.40	0.36	0.35	0.35	0.36	0.37	0.39	0.41	0.47	0.73	0.42

8.2. 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 the previous section. Table 8.7, Table 8.8, and Table 8.9 present marginal reliabilities of scale scores by grade and reporting category for ELA, Mathematics, and Science, respectively. Marginal reliability estimates for the total scores are well above 0.80 (.871 or higher), which is typically considered the minimally acceptable level of reliability. Because reliability for reporting categories are based on fewer items, they have lower reliability than total scores. Appendix M provides marginal reliability estimates for the total scores by demographic sub-group.

As shown in Table 8.10, reliability varies by overall score levels (i.e., deciles). Observed variance is from total score, and error variance is calculated for each decile. All students take the same number of items, but the information delivered by the items differs. The most information, and hence lower error and higher reliability, is found where the pool has the most items. The NSCAS item pools have more items in the middle than the both end and are easy relative to the population, resulting in lower reliability with higher scores (Deciles 9 and 10).

Table 8.7. Marginal Reliability of Scale Scores—ELA

Grade	N	Total Score	Reading Vocabulary	Reading Comprehension	Writing Skills
3	23,774	0.90	0.68	0.85	0.58
4	23,789	0.90	0.61	0.85	0.61
5	22,202	0.88	0.60	0.80	0.68
6	23,242	0.89	0.62	0.83	0.61
7	22,890	0.90	0.61	0.82	0.65
8	23,178	0.89	0.58	0.82	0.65

Table 8.8. Marginal Reliability of Scale Scores—Mathematics

Grade	N	Total Score	Number	Algebra	Geometry	Data
3	23,744	0.92	0.81	0.58	0.62	0.72
4	23,739	0.90	0.81	0.69	0.61	0.50
5	22,164	0.90	0.79	0.70	0.54	0.37
6	23,195	0.92	0.77	0.76	0.71	0.61
7	22,820	0.91	0.66	0.78	0.65	0.71
8	23,107	0.91	0.70	0.77	0.75	0.38

Table 8.9. Marginal Reliability of Scale Scores—Science

Grade	N	Total Score	Inquiry, Nature of Science, & Tech	Physical Science	Life Science	Earth/Space Sciences
5	22,140	0.87	0.54	0.62	0.59	0.61
8	23,056	0.91	0.62	0.70	0.69	0.66

Table 8.10. Marginal Reliability: Variance

Content Area	Grade	N	Variance	Overall	Deciles									
					1	2	3	4	5	6	7	8	9	10
ELA	3	23,774	5848.41	0.90	0.88	0.90	0.91	0.92	0.92	0.92	0.92	0.91	0.90	0.86
	4	23,789	5183.83	0.90	0.87	0.90	0.91	0.91	0.91	0.91	0.91	0.91	0.89	0.84
	5	22,202	4475.86	0.88	0.85	0.88	0.89	0.90	0.90	0.90	0.90	0.89	0.88	0.82
	6	23,242	4454.16	0.89	0.86	0.89	0.90	0.91	0.91	0.91	0.91	0.90	0.89	0.83
	7	22,890	5459.73	0.90	0.88	0.91	0.91	0.92	0.92	0.91	0.91	0.90	0.88	0.81
	8	23,178	4397.82	0.89	0.86	0.89	0.90	0.90	0.90	0.90	0.90	0.89	0.88	0.84
Mathematics	3	23,744	5062.10	0.92	0.92	0.93	0.94	0.94	0.94	0.94	0.93	0.92	0.91	0.80
	4	23,739	4486.11	0.90	0.91	0.93	0.93	0.93	0.93	0.93	0.92	0.91	0.89	0.77
	5	22,164	4382.85	0.90	0.92	0.93	0.93	0.93	0.93	0.93	0.92	0.91	0.89	0.71
	6	23,195	5237.46	0.92	0.93	0.94	0.94	0.94	0.94	0.94	0.93	0.93	0.91	0.80
	7	22,820	4546.53	0.91	0.91	0.92	0.93	0.93	0.93	0.93	0.93	0.92	0.91	0.80
	8	23,107	5099.24	0.91	0.92	0.93	0.94	0.94	0.94	0.94	0.93	0.92	0.90	0.76
Science	5	22,140	1085.46	0.87	0.90	0.91	0.91	0.91	0.91	0.90	0.88	0.87	0.83	0.67
	8	23,056	1296.22	0.91	0.92	0.94	0.94	0.94	0.94	0.92	0.92	0.91	0.88	0.74

8.3. Conditional Standard Error of Measurement (CSEM)

The CSEM, defined in Section 5.6.4. 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 8.11 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 from the achievement level cut scores; the mean CSEMs of students near the cut; and the standard deviation (SD) of the CSEMs. Science was not included in this table because they are fixed forms, so there is no need to compute CSEM with students ±10 scale score points from the achievement level cut scores.

Table 8.12 then presents the overall and by-decile CSEM. The overall CSEM is slightly higher for ELA (from 21.8 to 23.0) than for Mathematics (from 19.5 to 20.5). The low CSEM for Science is expected as its conversion slope is smaller than ELA or Mathematics. CSEM is also relatively similar between Deciles 2 and 9, while the CSEM tends to be higher at the first and last decile. This suggests that item pools have more items in the middle than the both end and that more difficulty items are needed for both ELA and Mathematics, which is consistent with reliability results. Appendix N presents scatterplots for scale score CSEM by reporting category for each content area and grade.

Table 8.11. CSEMs at the Proficient Cut Scores—ELA and Mathematics

		Developing/On Track Cut			On Track/CCR Benchmark Cut		
Content Area	**Grade**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**
ELA	3	2,448	22.0	0.1	1,650	24.0	0.7
	4	2,702	21.4	0.5	1,949	22.8	0.8
	5	2,655	21.0	0.2	1,823	22.5	0.5
	6	2,698	20.0	0.1	1,774	22.0	0.4
	7	2,304	21.5	0.5	1,473	25.1	0.7
	8	3,204	21.1	0.3	1,807	23.1	0.7
Mathematics	3	2,799	18.0	0.0	1,077	23.5	0.5
	4	2,745	18.0	0.1	1,190	24.5	0.6
	5	2,937	18.0	0.0	726	25.0	0.8
	6	2,747	18.0	0.0	1,293	23.0	1.1
	7	2,828	17.2	0.4	824	22.1	0.9
	8	2,539	18.0	0.0	1,096	25.0	1.0

Table 8.12. Mean CSEMs by Deciles

			Mean CSEM by Decile									
Content Area	**Grade**	**Mean CSEM**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
ELA	3	23.6	26.6	24.0	22.5	22.0	22.0	22.0	22.0	22.6	24.2	28.5
	4	23.0	25.8	22.5	22.0	21.7	21.4	21.4	21.5	22.0	23.4	28.4
	5	22.7	25.6	22.7	22.0	21.5	21.0	21.0	21.1	21.8	22.7	27.7
	6	21.8	24.5	21.8	20.8	20.1	20.0	20.0	20.3	21.2	22.4	27.1
	7	23.7	25.8	22.7	21.9	21.5	21.4	21.6	22.1	23.0	25.2	31.8
	8	22.4	24.6	22.0	21.2	21.0	21.1	21.0	21.2	22.0	23.0	26.5
Mathematics	3	20.1	20.3	18.3	18.0	18.0	18.0	18.0	18.7	19.5	21.5	30.2
	4	20.3	19.7	18.0	18.0	18.0	18.0	18.2	19.0	20.2	22.4	30.6
	5	20.4	19.0	18.0	18.0	18.0	18.0	18.0	18.8	19.8	22.2	33.6
	6	20.0	19.6	18.1	18.0	18.0	18.0	18.0	18.6	19.3	21.3	31.0
	7	19.5	20.6	18.6	18.0	17.7	17.3	17.3	17.9	18.6	20.3	28.5
	8	20.5	20.3	18.6	18.0	18.0	18.0	18.0	19.0	20.1	22.4	33.1
Science	5	11.5	10.4	10.0	10.0	10.0	10.0	10.5	11.3	12.0	13.6	18.5
	8	10.7	10.2	9.0	9.0	9.0	9.0	10.0	10.3	11.0	12.7	18.1

8.4. Classification Accuracy

Classification accuracy 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 utilizing a psychometric model to find true scores corresponding to observed scores. In other words, classification accuracy is a measure of how accurately test scores or sub-scores place students into reporting category levels. 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 8.13, Table 8.14, and Table 8.15 present the classification accuracy results by grade, achievement level, and reporting category. Overall classification accuracy ranges from 0.837 (for ELA Grade 8) to 0.881 (for Mathematics Grade 8). In general, classification accuracy is moderate to high. Considering that the magnitude of classification accuracy is influenced by key features of test design including the number of items, number of cut scores, and the reliability and associated SEM, the classification accuracy for 2018 suggests that accurate level classifications are being made for Nebraska students on the NSCAS assessments. Classification accuracy ranges by achievement level ranges from 0.727 (for ELA Grade On Track) to 0.924 (for ELA Grade 7 Developing). The On Track achievement level has lower accuracy than the other two levels because On Track is the middle category with two adjacent cells, whereas the other two levels have only one adjacent cell.

Table 8.13. Classification Accuracy by Achievement Level and Reporting Category—ELA

Grade	Achievement Level	N	%	Expected Proportion*			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
Overall								
3	Developing	11,083	0.47	0.43	0.04	0.00	0.916	0.852
	On Track	8,888	0.37	0.05	0.29	0.04	0.783	
	CCR Benchmark	3,798	0.16	0.00	0.03	0.13	0.825	
4	Developing	10,313	0.43	0.39	0.04	0.00	0.899	0.846
	On Track	9,640	0.41	0.05	0.32	0.04	0.793	
	CCR Benchmark	3,831	0.16	0.00	0.03	0.14	0.839	
5	Developing	10,773	0.49	0.44	0.04	0.00	0.911	0.839
	On Track	7,849	0.35	0.05	0.27	0.04	0.754	
	CCR Benchmark	3,577	0.16	0.00	0.03	0.13	0.807	
6	Developing	12,168	0.52	0.48	0.05	0.00	0.914	0.842
	On Track	7,063	0.30	0.05	0.22	0.04	0.727	
	CCR Benchmark	4,003	0.17	0.00	0.03	0.14	0.826	

Grade	Achievement Level	N	%	Expected Proportion*			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
7	Developing	11,979	0.52	0.48	0.04	0.00	0.924	0.852
	On Track	7,484	0.33	0.05	0.25	0.03	0.758	
	CCR Benchmark	3,408	0.15	0.00	0.03	0.12	0.805	
8	Developing	11,321	0.49	0.44	0.05	0.00	0.904	0.837
	On Track	8,612	0.37	0.05	0.28	0.04	0.753	
	CCR Benchmark	3,235	0.14	0.00	0.02	0.12	0.821	
Reading Vocabulary								
3	Developing	10,867	0.46	0.40	0.06	0.00	0.871	0.736
	On Track	6,852	0.29	0.08	0.15	0.06	0.503	
	CCR Benchmark	6,049	0.26	0.00	0.05	0.19	0.757	
4	Developing	10,144	0.43	0.36	0.06	0.00	0.841	0.709
	On Track	6,916	0.29	0.09	0.15	0.05	0.519	
	CCR Benchmark	6,722	0.28	0.00	0.07	0.20	0.703	
5	Developing	10,600	0.48	0.40	0.07	0.00	0.843	0.705
	On Track	6,600	0.30	0.09	0.13	0.08	0.441	
	CCR Benchmark	4,998	0.23	0.00	0.04	0.17	0.760	
6	Developing	11,879	0.51	0.43	0.07	0.00	0.843	0.724
	On Track	4,804	0.21	0.07	0.08	0.06	0.391	
	CCR Benchmark	6,547	0.28	0.00	0.05	0.21	0.752	
7	Developing	11,888	0.52	0.44	0.08	0.00	0.837	0.712
	On Track	5,826	0.26	0.07	0.11	0.07	0.447	
	CCR Benchmark	5,153	0.23	0.00	0.05	0.16	0.724	
8	Developing	11,368	0.49	0.40	0.08	0.00	0.823	0.699
	On Track	5,714	0.25	0.08	0.11	0.07	0.429	
	CCR Benchmark	6,084	0.26	0.00	0.06	0.19	0.719	
Reading Comprehension								
3	Developing	11,181	0.47	0.42	0.05	0.00	0.891	0.814
	On Track	8,694	0.37	0.06	0.26	0.05	0.716	
	CCR Benchmark	3,894	0.16	0.00	0.03	0.13	0.811	
4	Developing	10,384	0.44	0.39	0.05	0.00	0.883	0.810
	On Track	9,101	0.38	0.06	0.28	0.05	0.728	
	CCR Benchmark	4,298	0.18	0.00	0.04	0.15	0.801	
5	Developing	10,892	0.49	0.43	0.06	0.00	0.876	0.792
	On Track	7,735	0.35	0.06	0.24	0.05	0.681	
	CCR Benchmark	3,572	0.16	0.00	0.04	0.13	0.776	
6	Developing	11,935	0.51	0.46	0.05	0.00	0.895	0.802
	On Track	6,890	0.30	0.06	0.19	0.05	0.640	
	CCR Benchmark	4,409	0.19	0.00	0.04	0.15	0.800	
7	Developing	12,345	0.54	0.48	0.06	0.00	0.887	0.81
	On Track	7,278	0.32	0.05	0.22	0.05	0.692	
	CCR Benchmark	3,247	0.14	0.00	0.03	0.11	0.782	

Grade	Achievement Level	N	%	Expected Proportion*			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
8	Developing	11,313	0.49	0.43	0.06	0.00	0.883	0.794
	On Track	8,221	0.36	0.07	0.24	0.05	0.670	
	CCR Benchmark	3,634	0.16	0.00	0.03	0.13	0.796	
Writing Skills								
3	Developing	10,061	0.42	0.36	0.06	0.00	0.856	0.706
	On Track	9,728	0.41	0.12	0.22	0.08	0.533	
	CCR Benchmark	3,980	0.17	0.00	0.04	0.13	0.754	
4	Developing	10,889	0.46	0.36	0.09	0.00	0.788	0.706
	On Track	8,328	0.35	0.08	0.20	0.07	0.583	
	CCR Benchmark	4,564	0.19	0.00	0.05	0.14	0.734	
5	Developing	10,814	0.49	0.41	0.07	0.00	0.85	0.734
	On Track	6,847	0.31	0.08	0.16	0.06	0.532	
	CCR Benchmark	4,538	0.20	0.00	0.04	0.16	0.765	
6	Developing	12,866	0.55	0.45	0.10	0.00	0.812	0.721
	On Track	6,729	0.29	0.07	0.15	0.07	0.514	
	CCR Benchmark	3,637	0.16	0.00	0.03	0.12	0.777	
7	Developing	12,382	0.54	0.47	0.07	0.00	0.861	0.757
	On Track	5,847	0.26	0.07	0.14	0.05	0.539	
	CCR Benchmark	4,638	0.20	0.00	0.04	0.15	0.754	
8	Developing	11,685	0.50	0.42	0.08	0.00	0.827	0.727
	On Track	8,367	0.36	0.08	0.21	0.07	0.571	
	CCR Benchmark	3,115	0.13	0.00	0.03	0.10	0.776	

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

Table 8.14. Classification Accuracy by Achievement Level and Reporting Category—Mathematics

Grade	Achievement Level	N	%	Expected Proportion*			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
Overall								
3	Developing	11,893	0.50	0.46	0.04	0.00	0.920	0.876
	On Track	9,395	0.40	0.04	0.33	0.02	0.843	
	CCR Benchmark	2,454	0.10	0.00	0.02	0.08	0.786	
4	Developing	11,905	0.50	0.46	0.04	0.00	0.918	0.874
	On Track	9,358	0.39	0.04	0.33	0.02	0.843	
	CCR Benchmark	2,475	0.10	0.00	0.02	0.08	0.779	
5	Developing	10,944	0.49	0.45	0.04	0.00	0.909	0.868
	On Track	9,119	0.41	0.05	0.34	0.02	0.835	
	CCR Benchmark	2,098	0.10	0.00	0.02	0.08	0.800	
6	Developing	10,465	0.45	0.42	0.04	0.00	0.920	0.879
	On Track	10,338	0.45	0.04	0.38	0.03	0.845	
	CCR Benchmark	2,389	0.10	0.00	0.02	0.09	0.845	

Grade	Achievement Level	N	%	Expected Proportion*			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
7	Developing	11,520	0.51	0.46	0.04	0.00	0.915	0.879
	On Track	8,956	0.39	0.04	0.33	0.02	0.845	
	CCR Benchmark	2,337	0.10	0.00	0.02	0.09	0.833	
8	Developing	11,407	0.49	0.46	0.04	0.00	0.923	0.881
	On Track	9,503	0.41	0.04	0.35	0.03	0.842	
	CCR Benchmark	2,192	0.10	0.00	0.02	0.08	0.832	
Number								
3	Developing	11,926	0.50	0.45	0.06	0.00	0.886	0.816
	On Track	9,186	0.39	0.06	0.28	0.05	0.724	
	CCR Benchmark	2,628	0.11	0.00	0.02	0.09	0.820	
4	Developing	11,715	0.49	0.44	0.06	0.00	0.885	0.813
	On Track	8,483	0.36	0.06	0.27	0.04	0.745	
	CCR Benchmark	3,540	0.15	0.00	0.04	0.11	0.738	
5	Developing	11,036	0.50	0.44	0.06	0.00	0.873	0.81
	On Track	8,394	0.38	0.07	0.28	0.03	0.744	
	CCR Benchmark	2,731	0.12	0.00	0.03	0.09	0.756	
6	Developing	10,969	0.47	0.40	0.07	0.00	0.850	0.790
	On Track	8,838	0.38	0.06	0.28	0.05	0.724	
	CCR Benchmark	3,384	0.15	0.00	0.03	0.11	0.767	
7	Developing	10,772	0.47	0.41	0.06	0.00	0.864	0.766
	On Track	8,181	0.36	0.08	0.24	0.04	0.680	
	CCR Benchmark	3,856	0.17	0.00	0.05	0.11	0.675	
8	Developing	11,089	0.48	0.42	0.06	0.00	0.869	0.784
	On Track	8,579	0.37	0.07	0.26	0.04	0.712	
	CCR Benchmark	3,433	0.15	0.00	0.04	0.10	0.691	
Algebra								
3	Developing	11,706	0.49	0.42	0.08	0.00	0.842	0.733
	On Track	8,216	0.35	0.09	0.21	0.05	0.607	
	CCR Benchmark	3,817	0.16	0.00	0.05	0.11	0.671	
4	Developing	11,628	0.49	0.42	0.07	0.00	0.859	0.770
	On Track	9,417	0.40	0.08	0.26	0.06	0.657	
	CCR Benchmark	2,693	0.11	0.00	0.02	0.09	0.779	
5	Developing	10,922	0.49	0.42	0.07	0.00	0.848	0.767
	On Track	8,006	0.36	0.08	0.24	0.04	0.673	
	CCR Benchmark	3,233	0.15	0.00	0.04	0.11	0.726	
6	Developing	10,677	0.46	0.40	0.06	0.00	0.874	0.798
	On Track	9,944	0.43	0.07	0.31	0.05	0.720	
	CCR Benchmark	2,571	0.11	0.00	0.02	0.09	0.784	
7	Developing	11,263	0.49	0.43	0.07	0.00	0.862	0.802
	On Track	8,763	0.38	0.07	0.28	0.04	0.737	
	CCR Benchmark	2,786	0.12	0.00	0.03	0.09	0.762	

Grade	Achievement Level	N	%	Expected Proportion*			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
8	Developing	11,329	0.49	0.43	0.06	0.00	0.882	0.803
	On Track	8,797	0.38	0.06	0.27	0.05	0.711	
	CCR Benchmark	2,975	0.13	0.00	0.03	0.10	0.775	
Geometry								
3	Developing	12,192	0.51	0.43	0.08	0.00	0.840	0.748
	On Track	7,519	0.32	0.08	0.21	0.03	0.666	
	CCR Benchmark	4,029	0.17	0.00	0.06	0.11	0.618	
4	Developing	12,038	0.51	0.43	0.08	0.00	0.842	0.750
	On Track	7,603	0.32	0.07	0.21	0.04	0.650	
	CCR Benchmark	4,096	0.17	0.00	0.05	0.12	0.665	
5	Developing	9,949	0.45	0.39	0.06	0.00	0.869	0.724
	On Track	9,094	0.41	0.10	0.25	0.06	0.602	
	CCR Benchmark	3,114	0.14	0.00	0.04	0.09	0.617	
6	Developing	10,137	0.44	0.38	0.06	0.00	0.858	0.754
	On Track	9,711	0.42	0.08	0.27	0.06	0.652	
	CCR Benchmark	3,343	0.14	0.00	0.04	0.11	0.736	
7	Developing	11,390	0.50	0.41	0.09	0.00	0.824	0.763
	On Track	8,865	0.39	0.07	0.27	0.05	0.689	
	CCR Benchmark	2,556	0.11	0.00	0.03	0.08	0.750	
8	Developing	11,621	0.50	0.44	0.06	0.00	0.877	0.802
	On Track	7,828	0.34	0.06	0.25	0.03	0.740	
	CCR Benchmark	3,652	0.16	0.00	0.05	0.11	0.696	
Data								
3	Developing	11,623	0.49	0.43	0.06	0.00	0.867	0.779
	On Track	8,065	0.34	0.07	0.23	0.04	0.688	
	CCR Benchmark	4,051	0.17	0.00	0.05	0.12	0.702	
4	Developing	9,325	0.39	0.35	0.04	0.00	0.885	0.698
	On Track	9,305	0.39	0.14	0.22	0.03	0.551	
	CCR Benchmark	5,108	0.22	0.00	0.07	0.13	0.623	
5	Developing	11,479	0.52	0.42	0.10	0.00	0.807	0.696
	On Track	6,873	0.31	0.08	0.17	0.06	0.555	
	CCR Benchmark	3,805	0.17	0.00	0.05	0.11	0.616	
6	Developing	10,118	0.44	0.38	0.06	0.00	0.872	0.737
	On Track	8,844	0.38	0.09	0.23	0.06	0.606	
	CCR Benchmark	4,227	0.18	0.00	0.05	0.13	0.692	
7	Developing	11,546	0.51	0.43	0.08	0.00	0.846	0.782
	On Track	7,419	0.33	0.07	0.23	0.04	0.692	
	CCR Benchmark	3,847	0.17	0.00	0.04	0.13	0.763	
8	Developing	7,766	0.34	0.30	0.03	0.00	0.905	0.669
	On Track	11,760	0.51	0.18	0.25	0.08	0.499	
	CCR Benchmark	3,568	0.15	0.00	0.03	0.11	0.721	

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

Table 8.15. Classification Accuracy by Achievement Level and Reporting Category—Science

Grade	Achievement Level	N	%	Expected Proportion*			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
Overall								
5	Below	6,663	0.30	0.27	0.03	0.00	0.890	0.845
	Meets	12,042	0.54	0.05	0.45	0.04	0.833	
	Exceeds	3,433	0.16	0.00	0.03	0.12	0.800	
8	Below	7,587	0.33	0.30	0.03	0.00	0.900	0.864
	Meets	10,960	0.48	0.04	0.41	0.03	0.859	
	Exceeds	4,498	0.20	0.00	0.04	0.16	0.815	
Inquiry, Nature of Science, & Tech								
5	Below	7,173	0.32	0.28	0.05	0.00	0.855	0.718
	Meets	9,829	0.44	0.10	0.28	0.06	0.626	
	Exceeds	5,136	0.23	0.00	0.06	0.16	0.703	
8	Below	7,305	0.32	0.27	0.05	0.00	0.836	0.730
	Meets	8,576	0.37	0.09	0.24	0.04	0.651	
	Exceeds	7,163	0.31	0.00	0.08	0.22	0.717	
Physical Science								
5	Below	6,244	0.28	0.24	0.04	0.00	0.848	0.732
	Meets	11,029	0.50	0.10	0.34	0.05	0.687	
	Exceeds	4,865	0.22	0.00	0.07	0.15	0.686	
8	Below	7,543	0.33	0.26	0.07	0.00	0.795	0.761
	Meets	10,967	0.48	0.06	0.35	0.06	0.737	
	Exceeds	4,534	0.20	0.00	0.05	0.15	0.761	
Life Science								
5	Below	7,946	0.36	0.29	0.07	0.00	0.811	0.728
	Meets	9,095	0.41	0.07	0.26	0.08	0.635	
	Exceeds	5,097	0.23	0.00	0.05	0.18	0.765	
8	Below	8,280	0.36	0.31	0.05	0.00	0.866	0.762
	Meets	8,646	0.38	0.07	0.26	0.04	0.693	
	Exceeds	6,118	0.27	0.00	0.07	0.19	0.721	
Earth/Space Sciences								
5	Below	5,772	0.26	0.22	0.04	0.00	0.843	0.726
	Meets	11,441	0.52	0.11	0.35	0.06	0.673	
	Exceeds	4,925	0.22	0.00	0.06	0.16	0.712	
8	Below	7,303	0.32	0.27	0.04	0.00	0.861	0.748
	Meets	10,529	0.46	0.09	0.31	0.05	0.687	
	Exceeds	5,212	0.23	0.00	0.06	0.16	0.712	

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

8.5. 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 plus error variance). Clearly, 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 to 1, where 0.80 is typically considered the minimally acceptable level of reliability for assessments like the NSCAS. While sensitive to random error associated with content sampling variability, the index is not sensitive to other types of errors, such as temporal stability or variability in performance that might occur across different testing occasions. Cronbach’s alpha is computed as follows (Crocker & Algina, 1986):

$$\hat{\alpha} = \frac{k}{k-1}\left(1-\frac{\sum\sigma_j^2}{\sigma_X^2}\right) \tag{8.1}$$

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:

$$\text{SEM} = SD\sqrt{1-\hat{\alpha}} \tag{8.2}$$

where SD represents the standard deviation of the raw score distribution and $\hat{\alpha}$ represents Cronbach’s alpha, as expressed in Equation 8.1. The overall SEM is expressed in raw score units and is a test-level statistic. Table 8.16 presents Cronbach’s alpha reliability coefficients by demographics for the Science fixed forms, along with the SEMs. The alpha reliability coefficients are similar to marginal reliability (reported in Table 8.9) and to the 2017 results.

Table 8.16. Cronbach’s Alpha (Internal Consistency) by Demographics for Science Fixed Forms

Grade	Demographic Group*		#Items	Reliability	SEM
5		**Grade 5 Overall**	**50**	**0.89**	**10.93**
	Gender	Female	50	0.88	11.17
		Male	50	0.89	11.11
	Ethnicity	AI/AN	50	0.88	10.60
		Asian	50	0.92	10.76
		Black or African American	50	0.87	10.56
		Hispanic	50	0.87	10.49
		NH/PI	50	0.92	10.96
		White	50	0.87	11.25
		Two or More Races	50	0.88	10.88
	FRL	Yes	50	0.88	10.57
		No	50	0.86	11.42
	LEP	Yes	50	0.87	10.60
		No	50	0.88	11.14
	SPED	Yes	50	0.88	10.70
		No	50	0.87	11.25

Grade	Demographic Group*		#Items	Reliability	SEM
8		**Grade 8 Overall**	**60**	**0.92**	**10.18**
	Gender	Female	60	0.91	10.19
		Male	60	0.92	10.68
	Ethnicity	AI/AN	60	0.90	10.37
		Asian	60	0.93	10.73
		Black or African American	60	0.89	9.73
		Hispanic	60	0.89	10.19
		NH/PI	60	0.91	10.45
		White	60	0.91	10.37
		Two or More Races	60	0.91	10.06
	FRL	Yes	60	0.90	10.31
		No	60	0.90	10.69
	LEP	Yes	60	0.89	10.29
		No	60	0.91	10.64
	SPED	Yes	60	0.89	10.39
		No	60	0.90	10.82

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

Section 9: Validity

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

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

9.1. Intended Purposes and Uses of Test Scores

The purposes of the NSCAS Summative assessment are as follows:

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

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

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

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

The unintended uses of the NSCAS are as follows:

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

9.2. Sources of Validity Evidence

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

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

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

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

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

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

9.3. Evidentiary Validity Framework

Table 9.1 presents an overview of the validity components covered in this technical report. Table 9.2 – Table 9.5 then examine the types of evidence available for each intended purpose of the NSCAS Summative assessments.

Table 9.1. Sources of Validity Evidence for Each NSCAS Test Purpose

	Sources of Validity Evidence			
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 9.2. Sources of Validity Evidence based on Test Content

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, 8
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 for ELA and Mathematics 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
3. Measure students' annual progress toward college and career readiness in ELA and Mathematics.	• Nebraska's College and Career Ready Standards for ELA and Mathematics 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

Purpose	Summary of Evidence	Tech Report Sections
4. Inform teachers how student thinking differs along different areas of the scale as represented by the ALDs as information to support instructional planning.	• 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.	2,4,6
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 sub-groups. • Assessments are administered with appropriate accommodations.	2,3,5,8

Table 9.3. Sources of Validity Evidence based on Response Process

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 sub-groups. • Assessments are administered with appropriate accommodations.	2,3,5,8

Table 9.4. Sources of Validity Evidence based on Internal Structure

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.	5, 7, 8
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.	5, 6
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.	5, 6, 8
4. Inform teachers how student thinking differs along different areas of the scale as represented by the ALDs as information to support instructional planning.	• 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.	2,6
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.	5,8

Table 9.5. Sources of Validity Evidence based on Other Variables

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.	7
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.	This will be addressed in future studies of annual observed growth.	
4. Inform teachers how student thinking differs along different areas of the scale as represented by the ALDs as information to support instructional planning.		

Purpose	Summary of Evidence	Tech Report Sections
5. Assess students’ construct relevant achievement in ELA, Mathematics, and Science for all students and subgroups of students.		

9.4. Interpretive Argument Claims

The test scores for the 2018 NSCAS support their intended purpose, and the interpretation of the test scores after the careful development of the Reporting ALDs support that the test scores describe where the students were in their learning at the end of the year based on the Nebraska College and Career Ready standards. The claims to support this documented in the technical report are shown in Table 9.6.

Table 9.6. Interpretive Argument Claims, Evidence to Support the Essential Validity Elements

Arguments	Tech Report Sections	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.	8. Reliability 5. Psychometric Analyses	Simulations, analysis of test information, conditional standard errors of measurement, classification accuracy, and reliability estimates; blueprint comparability across students; item analysis, IRT model, vertical scaling and equating procedures.
Test administrations were secure and standardized.	3. Test Administration and Security	Test administration procedures, including administration training, test accommodations, test security, and availability of help desk during testing window.
Scoring was standardized and accurate.	4. Scoring and Reporting	Scoring rules and procedures; quality control of operational scoring,
Achievement standards were rigorous and technically sound.	6. 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 5. Psychometric Analyses	Accommodation policy and implementation, sensitivity review, availability of translations, and DIF analyses.

9.5. The NSCAS Validity Argument

The test development and technical quality of the NSCAS Summative assessments supports the intended test score interpretations that are provided through the Reporting ALDs and scale scores. The TOS, passage specifications, item specifications, and ALD development process show that the NSCAS Summative assessments are aligned to grade-level content. For ELA and Mathematics there is evidence that the student response processes associated with cognitive complexity specified in the standards and TOS is behaving as intended. As an added dimension for adaptive testing, the NSCAS Summative ELA and Mathematics assessments demonstrated that the tests administered to students conform to the TOS during the constraint 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 NSCAS Summative test's internal structure, this report includes principal component analysis (PCA) that shows one dominant dimension, as well as indices of measurement precision such as test reliability, classification accuracy, CSEMs, test information, and DIF. The high correlations between NSCAS and MAP Growth show a strong relationship between the two test scores and provide concurrent evidence based on other variables. Future studies may include a predictive validity study using ACT or SAT, as well as a concurrent validity study using NAEP.

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

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

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.

CRESST. (2015, June). *Simulation-based evaluation of the Smarter Balanced summative assessments*. National Center for Research on Evaluation, Standards, & Student Testing. Retrieved from https://portal.smarterbalanced.org/library/en/simulation-based-evaluation-of-the-smarter-balanced-summative-assessments.pdf.

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

Data Recognition Corporation (DRC). (2017). *Spring 2018 Nebraska State Accountability (NeSA) ELA, mathematics, and science technical report*. Retrieved from https://cdn.education.ne.gov/wp-content/uploads/2017/11/Final-NeSA-2017-Technical-Report.pdf.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (ETS Research Report No. ETS-RR-91-47). Princeton, NJ: Educational Testing Service.

EdMetric. (2018a). *Nebraska Student-Centered Assessment System – mathematics standard setting technical report*.

EdMetric. (2018b). *Nebraska Student-Centered Assessment System – English language arts cut score review technical report*.

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice and a proposed framework. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). New York: Routledge.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Hingham, MA: Kluwer-Nijhoff Pub.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, and H. I. Brown (Eds.), *Test Validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Huynh, H. (2000). Guidelines for Rasch linking for PACT. Memorandum to Paul Sandifer on June 18, 2000. Columbia, SC: Available from Author.

Huynh, H., & Rawls, A. (2009). A comparison between Robust *z* and 0.3-logit difference procedures in assessing stability of linking items for the Rasch model. In E. V. Smith, Jr., & G. E. Stone (Eds.) *Applications of Rasch measurement in criterion-referenced testing*. (pp. 429–442). Maple Grove, MN: JAM Press.

Huynh, H., & Meyer, P. (2010). Use of Robust Z in detecting unstable items in item response theory models. *Practical Assessment, Research & Evaluation*, *15*(2), 1–8.

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

Kolen, M. J. (2011). *Issues associated with vertical scales for PARCC assessments* [White paper]. Retrieved from https://parcc-assessment.org/content/uploads/2017/11/Vertical-Scales-Kolen.pdf.

Linacre, J. M. (2015). Winsteps® Rasch measurement computer program (V3.91.0.0). Beaverton, OR: winsteps.com.

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

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

McDonald, R. P. (1979). The structural analysis of multivariate data: A sketch of general theory. *Multivariate Behavioral Research*, *14*, 21–38.

Mulkey, J., Maynes, D., & Scott, M. (2018, Dec. 12). *Data forensics report: Nebraska Student-Centered Assessment System, grades 3–8 English language arts and math, and grades 5 & 8 science, spring 2018 administration*. Midvale, UT: Caveon.

NWEA. (2018a, January). *2017–2018 CAT engine simulation report for the Nebraska grades 3–8 ELA and mathematics assessments*. Report provided to the NDE. Portland, OR: NWEA.

NWEA (2018b, June). *2018 operational CAT engine evaluation report for the NSCAS-general summative ELA and mathematics assessments*. Report provided to the NDE. Portland, OR: NWEA.

NWEA (2018c, November). *2018 linking study: Predicting performance on the NSCAS summative ELA and mathematics assessments based on MAP Growth scores*. Portland, OR: NWEA. Retrieved from https://www.nwea.org/resources/nebraska-linking-study/.

Patz, R.J. (2007). *Vertical scaling in standards-based educational assessment and accountability systems.* Paper prepared for the Technical Issues in Large Scale Assessment State Collaborative on Assessment and Student Standards of the Council of Chief State School Officers.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments, *Educational Psychologist* *51*(1), p. 59–81.

Perie, M., & Huff., K. (2016). Determining the content and cognitive demand for achievement tests. In S. Lane, M. Raymond, & T. Haladyna. *Handbook of Test Development (2nd Ed).* pp. 119–143. New York, NY: Routledge.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 221–261). New York: American Council on Educational and Macmillan.

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

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

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

Smarter Balanced Assessment Consortium (SBAC). (2016). *Smarter Balanced Assessment Consortium: 2014-15 technical report*. Retrieved from https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf.

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2*(1), 66–78.

U.S. Department of Education (2015). *U.S. Department of Education peer review of state assessment systems: Non-regulatory guidance for states for meeting requirements of the Elementary and Secondary Education Act of 1965, as amended*. Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education.

Webb, N. (1997). *Research monograph number 6: Criteria for alignment of expectations and assessments on mathematics and science education*. Washington, D.C.: CCSSO.

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

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

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

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