NSCAS NEBRASKA STUDENT-CENTERED ASSESSMENT SYSTEM

Spring 2020 NSCAS General Summative ELA, Mathematics, and Science Technical Report

nwea

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

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

ALD	achievement level descriptor
CAP	Comprehensive Assessment Platform
CCC	Crosscutting Concept
CCR	College and Career Readiness
DCI	Disciplinary Core Idea
DIF	differential item functioning
DOK	Depth of Knowledge
DRC	Data Recognition Corporation

EDS	Educational Data Systems
ELA	English Language Arts
ELL	English language learner
ESEA	Elementary and Secondary Education Act
ESC	Education Strategy Consulting
ESU	educational service unit
ETS	Educational Testing Service
FT	field test
HL	horizontal linking
ID	Item-Descriptor
ISR	Individual Student Report
IEP	Individualized Education Plan
IRT	item response theory
IWW	item writer workshop
LOSS	lowest obtainable scale score
MC	multiple-choice
MLE	maximum likelihood estimation
NCCRS-S	Nebraska College and Career Ready Standards for Science
NCLB	No Child Left Behind
NDE	Nebraska Department of Education
NeSA	Nebraska State Accountability
NSCAS	Nebraska Student-Centered Assessment System
OIB	ordered item book
OP	operational
PP	paper-pencil
RAEL	Recently Arrived Limited English Proficient
SD	standard deviation
SEM	standard error of measurement
SEP	Science and Engineering Practice
SFTP	Secure File Transfer Protocol
STARS	School-based Teacher-led Assessment and Reporting System
TAC	Technical Advisory Committee
TAM	Test Administration Manual
TCC	test characteristic curve
TEI	technology-enhanced item
TOS	Table of Specifications
TTS	text-to-speech
UAT	user acceptance testing
UDL	Universal Design for Learning
VL	vertical linking
VOIP	Voice Over Internet Protocol

Executive Summary

The Spring 2020 Nebraska Student-Centered Assessment System (NSCAS) General Summative testing was cancelled due to COVID-19. This technical report documents the processes and procedures that had been implemented to support the Spring 2020 assessments prior to the cancellation. Below is a high-level summary of each section in the technical report.

Section 1: Introduction

The NSCAS General Summative assessments are administered in English language arts (ELA) and mathematics in Grades 3–8 and in science in Grades 5 and 8. The science assessment is being transitioned to the Nebraska College and Career Ready Standards for Science (NCCRS-S). A full-scale field test was planned for Spring 2020 but will now take place in Spring 2021. The purposes of the NSCAS assessments are to measure and report Nebraska students' depth of achievement regarding the Nebraska College and Career Ready Standards; to report if student achievement is sufficient academic proficiency to be on track for achieving college readiness; to measure students' annual progress toward college and career readiness; to inform teachers how student thinking differs along different areas of the scale as represented by the achievement level descriptors (ALDs) as information to support instructional planning; and to assess students' construct-relevant achievement in ELA, mathematics, and science for all students and subgroups of students. Students taking the NSCAS tests are placed into one of the following achievement levels: Developing, On Track, or College and Career Readiness (CCR) Benchmark.

Section 2: Test Design and Development

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

Section 3: Test Administration and Security

The Spring 2020 NSCAS testing window was scheduled from March 16 to April 24, 2020. However, the 2020 administration was cancelled due to COVID-19. Prior to the cancellation, user acceptance testing (UAT) was conducted prior to the operational administration to make sure the technology and item functionality were working properly, and the appropriate test security measures were put in place.

Section 4: Scoring and Reporting

Scoring and reporting did not take place in 2020 due to the administration cancellation. As a result, student test data were not collected and there were no answer sheets to scan. Report mockups were created prior to the cancellation and are provided in Appendix C. Even though 2020 testing was cancelled, Education Strategy Consulting (ESC) maintained the Matrix with historical information for reference. Users still had access to this tool, but it was not reporting what was completed in 2020.

Section 5: Constraint-Based Engine

The NWEA constraint-based engine administers items adaptively to match the ability level of each individual student. It has two stages of consideration as it selects the next item that conforms to the blueprint while providing the maximum information about the student based on the student's momentary ability estimate: the shadow test approach followed by a variation of the weighted penalty model. Pre-administration simulations were conducted prior to the Spring 2020 testing window to evaluate the constraint-based engine's item selection algorithm and estimation of student ability based on the blueprint. Because summative testing was cancelled, a post-administration evaluation study was not conducted.

Section 6: Psychometric Analyses

Psychometric analyses were not conducted for Spring 2020 due to the administration cancellation.

Section 7: Standard Setting

No standard setting was held in 2019–2020. If testing and scoring had occurred in 2020, the cut scores would have been the same as in 2018 and 2019. Nebraska's statewide assessment system for ELA and mathematics underwent significant changes between 2016 and 2017, so cut scores for ELA and mathematics were set following the Spring 2018 administration at standard setting and cut score review meetings from July 26–28, 2018, using the Item-Descriptor (ID) Matching method. The purpose of the standard setting was to set new cut scores for mathematics, whereas the purpose of the cut score review was to validate the existing cut scores for ELA. Standard setting will take place for the new NSCAS Science assessment following the first operational administration.

Section 8: Test Results

Test results are not provided for Spring 2020 due to the administration cancellation.

Section 9: Reliability

The reliability/precision of the Spring 2020 NSCAS assessments is not able to be properly evaluated due to the administration cancellation.

Section 10: Validity

Validating a test score interpretation is not a quantifiable property but an ongoing process, beginning at initial conceptualization of the construct and continuing throughout the entire assessment process. As the technical report progresses, it covers the different phases of the testing cycle and the procedures and processes applied in the NSCAS. The section revisits phases and summarizes relevant evidence and a rationale in support of any test score interpretations and intended uses based on the *Standards for Educational and Psychological Testing* (AERA et al., 2014). The validity argument begins with a statement of the assessment's intended purposes, followed by the evidentiary framework where available validity evidence is provided to support the argument that the test actually measures what it purports to measure (SBAC, 2016).

Section 1: Introduction

The Spring 2020 administration of the Nebraska Student-Centered Assessment System (NSCAS) General Summative assessments was cancelled due to COVID-19. The purpose of this technical report is to summarize the test development work that had occurred in support of the 2020 administration up until the cancellation. It does not include any test or psychometric analysis results based on empirical student data.

1.1. NSCAS Overview

NSCAS is a statewide assessment system that embodies Nebraska's holistic view of students and helps them prepare for success in postsecondary education, career, and civic life. It uses multiple measures throughout the year to provide educators and decision makers at all levels with the insights they need to support student learning. The NSCAS General 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 in English language arts (ELA) and mathematics and in Grades 5 and 8 in science.

Due to the suspension of the 2020 NSCAS General Summative test, no data were collected and no student scores were produced. The NSCAS assessments have typically been administered online with paper-pencil versions available as an accommodation. They include a variety of item types, including multiple-choice and technology-enhanced items. Student scores are reported as composite scale scores, reporting category scale scores, and achievement levels. The ELA and mathematics assessments are administered using a multi-stage adaptive design, whereas science is currently under development with the next step being a full-scale field test. Students taking the ELA and mathematics tests are placed into one of the following achievement levels based on their final test scores: Developing, On Track, and College and Career Readiness (CCR) Benchmark. The new science assessment will use these achievement levels as well.

Items for ELA and mathematics are aligned to the 2014 and 2015 College and Career Ready Standards, respectively, and come from the item bank that the Nebraska Department of Education (NDE) and Nebraska educators have built over the years. The tests also include newly developed field test items that are added to the operational pool depending on the field test data and data review. Tasks for the new NSCAS Science test were developed in Summer 2019 and are aligned to the Nebraska College and Career Ready Standards for Science (NCCRS-S; NDE, 2017). A full-scale field test was planned for Spring 2020 but will now take place in Spring 2021 due to the administration cancellation.

1.2. Background

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

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

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

NSCAS replaced the NeSA assessments beginning in 2017–2018. Spring 2019 was the second administration of the NSCAS ELA and Mathematics assessments that were administered adaptively, whereas science continued to be administered as a fixed-form assessment. The new NSCAS Science assessment aligned to the NCCRS-S was piloted in March 2019, with a full-scale field test scheduled for Spring 2020 and an operational launch in Spring 2021. However, due to COVID-19, the Spring 2020 NSCAS administration was cancelled. No testing occurred, which resulted in no field test items or science tasks being administered in any content area. As a result, reporting did not occur and no psychometric analyses using empirical student data were conducted in 2020.

1.3. Schedule of Major Events

Table 1.1 presents the major events that occurred for the 2020 NSCAS assessments, including the new science assessment. NDE involves educators throughout the development process to produce customized items and provide an invaluable professional development opportunity, including item/task writing and review meetings and achievement level descriptor (ALD) reviews.

Table 1.1. Schedule of Major Events for the Spring 2020 Administration

Event	Date(s)
ELA passage review	March 12, 2019
Science ALD workshop	May 1–2, 2019
ELA and mathematics item writing workshop	June 11–13, 2019
Science phenomena writing workshop	June 17–21, 2019
Science task writing workshop	July 8–12, 2019
ELA and mathematics content and bias review committee	July 23–25, 2019
Science content and bias review committee	September 9–12, 2019
Fall 2019 regional workshop	October 9–16, 2019
Summative test administration training	February 14–20, 2020
Technical Advisory Committee (TAC) meeting	March 13, 2020
Operational testing window (cancelled due to COVID-19)	March 16 – April 24, 2020
Make-up testing window (cancelled due to COVID-19)	April 27 – May 1, 2020

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

1.4. Principled Assessment Design

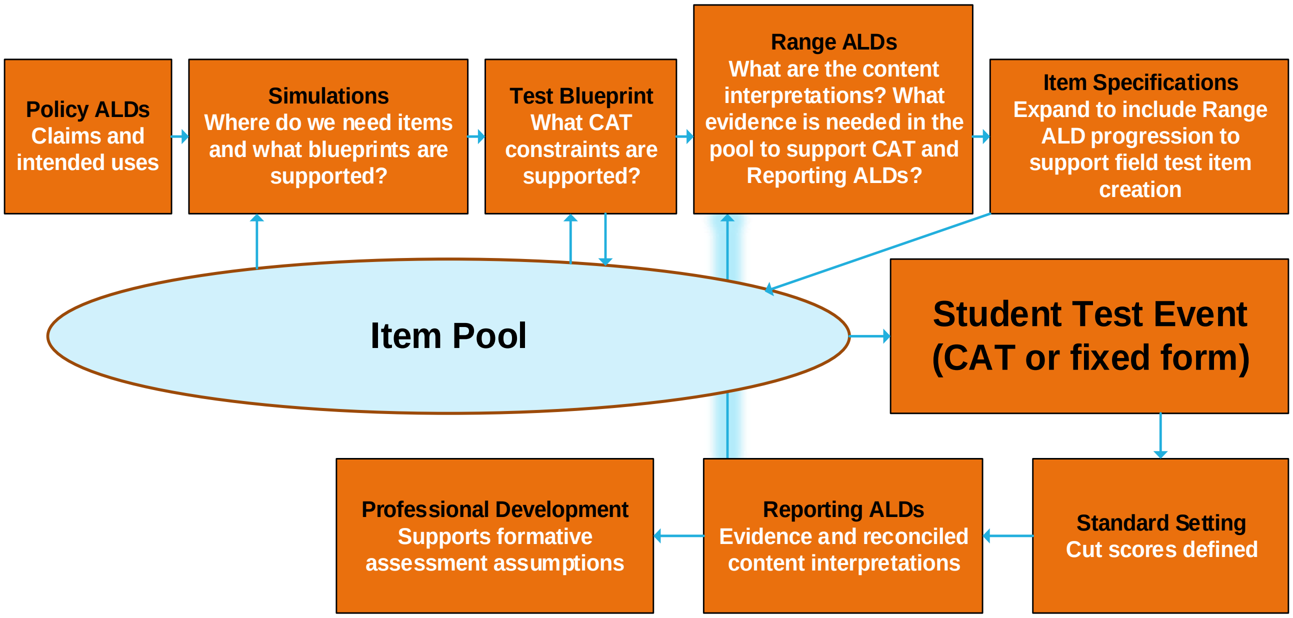
The NSCAS General Summative assessments have been developed based on a principled approach to test design that centers around ALDs and conceptualizing test score use as part of a broader solution to achieve important outcomes for test users. The evidence needed to draw a conclusion about where a student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

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

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

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

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


Policy ALDs
Claims and intended uses
Simulations
Where do we need items and what blueprints are supported?
Test Blueprint
What CAT constraints are supported?
Range ALDs
What are the content interpretations? What evidence is needed in the pool to support CAT and Reporting ALDs?
Item Specifications
Expand to include Range ALD progression to support field test item creation
Item Pool
Student Test Event (CAT or fixed form)
Professional Development
Supports formative assessment assumptions
Reporting ALDs
Evidence and reconciled content interpretations
Standard Setting
Cut scores defined

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

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

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

1.5. Intended Purposes and Uses of Test Results

The following are the intended purposes of the NSCAS assessments:

1. To measure and report Nebraska students' depth of achievement regarding the Nebraska College and Career Ready Standards
2. To report if student achievement is sufficient academic proficiency to be on track for achieving college readiness
3. To measure students' annual progress toward college and career readiness

4. To inform teachers how student thinking differs along different areas of the scale as represented by the ALDs as information to support instructional planning
5. To assess students' construct-relevant achievement in ELA, mathematics, and science for all students and subgroups of students

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

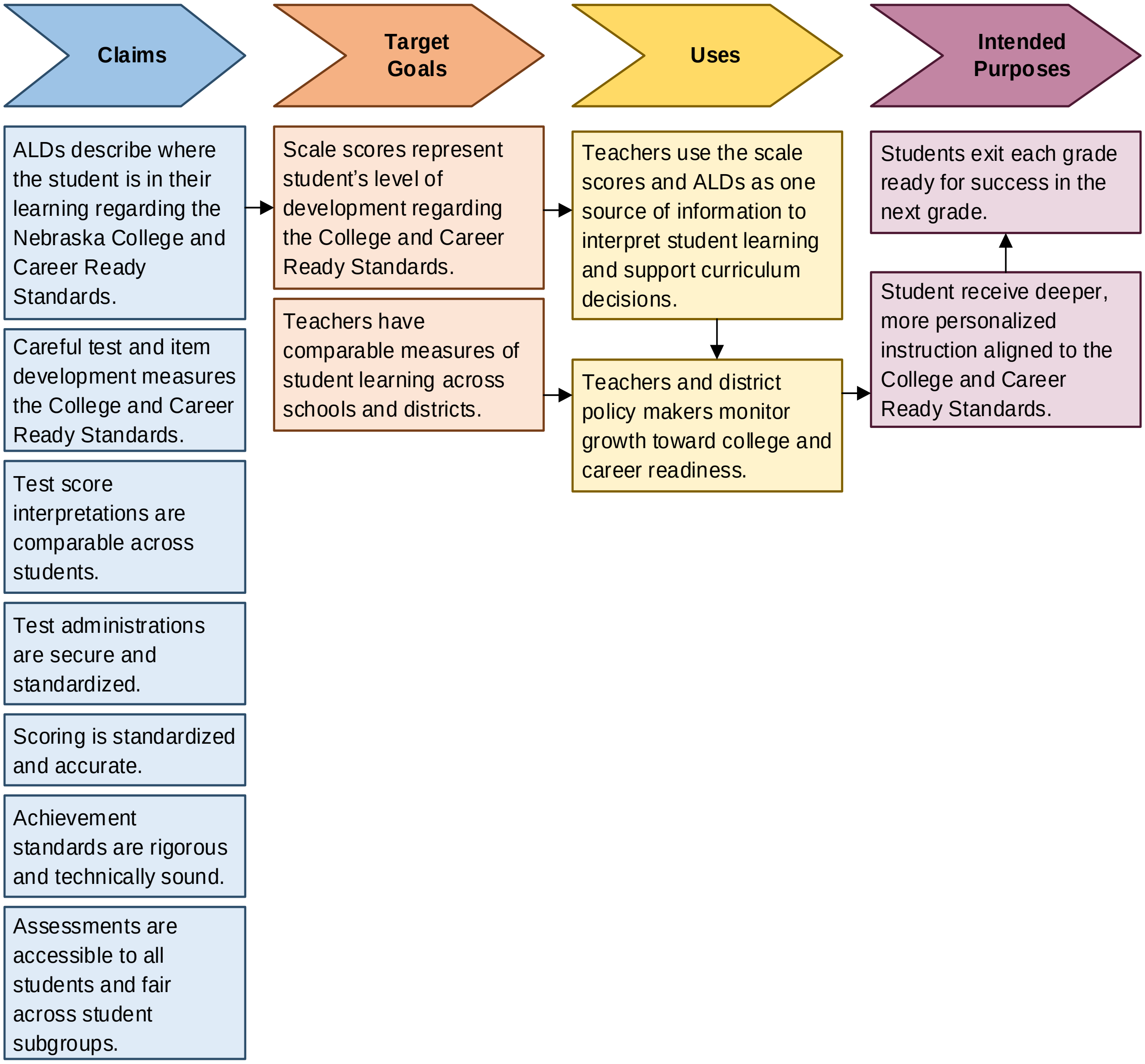
- To supplement teachers' observations and classroom assessment data
- To improve the decisions teachers make about sequencing instructional goals, designing instructional materials, and selecting instructional approaches for groups and individuals
- To identify individuals for summer school and other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

1.6. Theory of Action

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

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

Figure 1.2. NSCAS Theory of Action



Claims

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

Careful test and item development measures 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 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

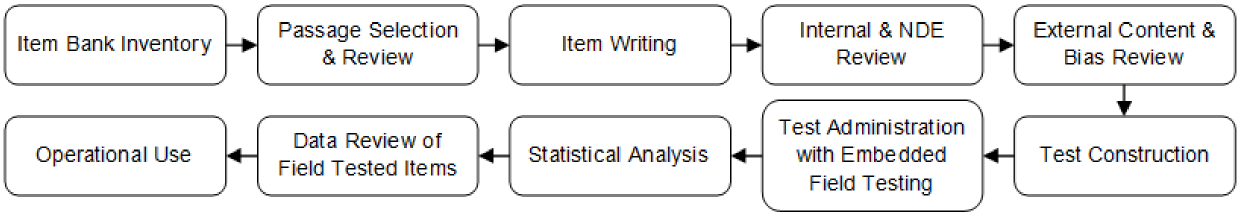
Students exit each grade ready for success in the next grade.

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

Section 2: Test Design and Development

This section describes the test design and development processes for the 2020 NSCAS General Summative assessments. As Nebraska transitioned to an adaptive administration for ELA and mathematics in 2017–2018, the need to build a large, robust item bank was a key requirement, and the development of new scales had to be accomplished concurrently with thinking about the development of ALDs. Development to support building of a bank to sufficiently support adaptive testing continued for 2019–2020 to have enough content available to populate field test slots in the Spring 2020 assessments. Items were written by educators in an item writing workshop (IWW) and by independent contractors. Passages were also developed by contractors and reviewed by Nebraska educators. Once initial item development was completed, all items were taken to content and bias review meetings with Nebraska educators. Items that survived these meetings were considered for the field test pool. Figure 2.1 outlines the general steps taken to develop the passages and items, although the test administration, statistical analysis, and data review will now occur in Spring 2021 for the items developed for the Spring 2020 administration.

Figure 2.1. Test Development Process


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

Content development for the new three-dimensional science assessment began in Summer 2018 with the pilot occurring in March 2019. The Spring 2020 full-scale field test was intended to be a next step from the pilot test from March 2019. However, due to the cancellation of the 2020 administration, the science field test will now occur in Spring 2021.

2.1. Test Designs

Table 2.1 summarizes the different versions of the NSCAS General Summative assessments available for 2020 (had the assessments been administered). Table 2.2 presents the number of items and points possible on each online and paper-pencil test form. The paper-pencil forms served as accommodated versions that contained only operational items and were slightly longer than the adaptive assessments to support comparable levels of test score precision. Science was to be administered as a full-scale field test in Spring 2020.

Table 2.1. Available NSCAS General Summative Assessments in 2020

	Available Assessments*				
Grade(s)	Online	PP	Spanish Online	Spanish PP	Breach
ELA					
3–8	Adaptive (53 total per grade, 41 OP + 7 FT/VL + 5 MAP Growth)	2018 PP form (with minimal updates); 1 form per grade (48 OP)	Fixed (translation of PP form)	Same form as Spanish online	2019 PP form

Grade(s)	Available Assessments* Online	PP	Spanish Online	Spanish PP	Breach
Mathematics					
3–8	Adaptive (53 total per grade, 41 OP + 7 FT/VL + 5 MAP Growth)	2018 PP form (with minimal updates); 1 form per grade (48 OP)	Fixed (translation of PP form)	Same form as Spanish online	2019 PP form
Science					
5	FT only (42 prompts per form)	–	–	–	–
8	FT only (41 prompts per form)	–	–	–	–

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

Table 2.2. Number of Items and Points Per Test

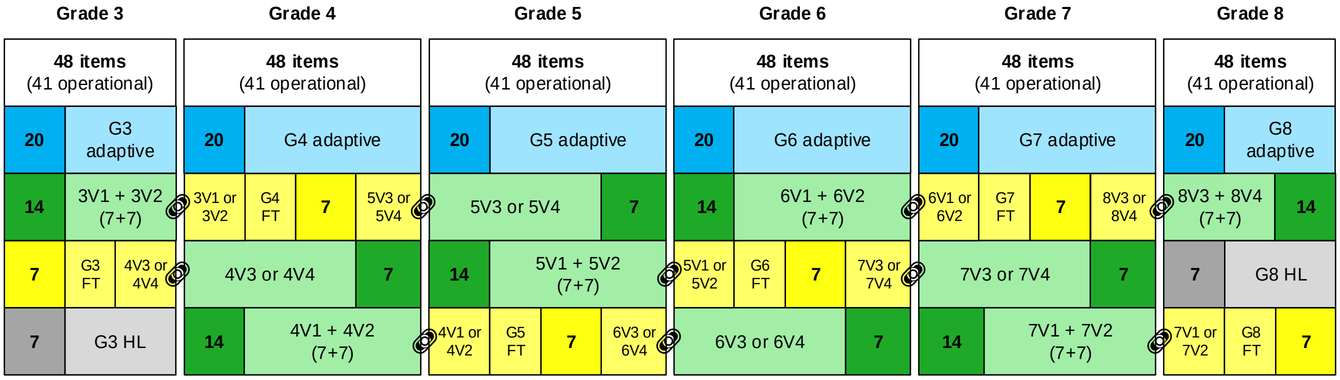
Grade	Online*									
	Operational		FT/VL		MAP Growth		Total		Paper-Pencil	
	#Items	#Points	#Items	#Points	#Items	#Points	#Items	#Points	#Items	#Points
ELA										
3	41	47–51	7	7–10	5	5	53	59–66	48	51
4	41	48–50	7	7–10	5	5	53	60–65	48	51
5	41	51–54	7	7–10	5	5	53	63–69	48	52
6	41	49–54	7	7–10	5	5	53	61–69	48	53
7	41	50	7	7–10	5	5	53	62–65	48	52
8	41	52–57	7	7–10	5	5	53	64–72	48	53
Mathematics										
3	41	45	7	7–9	5	5	53	57-59	48	49
4	41	45	7	7–9	5	5	53	57-59	48	48
5	41	45	7	7–9	5	5	53	57-59	48	48
6	41	45	7	7–9	5	5	53	57-59	48	52
7	41	45	7	7–9	5	5	53	57-59	48	51
8	41	45	7	7–9	5	5	53	57-59	48	50
Science										
5	–	–	42	42	–	–	–	–	–	–
8	–	–	41	41	–	–	–	–	–	–

*FT/VL = field test/vertical linking. Items in this slot are either FT or VT items for ELA and mathematics. The science test is a full-scale field test that will now occur in Spring 2021 and will be operational in Spring 2022. MAP Growth items are added at the end of the ELA and mathematics tests as non-operational items to build the through-year item bank.

2.1.1. ELA and Mathematics

Figure 2.2 illustrates the online adaptive test design for the NSCAS ELA and Mathematics assessments using both horizontal linking (HL) and vertical linking (VL) anchor items (without the additional five MAP Growth items added at the end of each test). All students see a total of 48 items (41 operational + 7 non-operational). Of the 41 operational items, 21 of them are non-adaptive pre-selected HL anchors. The remaining 20 operational items are selected adaptively based on student ability level. Thus, the test design is best classified as a multi-staged adaptive assessment in which students first receive the fixed anchor sets that act as a locater with which to begin adaptive selection for the second portion of the test. Each student also sees one set of 7 non-operational items that are either on-grade field test or off-grade VL items.

Figure 2.2. Adaptive Test Design with Horizontal and Vertical Linking



🟦 = 20 on-grade adaptive operational items

🟨 = 7 non-operational items that can be either on-grade FT or off-grade VL

⬜ = 7 on-grade pre-selected, non-adaptive operational HL items that were not used as VL items (Grades 3 and 8 only)

🟩 = 14 or 7 on-grade pre-selected, non-adaptive operational HL items also used as VL items in adjacent grades

- V1 & V2 = embedded as VL items into the grade above
- V3 & V4 = embedded as VL items into the grade below

🔗 = vertical linking, either from the lower to the upper grade or from the upper to the lower grade

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

Table 2.3. Horizontal Linking Configuration

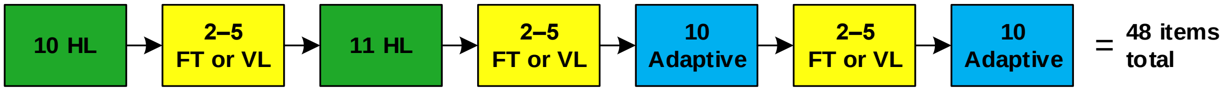
	Horizontal Form 1 (core)			Horizontal Form 2a (Set A)			Horizontal Form 2b (Set B)			Total #HL Items Per Student
Grade	Item Set(s)	#Items	%N	Item Set	#Items	%N	Item Set	#Items	%N	
3	HL	7	100%	V1	7	100%	V2	7	100%	21
4	V1+V2	14	100%	V3	7	50%	V4	7	50%	21
5	V1+V2	14	100%	V3	7	50%	V4	7	50%	21
6	V1+V2	14	100%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

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

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

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

Figure 2.3. General Item Sequence for ELA and Mathematics



2.1.2. Science Field Test

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

Each grade has three test forms, each with seven tasks and 4–8 associated prompts. One or two survey questions are also included at the end of each test to make the test lengths equal across forms at each grade level, allowing the constraint-based engine to properly administer the forms. The survey questions will also garner feedback from students in terms of their test-taking engagement. Each task is included on at least two test forms per grade, as shown in Table 2.4, to ensure a sufficient number of responses per task for item calibration and to allow an evaluation of how the prompts of the task are likely to function operationally. These common tasks across forms also serve as anchor sets to equate prompts across forms. For example, Task 2147 in Grade 5 appears on all three forms, and Task 2136 is common on Forms A and B.

Table 2.4. Science Field Test Form Design

Task Code	#Prompts	Form A	Form B	Form C
Grade 5				
2135	7	X		X
2136	6	X	X	
2139	4		X	X
2142	4	X		X
2143	8		X	X
2144	4	X	X	
2145	5		X	X
2146	6	X		X
2147	6	X	X	X
2149	8	X	X	
Survey Q1 (41176550)	1	X	X	X
Survey Q2 (41176560)	1			X
Total #Prompts		42	42	42
Total #Tasks		7	7	7
Grade 8				
2133	5	X	X	X
2150	6		X	X
2151	5	X		X
2154	6	X	X	
2155	5		X	X
2156	6	X	X	X
2158	6	X	X	
2160	7	X	X	X
2161	5	X		X
Survey Q1 (41176530)	1	X		X
Survey Q2 (41176540)	1			X
Total #Prompts		41	41	41
Total #Tasks		7	7	7

The order of prompts within a task is fixed, but the order of tasks on a form varies across students to reduce task position effect that can alter the quality of the data due to factors such as fatigue. For example, students might be tired at the end of a test and will not do as well as the beginning, so task positions vary across forms (e.g., a task can appear early on a form for some students but in a late position for others) to ensure an even opportunity for full student engagement. In addition, two tasks per grade with high content similarities (i.e., Tasks 2135 and 2136 for Grade 5 and Tasks 2156 and 2161 for Grade 8) were set to be non-adjacent on a test form (i.e., not situated next to each other) to avoid situations in which students may not realize the differences between the two tasks and use incorrect information to respond to the prompts.

2.2. Academic Content Standards

As stated in Nebraska Revised Statute 79-760.01[2] that was effective as of August 30, 2015[3]:

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

On September 5, 2014, the Nebraska State Board of Education adopted Nebraska's College and Career Ready Standards for ELA. On September 4, 2015, the Nebraska State Board of Education adopted Nebraska's College and Career Ready Standards for Mathematics. On September 8, 2017, the Nebraska State Board of Education approved the NCCRS-S that were implemented in the Spring 2019 pilot administration and will be implemented in the full-scale field test in Spring 2021.

2.3. Blueprints

The 2020 NSCAS blueprints for ELA and mathematics are embedded in the Table of Specifications (TOS) that indicate the range of test items included for each standards indicator. The adaptive test is constrained to make sure each student receives items within the identified ranges. The 2020 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. The ELA TOS for each grade is available online at https://www.education.ne.gov/assessment/nscas-general-summative-assessment/nscas-english-language-arts-ela/. The mathematics TOS for each grade is available online at https://www.education.ne.gov/assessment/nscas-general-summative-assessment/nscas-mathematics/. The blueprint for the new science assessment is currently in draft form and is available online at https://cdn.education.ne.gov/wp-content/uploads/2019/12/NE-Science-Draft-Public-Blueprint-V15.pdf. This document provides an expectation of the frequency of the DCIs, SEPs, and CCCs from the NCCRS-S. Each element from the DCIs, SEPs, and CCCs is assigned a frequency (i.e., frequent, infrequent, rare) that indicates how often the element will be assessed.

2.4. Item Types

Table 2.5 presents the item types available for the online ELA and mathematics adaptive tests. The paper-pencil tests include multiple-choice, multiselect, and composite items made up of multiple-choice and multiselect items. ELA assessments include passages that incorporate sets of items. Tasks to be field tested in science include phenomena and a set of items (i.e., prompts) using that phenomena that may include all of the available item types.

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

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

Table 2.5. Online Item Types

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

2.5. Depth of Knowledge (DOK)

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

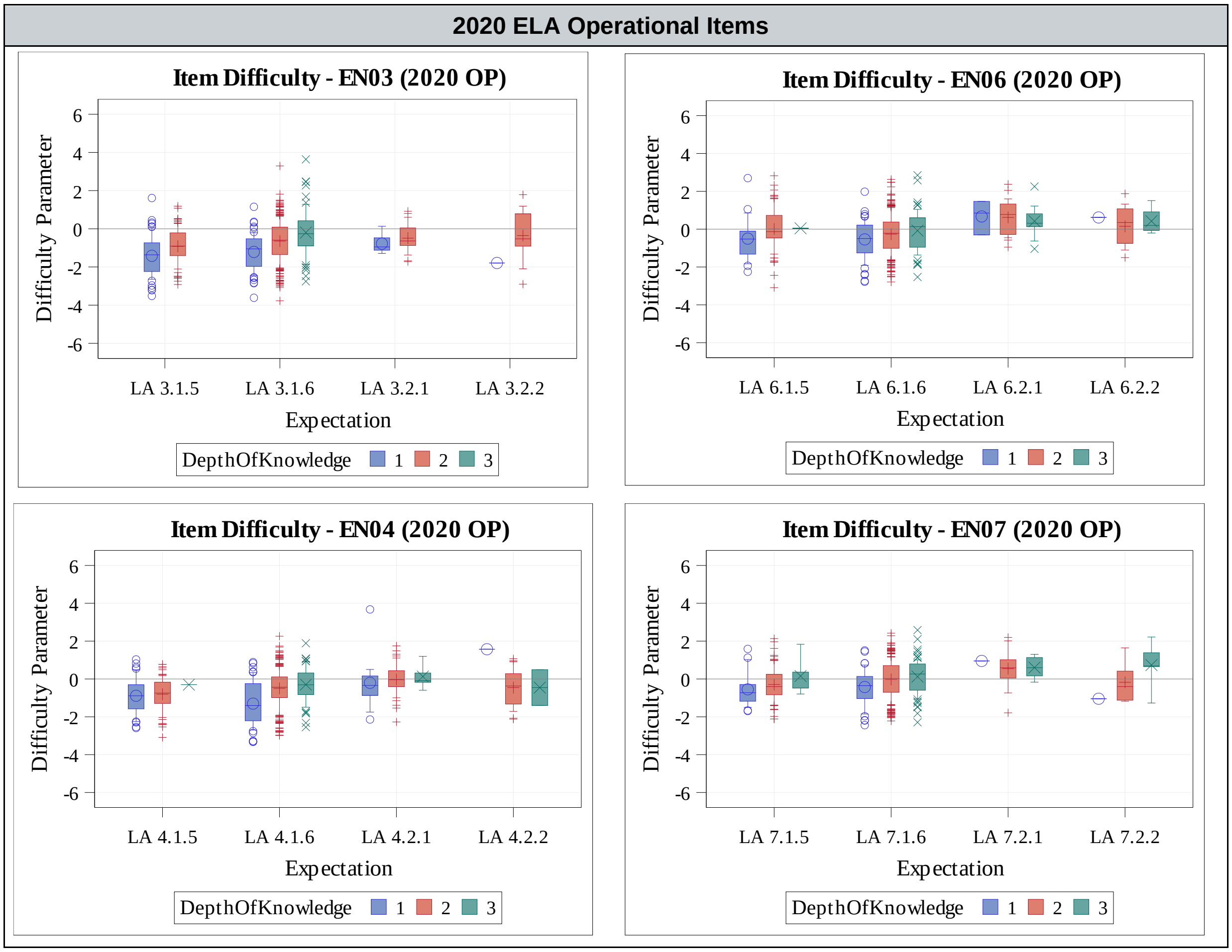
To ensure that the NSCAS assessments include a deep pool of items that span a full range of cognitive levels and skills, each item in ELA and mathematics was evaluated and tagged with one of the following DOK levels (Webb, 1997). DOK Level 4: Extended Thinking items are not included because the tests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

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

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

Figure 2.4 and Figure 2.5 present boxplots of item DOK levels based on the state's interpretation of DOK for the 2020 ELA and mathematics operational item pools, respectively. These results suggest the need to develop DOK 3 items in standards in the future based on the state policy decision in February 2018.

Figure 2.4. DOK Box Plots for 2020 Operational Items—ELA



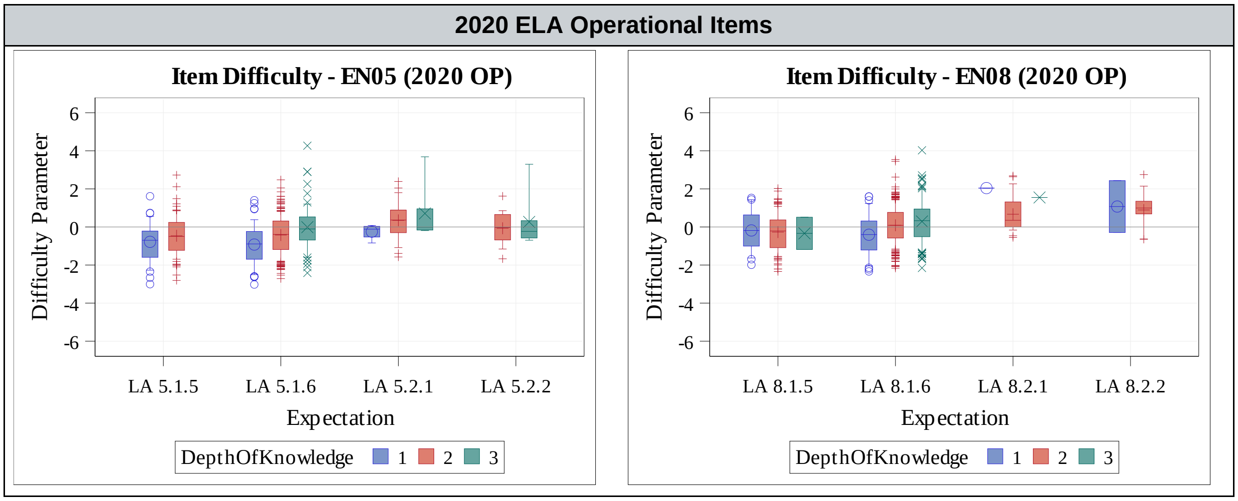
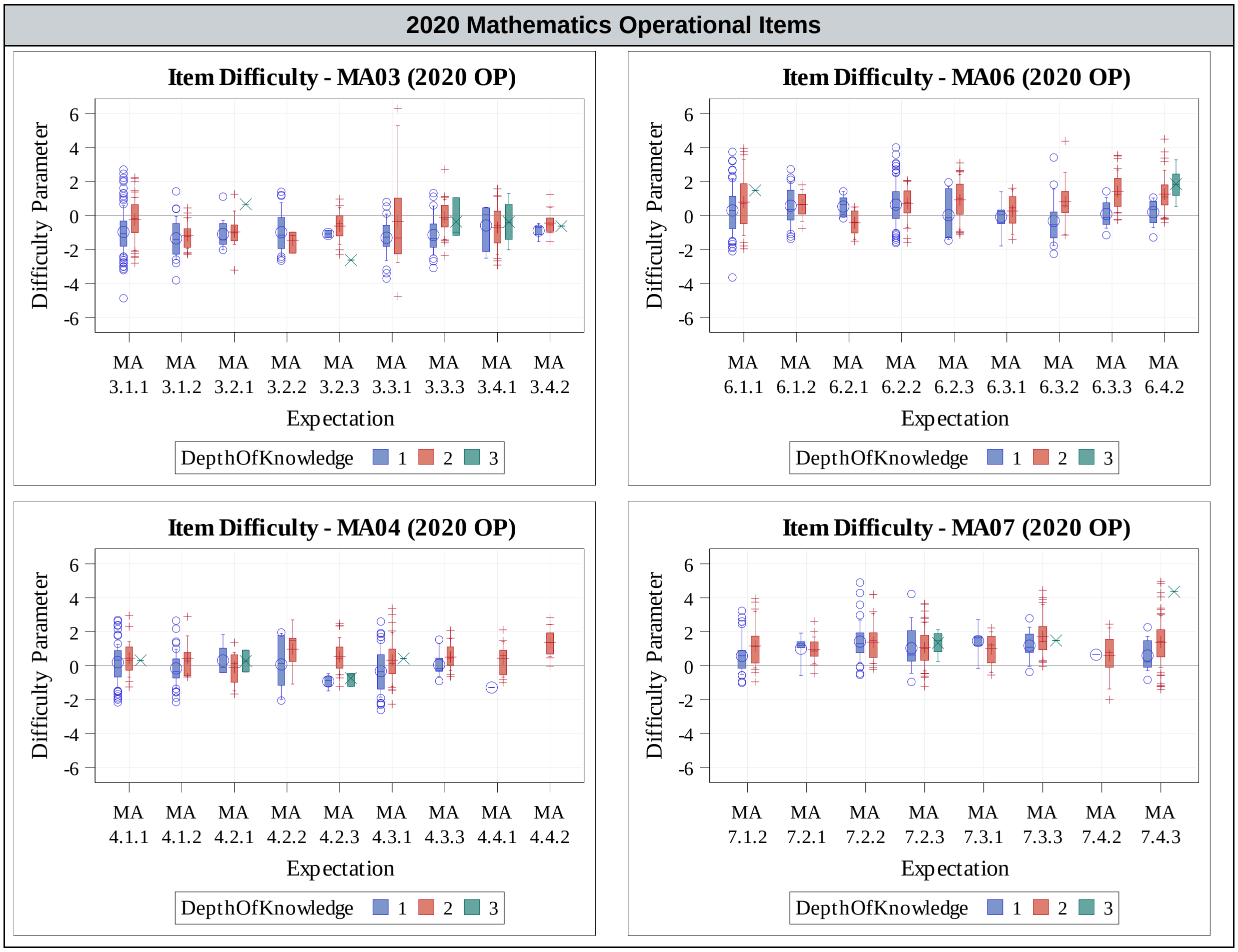
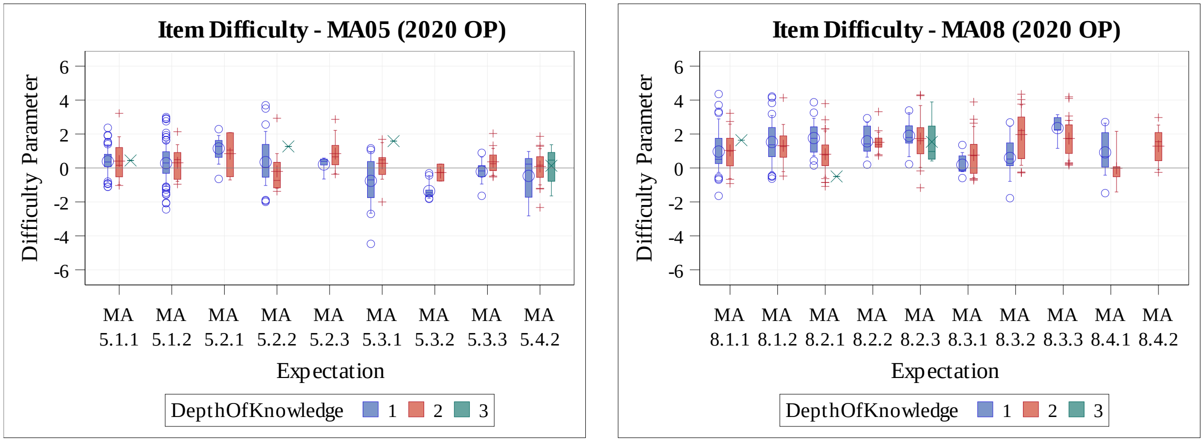


Figure 2.5. DOK Box Plots for 2020 Operational Items—Mathematics



2020 Mathematics Operational Items



2.6. ALD Development

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

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

2.6.1. Policy ALDs

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

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

2.6.2. Range ALDs

Range ALDs provide the intended content-based interpretations of what test scores within an achievement level represent and explicate observable evidence of achievement, demonstrating how the skill changes and becomes more sophisticated across achievement levels for each standard and achievement level on an assessment. Teachers can use the Range ALDs to determine how students with different scores within different achievement levels may differ in their abilities. Range ALDs for ELA were developed in 2017 and reviewed by NWEA in 2018. Range ALDs for mathematics were developed in 2018, including an educator review in Spring 2018. Both ELA and mathematics Range ALDs were refined during the July 2018 standard setting and cut score review meetings. Range ALDs have also been generated for the new science assessment aligned to the NCCRS-S, beginning with an ALD workshop in May 2019. These science ALDs are still in draft form.

2.6.2.1. ELA and Mathematics

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

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

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

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

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

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

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

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

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

2.6.2.2. Science

Before task development began in Summer 2019 for the new science assessment, it was essential to first develop the ALDs that correspond to the Developing, On Track, and CCR Benchmark achievement levels to guide development. The science Range ALDs are intended to describe students’ increasingly advanced three-dimensional reasoning on tasks that require students to apply and integrate SEPs and CCCs within and among the disciplines of science. The draft science ALDs are available online at https://cdn.education.ne.gov/wp-content/uploads/2020/02/NSCAS-Science-Summative-Achievement-Level-Descriptors-ALDs.pdf.

The NCCRS-S may be thought of as the broad content learning goals for students at each grade level that are intended to cue instruction in ways that emphasize active scientific reasoning, but there is complexity regarding how the standards are intended to be interpreted, taught, and assessed. Indicators found in the NCCRS-S are meant only to provide examples of ways the three-dimensional standards could be integrated on an assessment. Assessment tasks centered in the NCCRS-S are intended to measure a novel indicator based on the intersection of the grade-level DCI, CCC, and SEP through a task-based claim (i.e., students are applying SEPs to make sense of task phenomena using the intended DCIs and CCCs). Because a task-based claim represents a novel indicator, indicators can and likely will vary across alternate test forms of the state assessment. The ALDs must do two things:

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

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

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

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

2.6.3. Reporting ALDs

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

2.7. ELA Passage Development

Table 2.6 presents the number of passages developed for the NSCAS ELA assessments by passage type (literary vs. informational) and passage source (commissioned vs. public domain), including the development targets. As shown in the table, the targets were met with a total of 36 passages being developed, all of which were commissioned. All passages were reviewed during educator review meetings.

Table 2.6. ELA Passage Targets and Development by Passage Type and Source

		#Passages				
		Passage Type		Passage Source		
Grade	Targets	Literary	Informational	Commissioned	Public Domain	Total
3	6	6	–	6	–	6
4	6	3	3	6	–	6
5	6	3	3	6	–	6
6	6	2	4	6	–	6
7	6	2	4	6	–	6
8	6	3	3	6	–	6
Total	36	19	17	36	–	**36**

Passage specifications were updated prior to the start of passage development for ELA. Passages were not newly developed in any other content area. The document captures specifications such as what types of passages would be found or developed, as well as grade-level appropriateness, readability, word count, accuracy of facts within the passage, and bias, sensitivity, and fairness considerations.

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 text structure, levels of meaning, language features, demands on the reader, purpose, bias and sensitivity concerns, and ALD placement. The NWEA Text Complexity Qualitative Analysis Rubric was completed for each passage submitted for consideration.

The quantitative measures of a passage were also considered as a factor. 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 and 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	Lexile Range	Word Count
3	450L – 790L	200–700
4	745L – 980L	200–900
5	745L – 980L	300–1000
6	925L –1155L	400–1100
7	925L –1155L	400–1100
8	925L –1155L	400–1200

2.8. Item Development

Item/task development for 2019–2020 occurred for ELA, mathematics, and science. For ELA and mathematics, the adaptive and paper-pencil item pools are the same and therefore follow the same development processes. An in-person IWW generated 60% of the development for this cycle. Independent contractors were then used to offset gaps in the item bank to ensure that enough items were developed to fulfill the item development requirements. Development of the new three-dimensional science assessment began in July 2018 when a group of educators developed tasks and prompts for the March 2019 pilot test and continued in July 2019 with phenomena and task writing workshops. The tasks are currently awaiting field testing.

2.8.1. Item Specifications

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

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

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

2.8.2. ELA and Mathematics

2.8.2.1. <u>Development Targets</u>

Table 2.8 and Table 2.9 present the item development targets for ELA and mathematics, respectively. The item development plan included the development of 1,137 items across both content areas (777 + 360). Technology-enhanced items (TEIs) are any item type that is not a multiple-choice (MC) item and can be worth 1 or 2 points. The ELA item development focused on passage-dependent items. After the mathematics item bank realignment was complete, a review was done in 2019 prior to development. The item development plan is based on this review. Grades had different development targets across domains based on the needs of each grade.

Table 2.8. Item Development Targets—ELA

	Item Targets								
	Reading			Writing			Overall		
Grade	MC	TEI	Total	MC	TEI	Total	MC	TEI	Total
3	76	31	107	20	12	32	96	43	139
4	73	33	106	20	12	32	93	45	138
5	71	33	104	20	10	30	91	43	134
6	61	34	95	17	12	29	78	46	124
7	57	36	93	17	12	29	74	48	122
8	57	35	92	16	12	28	73	47	120
Total	395	202	662	110	70	180	505	272	777

Table 2.9. Item Development Targets—Mathematics

	Item Targets				
		TEI			
Grade	MC	1-pt.	2-pt.	Total	Overall
3	24	18	18	36	60
4	24	18	18	36	60
5	24	18	18	36	60
6	24	18	18	36	60
7	24	18	18	36	60
8	24	18	18	36	60
Total	144	108	108	216	360

2.8.2.2. Item Writer Workshop (IWW)

The IWW from June 11–13, 2019, provided a professional development opportunity to educators and allowed them to be a part of the item development process for ELA and mathematics. Table 2.10 presents the number of participants in each panel who were recruited and selected by NDE. The expertise of Nebraska teachers was critical to the item writing process. Nebraska educators wrote test items that were featured on the assessments. This ensured content that seems familiar to students as they take the tests; they will not see unfamiliar wording or approaches that might negatively impact performance.

Table 2.10. IWW Panel Composition

Panel	#Panelists
ELA 3	9
ELA 4	8
ELA 5	8
ELA 6	8
ELA 7	8
ELA 8	8
Math 3	5
Math 4	8
Math 5	8
Math 6	8
Math 7	9
Math 8	9
Total	96

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

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

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

2.8.2.3. Item Development Results

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

Table 2.11. Item Development Results—ELA

	#Items		
Grade	MC	TEI	Total
3	83	31	114
4	67	36	103
5	69	37	106
6	63	43	106
7	61	47	108
8	72	36	108
Total	415	230	645

Table 2.12. Item Development Results—Mathematics

	#Items				
		TEI			
Grade	MC	1-pt.	2-pt.	Total	Overall
3	24	18	18	36	60
4	24	18	18	36	60
5	24	18	18	36	60
6	24	18	18	36	60
7	24	18	18	36	60
8	24	18	18	36	60
Total	144	144	144	216	360

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

Grade	Target #Items	#Items Developed	Difference to be Added to the Next Development Cycle
ELA			
3	139	114	25
4	138	103	35
5	134	106	28
6	124	106	18
7	122	108	14
8	120	108	12
Mathematics			
3	60	60	–
4	60	60	–
5	60	60	–
6	60	60	–
7	60	60	–
8	60	60	–

2.8.2.4. External Content and Bias Review

Nebraska educators convened from July 23–25, 2019, 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. Educator involvement in item reviews provided another opportunity to make sure that the material was appropriate and to provide a valuable professional development opportunity. Participants received training, delivered collaboratively by NDE and NWEA, at the beginning of each review session and were provided checklists to refer to during the reviews.

Participants in item content review learned to review items for qualities such as proper alignment and cognitive complexity, clear and concise wording, and presence of a correct answer. Participants in item bias review learned to review items for qualities such as diversity of background and cultural representation, avoidance of stereotypes, avoidance of topics that may cause discomfort to students, 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: accept the item as is, accept the item with proposed modifications, and reject the item. Only items that were accepted during both reviews are eligible for field testing.

Table 2.14 presents the panel compositions for both the bias and content review meetings. Table 2.15 presents the number of items accepted, modified, or rejected results at the external content and bias review meeting. For ELA, 94.4% of items were either accepted or accepted with modifications, with the remaining 5.6% of items being rejected. For mathematics, 100% of items were either accepted or accepted with modifications.

Table 2.14. Item Review Meeting Panel Composition

Item Review Meeting	Panel	#Panelists
Bias Review	ELA 3–5	5
	ELA 6–8	2
	Math 3–5	4
	Math 6–8	5
	Total	16
Content Review	ELA 3–4	4
	ELA 5–6	3
	ELA 7–8	5
	Math 3–4	5
	Math 5–6	5
	Math 7–8	4
	Total	26
	Grand Total	**42**

Table 2.15. External Item Review Results

	#Items			
Grade	**Accepted**	**Modified**	**Rejected**	**Total**
ELA				
3	60	53	1	114
4	63	36	4	103
5	48	57	3	106
6	69	35	2	106
7	61	44	3	108
8	83	24	1	108
Total	384	249	14	645
Mathematics				
3	20	40	–	60
4	14	46	–	60
5	23	37	–	60
6	23	37	–	60
7	35	25	–	60
8	25	35	–	60
Total	140	220	–	360

2.8.2.5. Item Retirement

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

2.8.3. Science

Nebraska teachers were recruited by NDE and brought together from June 17–21, 2019, for a phenomena writing workshop and from July 8–12, 2019, for a task writing workshop. A total of 20 teachers participated, five in each grade per workshop. Table 2.16 presents the number of phenomena and tasks developed at these workshops. Each task included 4–8 prompts.

Table 2.16. Task Development Results—Science

Grade	#Phenomena Written	#Tasks Completed
5	15	10
8	17	11

The writers were guided in the vision of the new NSCAS Science assessment and began the development process by identifying a phenomenon that met NDE’s criteria (e.g., it is observable, accessible, engaging, and explainable using grade-level appropriate science core ideas). Writers then thought about the steps needed for students to make sense of the phenomenon and identified SEPs and CCCs students would use in the sense-making process. A task was built by introducing the phenomenon in a scenario that was bimodal (e.g., it had text and graphics) followed by prompts that were minimally two-dimensional. When additional information was needed, it was presented with another mini-scenario. Each task had at least one three-dimensional prompt. The newly developed tasks and prompts were further refined by a task review committee that met from Sept. 9–12, 2019, and consisted of NDE staff, NWEA staff, and 15 educators recruited by NDE who were not involved in writing the tasks. The tasks and prompts were reviewed for content and bias concerns. NWEA content specialists and psychometricians created three forms. Each task developed is present on at least two forms.

2.9. Content Alignment

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

2.9.1. Alignment and Adaptive Testing

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

As 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 et al., 2012) is that DOK levels follow the indicator progression. While DOK may increase across achievement levels, the DOK level should not automatically increase with the achievement level increase. What may be required from a learning theory perspective is that students have support accessing the standards, such as with visual supports demarcating a manipulation of an item context feature. They then may access the standards without the visual aids, followed by accessing the standards at a higher DOK level. Thus, if the item development is purposeful to the progression, DOK specifications are not required as a constraint conditional that items are measuring what the ALDs say they are.

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

2.9.2. 2019 Mathematics Alignment Study

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

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

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

2.10. Universal Design

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

2.11. Sensitivity and Fairness

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

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

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

- Distract, upset, or confuse in any way
- Contain inappropriate or offensive topics
- Require construct-irrelevant knowledge or specialized knowledge
- Favor students from certain language communities
- Favor students from certain cultural backgrounds
- Favor students based on gender

- Favor students based on social economic issues
- Employ idiomatic or regional phrases and expressions
- Stereotype certain groups of people or behaviors
- Favor students from certain geographic regions
- Favor students who have no visual impairments
- Use height, weight, test scores, or homework scores as content or data in an item

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

2.12. Test Construction (ELA and Mathematics)

The 2020 ELA and mathematics paper-pencil forms were based on the 2018 forms, with a few items being replaced as needed. The online adaptive tests were produced by selecting the item pools, building the test models that configured the engine and provided the constraints, running simulations, approving the results, and conducting user acceptance testing (UAT).

2.12.1. Fixed-Forms

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

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

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

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

2.12.2. MAP Growth Item Selection

To ensure a successful transition to a through-year solution, a linking study between NSCAS and MAP Growth is needed. The goals of the linking study are to (1) investigate the degree to which MAP Growth items could be brought onto the NSCAS scale and achieve comparable results to NSCAS and (2) project a MAP Growth RIT score from the NSCAS items. A common item linking study between NSCAS and MAP Growth was planned to be conducted using the 2020 data but could not be completed due to testing cancellation. The study will be conducted in 2021.

NSCAS and MAP Growth use different item players, which means ELA reading passages are formatted differently; mathematics items have different calculator rules regarding when calculators can be used and different types of calculator; and item display settings such as color, text font, and layout are different. As a result, embedding MAP Growth items directly into the NSCAS player would not allow the linking constant from NSCAS to MAP Growth to be obtained. Therefore, a subset of items on MAP Growth tests that are the least different in formatting from NSCAS were selected for the common item linking study.

Following NDE's approval, NWEA selected the most NSCAS-like items in the MAP Growth item pool to be placed at the end of the 2020 NSCAS forms. These items will be spiraled from the pool instead of being embedded in the typical field test slots within the operational test. Including the MAP Growth items at the end of the forms made the 2020 NSCAS tests slightly longer (i.e., from 48 to 53 items), but any cognitive confusion over formatting differences would not affect operational scores as they would be presented after all the NSCAS items.

To include the most NSCAS-like items, MAP Growth Reading items were included if they are associated with passages, and mathematics items were included if their calculator use is aligned with that of NSCAS. Specifically, reading items were removed if they were not associated with any passage or if any passages had less than three items because all NSCAS Reading Vocabulary and Reading Comprehension items are associated with passages. Mathematics items were removed if they were flagged during review for being marked as not at grade level in the recent EdMetric alignment study (EdMetric, 2019), marked with "calculator at a grade NSCAS does not allow a calculator," or "wrong calculator for the grade for NSCAS." It resulted in no calculator items in mathematics across all grades.

Further, there was a difference in the percentage of items for each reporting category between the Nebraska MAP Growth item pool and the NSCAS assessments based on the blueprint. A decision was made to select MAP Growth items based on the percentage of items for each reporting category of the Nebraska MAP Growth item pool so that the selected MAP Growth items for the 2020 NSCAS forms will represent the item distribution in the Nebraska MAP Growth item pool.

Approximately 150 of those items per grade and content area were then selected for inclusion on the Spring 2020 NSCAS forms. Specifically, 110 MAP Growth Reading and the 40 MAP Growth Language Usage are included for NSCAS ELA, and 150 MAP Growth Mathematics are included for each grade. The targeted minimum n-count for each MAP Growth item is 750, and a total of approximately 900 MAP Growth items are included across each grade and content area.

2.13. Data Review

Data review did not occur in 2020 due to the administration cancellation.

Section 3: Test Administration and Security

The Spring 2020 NSCAS General Summative testing window was scheduled from March 16 to April 24, 2020, and the make-up testing window was scheduled from April 27 to May 1, 2020. The tests were to be untimed and administered online via the NWEA Comprehensive Assessment Platform (CAP). However, the 2020 administration was cancelled due to COVID-19. This chapter summarizes the events that occurred prior to the cancellation such as administration training and user acceptance testing (UAT).

3.1. User Roles and Responsibilities

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

Table 3.1. User Roles and Responsibilities

User	Roles and Responsibilities
District Assessment Contact	Responsible for coordinating the testing activities of all schools within their districts, including coordinating the test schedules of the schools within the district and setting up test sessions.
School Assessment Coordinator	Responsible for coordinating the testing activities within their schools, including the 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.
Proctor	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 the following:

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

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

- Test schedule
- Administration preparation
- Students will special needs
- Testing conditions
- Scratch paper and reference sheets
- Security

3.2. Administration Training

In addition to district- and school-held trainings, NWEA, in collaboration with NDE, held two trainings for district leaders in advance of testing. The Fall 2019 regional workshops were half-day, in-person workshops held across multiple regions of the state from October 9–16, 2019. Information on the spring summative administration including test sessions, accessibility, and student rostering was presented. The three summative test administration workshops in February 2020 were two-hour virtual sessions that provided important information on the NSCAS assessments. Table 3.2 presents the locations and number of participants based on the registration numbers for the Fall 2019 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 B presents the PowerPoint presentations for each training.

Table 3.2. Fall 2019 Regional Workshop Locations and Participation

Date	Location	Approximate #Participants
Oct. 9, 2019	Scottsbluff	37
Oct. 10, 2019	Kearney	75
Oct. 11, 2019	Norfolk	40
Oct. 15, 2019	Lincoln	35
Oct. 16, 2019	Omaha	35

Table 3.3. Summative Test Administration Workshop Dates and Participation

Date	#Participants
Feb. 14, 2020	49
Feb. 17, 2020	39
Feb. 20, 2020	26

3.3. Item Type Samplers

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

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

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

3.4. Accommodations and Accessibility Features

Table 3.4 presents the accessibility supports intended to be available for the Spring 2020 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 General Summative & Alternate Accessibility Manual (NDE, 2019).

- Accommodations are changes in procedures or materials that ensure equitable access to instructional and assessment content and generate valid assessment results for students who need them. Embedded accommodations (e.g., TTS) 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 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 General Summative & Alternate Accessibility Manual.

Table 3.4. Accommodations and Universal Features

Support	Description
Embedded Accommodations	
Text-to-speech (TTS)	A student can use this feature to hear audio of the item content.
Non-Embedded Accommodations	
Paper-pencil	A student takes the assessment on paper instead of online.
Mathematical 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)

Support	Description
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.
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.

Support	Description
Non-Embedded Universal Features	
Alternate location	Student takes test at home or in a care facility (e.g., hospital) with direct supervision. For facilities without internet, a paper-pencil test will be allowed.
Directions	Test administrator rereads, simplifies or clarifies directions aloud for student as needed.
Color contrast	Background color can be adjusted based on student's need.
Cultural considerations	The student receives a paper-pencil form due to specific belief or practice that objects to the use of technology. This student does not use technology for any instructional related activities. Districts must contact NDE to request this accessibility feature.
Noise buffer/headphones	The student uses noise buffers to minimize distraction or filter external noise during testing.
Redirection	Test administrator directs/redirects student focus on test as needed.
Scratch paper (plain or graph)	The student uses blank scratch paper, blank graph paper, or an individual erasable whiteboard to make notes or record responses.
Setting	The student is provided a distraction-free space or alternate, supervised location (e.g., study carrel, front of classroom, alternate room).
Student reads test aloud	The student quietly reads the test content aloud to self. This feature must be administered in a setting that is not distracting to other students.

3.5. User Acceptance Testing (UAT)

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

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

From February 10–12, 2020, 29 testers participated in UAT. Each were assigned 1–9 tests. Testers are typically NWEA staff who are at least somewhat familiar with how the functionality is supposed to interact. In addition to a training and kick-off on the process and a checklist of tasks, technical product managers are present at the kick-off meeting to describe the UAT process overall, expected enhancements to functionality, and known issues. Use cases describing each item feature and other support documentation are provided to testers to review prior to UAT. Testers should spend 1–2 hours reviewing existing documentation prior to performing testing. They are also encouraged to explore the item type sampler beforehand.

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

3.6. Student Participation

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

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

3.6.1. Paper-Pencil Participation Criteria

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

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

3.6.2. Participation of English Language Learners (ELLs)

According to the Elementary and Secondary Education Act (ESEA), ELLs are students who have a native language other than English, OR who came from an environment where a language other than English has had a significant impact on their level of English proficiency, AND whose difficulties in speaking, reading, writing, or understanding the English language may be sufficient to deny the individual (i) the ability to meet the state's proficient level of achievement on state assessments, (ii) the ability to successfully achieve in classrooms where the language of instruction is English, or (iii) the opportunity to participate fully in society.

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 General Summative. Districts should have reviewed the following guidelines before testing:

- In determining appropriate linguistic supports for students in the NSCAS system, districts should use the NSCAS General Summative & Alternate Accessibility Manual (NDE, 2019).
- Districts must be aware of the difference between linguistic supports (accommodations for ELLs) and modifications.
- For students learning the English language, linguistic supports are changes to testing procedures, testing materials, or the testing situation that allow the students meaningful participation in the assessment. Effective linguistic supports for ELL students address their unique linguistic and socio-cultural needs. Linguistic supports for ELL students may be determined appropriate without prior use during instruction throughout the year.
- Modifications are adjustments or changes in the test or testing process that change the test expectation, grade level, construct, or content being measured. Modifications are not acceptable in the NSCAS assessments.

3.6.3. Participation of Recently Arrived Limited English Proficient Students

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

3.7. Test Security

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

As part of NDE’s security policy, the principal of each school participating in the NSCAS General Summative assessments was required to complete a Building Principal Security Agreement and return it to the Statewide Assessment Office by Nov. 8, 2019. District Assessment Contacts were required to complete and sign the District Assessment Contact Confidentiality of Information Agreement and return it to the Statewide Assessment Office by Nov. 9, 2019. School districts were bound to hold all certificated staff members 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.

Due to the cancellation of Spring 2020 testing due to COVID-19, online test security, paper-pencil test security, and Caveon test monitoring procedures were not executed. Following the suspension of testing, student test tickets, generated after test session creation by a School Assessment Coordinator or District Administrator Contact, that contained student-level password information were to be securely destroyed by districts. Districts were also instructed to securely destroy paper test materials or securely return test materials as instructed in the Paper/Pencil Test Administration manual. EDS and Caveon test security procedures that would have been implement for the 2020 administration are described below

3.7.1. EDS Test Security

3.7.1.1. <u>Physical Warehouse Security</u>

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 General 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. Documents are stored until the second week of January following the test administration or until NDE provides express written consent to destroy them.

3.7.1.2. <u>Secure Destruction of Test Materials</u>

EDS will manage the secure destruction of test materials during the first two weeks of January 2021. Using the information from the long-term storage database, EDS will retrieve the documents and systematically destroy them through a secure shredding process. The shredding company uses a high-capacity mobile onsite document destruction vehicle that provides the most advanced document destruction technology in the industry. The shred trucks, equipped with a 20-inch monitors so EDS staff may monitor the documents going into and being expelled in a pulverized state, provide the quickest, most complete, and most confidential destruction of sensitive documents. Every sensitive document is pulverized using a *hammermill* process that creates the smallest pieces in the document destruction industry.

After the test materials destruction process is complete, the shredding company provides a certificate of destruction that will remain on file at EDS. The long-term storage database will be updated to reflect that the materials have been destroyed. During the first two weeks of January 2021 and upon written approval, EDS will also delete the answer document and test book images from the server hard drive and all backup drives. The deletion process will securely erase the data to ensure that the images cannot be retrieved through data restorative means. EDS will provide NDE with archives of all data files prior to deletion, upon request.

3.7.1.3. Shipping Security

Hardcopies of the prepress test materials for proof approval were provided to NWEA via traceable courier and tracked to ensure arrival. All proofs arrived with no incident. For district shipments, EDS used the secure and trackable UPS ground and two-day shipping services to send materials to and receive materials from districts. The system interfaced with the in-house UPS shipping system, thus making certain that deliveries were made to accurate and correct addresses. Address verification was used to ensure that the materials were shipped to known UPS addresses before shipping.

To ensure correct deliveries to all sites, all boxes belonging to a school or district were numbered and labeled with unique barcode numbers tracked in the system. Every box was assigned a unique UPS tracking number and the numbers were uploaded to the Materials Tracking module allowing EDS, districts, NWEA, and NDE to track all shipments and diagnose problems early. One-hundred percent of shipments containing test documents were tracked and monitored to and from sites. EDS resolved all shipping issues in a timely manner and no material reships were required.

3.7.1.4. Electronic Security of Test Materials and Data

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

3.7.2. Caveon Test Security

Caveon Web Patrol intended to investigate the NSCAS assessments online with the primary goals of detecting, reporting, and eliminating, where possible, exposures and infringing content from the individual assessments. During the administration windows, Caveon Core was used as a secure incident reporting and encrypted materials storage platform for NWEA or NDE. Live test items provided to Caveon Web Patrol by NWEA were protected by placing them securely on a non-networked air-gapped computer. Access to those live items was only authorized to be used by Caveon’s Executive Web Patrol Manager.

Live items were never intended to be used for searching but only for verification in the case of potential infringements. Use of materials, other than live test items, were also limited to only Caveon Web Patrol employees assigned to this project. Each employee signed non-disclosure agreements before engaging in work for NWEA and NDE and was trained in how to protect their security online using anonymous email addresses, Virtual Private Networks, and prescribed processes for accessing, transferring, and handling of secure client files and associated information.

3.8. Partner Support

The NWEA Partner Support Services team provided implementation and technical support throughout the 2019–2020 school year for the NSCAS General Summative assessments. This team provided resources to support Nebraska and its educators, assisting with generating roster files, configuration of the assessment program, accessing online reports, and general questions with the use of the online assessment system. NWEA provided phone, email, and chat support to schools and educators from 8:00 a.m. to 5:00 p.m. Central Time (CT) Monday through Friday, and 7:00 a.m. to 5:00 p.m. CT during the testing windows, as described in Table 3.5. Table 3.6 presents the number of cases presented to the Partner Support team by case type for the entire 2019–2020 school year from July 2019 to June 2020 for the NSCAS tests. More than half of the cases were related to testing (i.e., administration questions).

Table 3.5. Partner Support Communication Options

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

Table 3.6. Number of NSCAS Cases to Partner Support in 2019–2020

Case Type	#Cases	% of Total Cases
Student Mobility	1	0.6
Reports	47	28.3
Navigation	8	4.8
Setup and Management	66	39.8
Testing	44	26.5
Total	**166**	**100.0**

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

Section 4: Scoring and Reporting

Scoring and reporting did not take place in 2020 due to the administration cancellation. As a result, student test data were not collected and there were no answer sheets to scan. This chapter summarizes the decisions and processes that still occurred in 2020 such as scoring rules and the continued use of the Matrix.

4.1. Scoring Rules

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

Table 4.1. Attemptedness Rules for Scoring

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

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

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

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

4.2. Paper-Pencil Scoring

Due to the administration cancellation, there were no answer sheets to scan.

4.3. Score Reporting Methods

Student performance on the NSCAS assessment is reported as a scale score and achievement level. Scale scores range from 2220 to 2890 for ELA and 1000 to 1550 for mathematics, as shown in Table 4.2. Science was intended to be a field test and no score data were to be produced. 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

	Scale Score Ranges*		
Grade	Developing	On Track	CCR Benchmark
ELA			
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			
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 as intended to be a field test and no score data were to be produced.

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

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

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

Table 4.3. Achievement Level Descriptions

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

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

Table 4.4. Reporting Categories

Content Area	Reporting Categories
ELA	• Reading Vocabulary • Reading Comprehension • Writing Skills
Mathematics	• Number • Algebra • Geometry • Data

4.4. Report Summary

The following reports were prepared for the 2020 NSCAS test administration, although they were never used due to the administration cancellation. Appendix C presents examples of each report.

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

ISRs show a student's performance on the NSCAS General Summative tests. If a non-tested code (NTC) is applied to any content area, the student's achievement level scores and proficiency by reporting category within the respective content area are reported as affected by the NTC, as defined in Table 4.5. If a student has an NTC of INV, PAR, SAE, STR, or UTT assigned to their test, the automatically assigned score displays with a score of the lowest scale score for that grade and content area.

Table 4.5. Non-Tested Codes (NTCs)

Code	Translation	Description	Score / Reporting
ALT	Alternate Assessment	Student took the NSCAS Alternate assessment and is not included in results from this testing vendor.	• No scale score provided for a test with this code • Score suppressed • State data file only
EMW	Emergency Medical Waiver	Student was not tested because of an approved emergency medical waiver.	• No scale score provided for a test with this code • Score suppressed • NTC reported
EXP	Exception	Due to testing irregularities, the assessment was not scored.	• Score not included in any reports or calculations
INV	Invalid	Student's assessment was invalidated, such as a security breach.	• Score as LOSS • NTC + LOSS on a specific report(s)

Code	Translation	Description	Score / Reporting
NLE	No Longer Enrolled	Student was not enrolled in the district/school during the testing window(s.)	• No scale score provided for a test with this code • NTC on ISR • Exclude from aggregate reports
OTH	Other	Student's score was removed from performance for reasons not covered by other descriptions.	• Score suppressed • Data file only
PAR	Parental Refusal	Student was not tested because of a written request from parent or guardian.	• Score as LOSS • NTC + LOSS on ISR • Include in aggregate reports
PPE	Paper-Pencil Expected	A separate paper-pencil test event is expected for this student. This test event should not be included in reports. Refer to the paper-pencil test event for this student instead.	• Score not included in any reports or calculations
RMV	Removed	Student was removed from the file for reasons not otherwise covered.	• Score suppressed • Suppress from all reports or calculations
SAE	Student Absent for Entire Testing Window	Student was absent from school for the entire testing window(s).	• Score as LOSS • NTC + LOSS on ISR • Include in aggregate reports
STR	Student Refusal	Student was not tested due to student refusal to participate	• Score as LOSS • NTC + LOSS on ISR • Include in aggregate reports
TXP	Tested at External Program	Student is attending an external program and test scores should be transferred to the district/school of accountability.	• Score not included in any reports or calculations
UTT	District Unable to Test Student	District was unable to test student during the testing window and none of the other NTCs are applicable.	• Score as LOSS • NTC + LOSS on ISR • Include in aggregate reports

The School Roster report lists students required to take the NSCAS tests and presented a report of their performance. The size of this document depends on the class size. The School Achievement Level Summary report presents a summary of performance and demographics for all students at a school by grade required to take the NSCAS tests. The District Achievement Level Summary report is for internal district use only and is required for state and federal reporting purposes. The State Achievement Level Summary report presents the average state performance based on demographics for the NSCAS tests.

4.5. Reporting Process

Reporting did not occur in 2020 due to the testing cancellation.

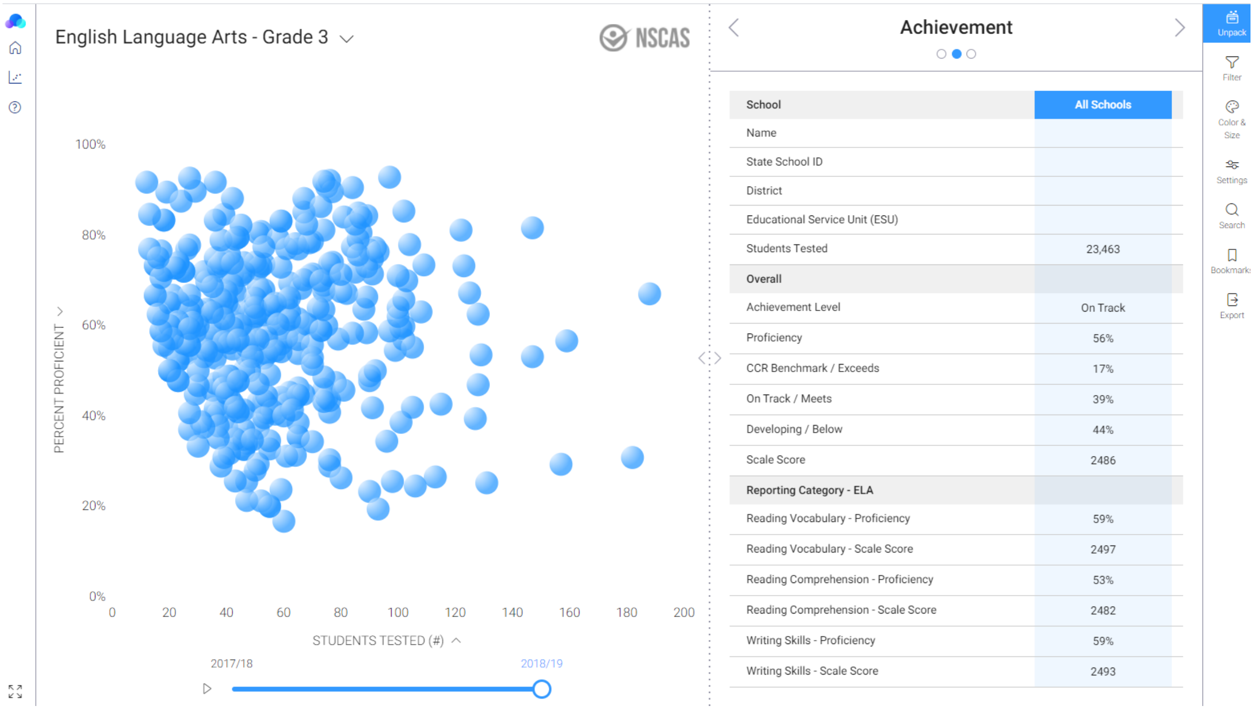
4.6. Matrix

Even though 2020 testing was cancelled, Education Strategy Consulting (ESC) is maintaining the Matrix with historical info for reference. Users still have access to this tool, but it is not reporting what was completed in 2020.

NWEA used ESC's tools to view web-based visualizations for the NSCAS assessments, including combinations of aggregate and disaggregate information of results by demographics and other filtering options. This web portal, referred to as the Matrix, allows users to save and print specific plot and screen images from the interactive visualization. Users can interact with and explore many different levels of information to answer targeted questions about their district, school, or state. The main feature of this tool is an interactive scatterplot designed to display longitudinal data, as shown in Figure 4.1, Figure 4.2, and Figure 4.3. The X and Y axes are modifiable. Users can construct a spreadsheet from all the available variables within the visualization via the export function. This feature allows for easy access to high-quality data that has gone through rigorous auditing. Users can then explore and sort data to meet their individual needs. Suppression rules are applied to the data for all users. For example, all data is suppressed for a school if the number of tested students was less than 10.

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

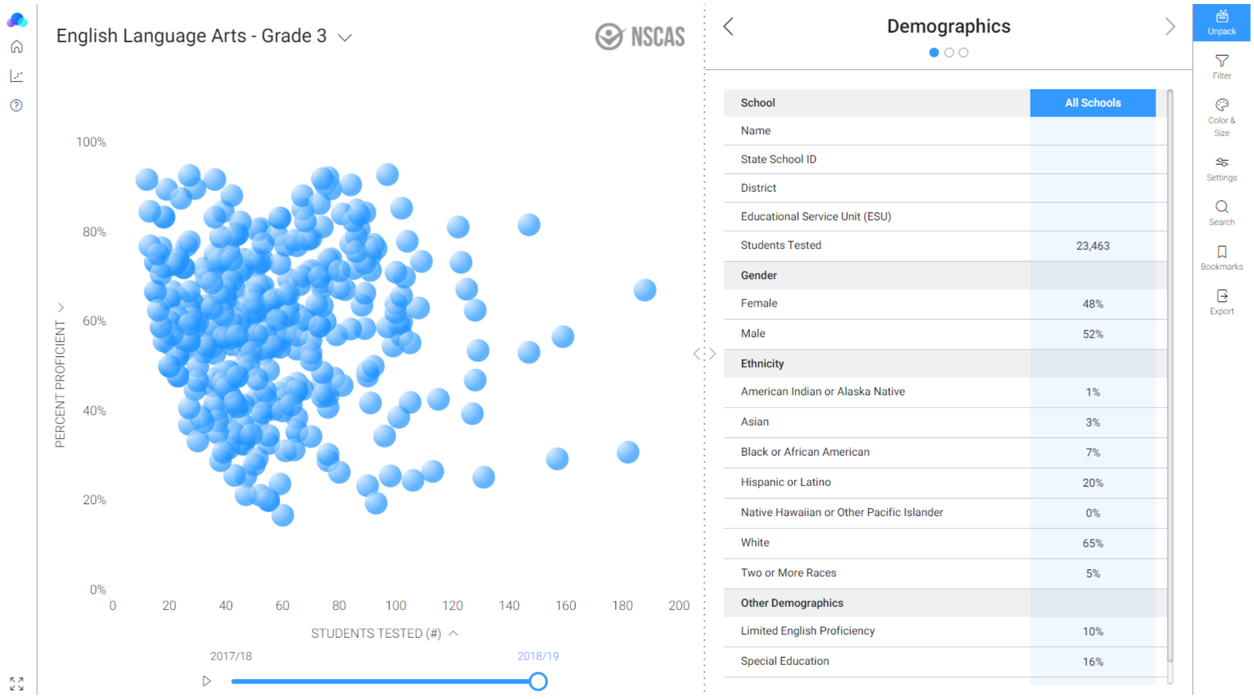
Figure 4.1. Matrix Example: Percent Proficient



Achievement

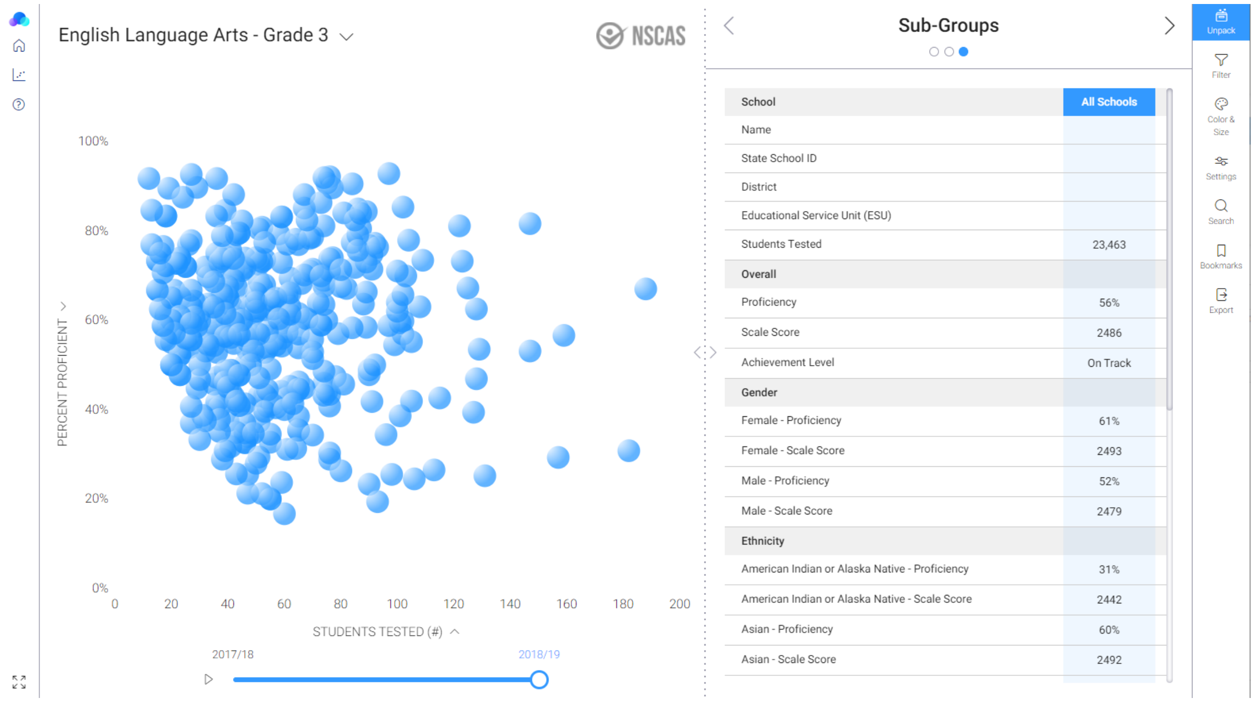
School	All Schools
Name	
State School ID	
District	
Educational Service Unit (ESU)	
Students Tested	23,463
Overall	
Achievement Level	On Track
Proficiency	56%
CCR Benchmark / Exceeds	17%
On Track / Meets	39%
Developing / Below	44%
Scale Score	2486
Reporting Category - ELA	
Reading Vocabulary - Proficiency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

Figure 4.2. Matrix Example: Scale Score by Demographics



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

Figure 4.3. Matrix Example: Scale Score by Sub-Groups



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

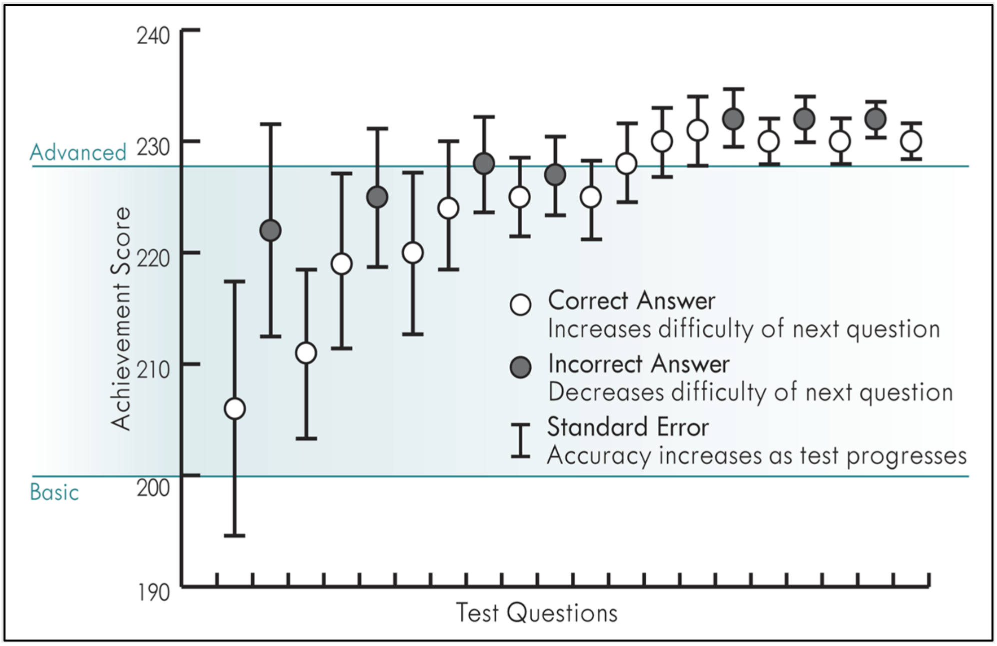
Section 5: Constraint-Based Engine

Since there was no empirical student for Spring 2020, a post-administration evaluation study did not occur. This section only includes results from the 2020 simulation studies.

5.1. Overview

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

Figure 5.1. Adaptive Engine Overview

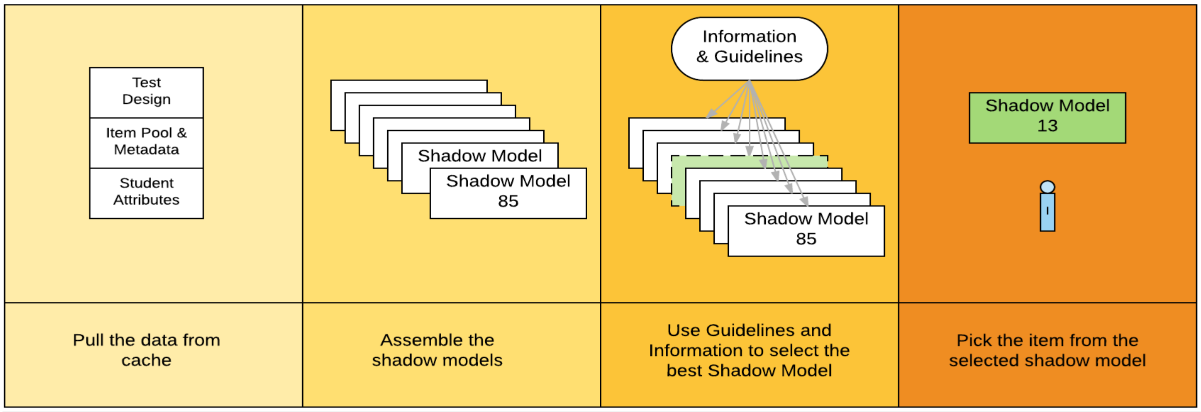


Items were selected based on item difficulty. The goal of the constraint-based engine's item selection was to provide a test that meets "must-have" (hard) constraints and "nice-to-have" (soft) constraints. Examples of hard constraints are all item selection constraints, such as all levels of standards, 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 blueprint while providing the maximum information about the student based on the student's momentary ability estimate: (1) shadow test approach and (2) a variation of the weighted penalty model.

As shown in Figure 5.2, the shadow test approach (Van der Linden & Reese, 1998) selects items based on the required aspects of the blueprint, and a new valid shadow model is selected upon each update to the student's momentary theta. In other words, this approach uses the student's answer to the last item to create shadow models that are waiting "in the shadows" while the student answers the current item. When the student responds to the item, that answer is used to select the next correct shadow model. Because multiple shadow models can be drawn from an item pool, a variation of the weighted penalty model (Segall & Davey, 1995) then selects which shadow model is optimal based on additional content guidelines while ensuring the most representative sample for linking and field test items. The shadow model with the smallest penalty is selected when multiple shadow tests meet the required attributes of the test and have similar information.

Figure 5.2. Shadow Test Approach



5.2. Engine Simulations: ELA and Mathematics

Pre-administration engine simulations are important evidence, along with post-administration evaluation studies and analyses when applicable, for confirming interpretation and test score use arguments regarding student proficiency with the state standards. Pre-administration simulations were conducted prior to the Spring 2020 operational testing window to evaluate the constraint-based engine's item selection algorithm and estimation of student ability based on the blueprint. The simulation tool used the operational constraint-based engine, thereby providing results with the same properties and functionality as what would be seen operationally. Detailed information regarding the simulation study can be found in the full report (NWEA, 2020).

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

5.2.1. Evaluation Criteria

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

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

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

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

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

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

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

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

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

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:

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

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

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

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

where σ is defined as:

$$\sigma = \mathrm{E}\left\{[I(\theta)]^{-1/2}\right\} \tag{5.8}$$

5.2.2. Blueprint Constraint Accuracy

Table 5.1 presents the blueprint constraint results at the reporting category level for the pre-administration simulation study. The number of items and points at the reporting category level resulted in a 100% match for all grades based on the blueprint. Results were also provided at the indicator level by passage type selection, DOK level, and item range requirements (NWEA, 2020). While most DOK levels also resulted in a 100% match, some indicators did not because the constraint-based engine used DOK level as a guideline or a “nice to have” given the limited number of items at a specified DOK level for some indicators. Passage type for ELA also resulted in a less-than 100% match for some indicators.

Table 5.1. Blueprint Constraint by Reporting Category—Simulations

		#Items			#Points		
Grade	**Reporting Category**	**Min.**	**Max.**	**%Match**	**Min.**	**Max.**	**%Match**
ELA							
3	Reading Vocabulary	8	9	100.0	9	10	100.0
	Reading Comprehension	22	24	100.0	26	28	100.0
	Writing Skills	8	8	100.0	12	13	100.0
4	Reading Vocabulary	9	9	100.0	9	10	100.0
	Reading Comprehension	24	24	100.0	28	28	100.0
	Writing Skills	8	8	100.0	11	12	100.0
5	Reading Vocabulary	8	9	100.0	9	10	100.0
	Reading Comprehension	22	24	100.0	28	30	100.0
	Writing Skills	10	10	100.0	14	14	100.0
6	Reading Vocabulary	8	9	100.0	9	10	100.0
	Reading Comprehension	22	23	100.0	27	29	100.0
	Writing Skills	9	10	100.0	13	16	100.0
7	Reading Vocabulary	9	9	100.0	10	10	100.0
	Reading Comprehension	22	22	100.0	28	28	100.0
	Writing Skills	10	10	100.0	12	12	100.0
8	Reading Vocabulary	9	9	100.0	9	10	100.0
	Reading Comprehension	21	24	100.0	28	31	100.0
	Writing Skills	11	11	100.0	15	16	100.0
Mathematics							
3	Number	16	16	100.0	17	18	100.0
	Algebra	6	6	100.0	7	8	100.0
	Geometry	11	11	100.0	12	12	100.0
	Data	8	8	100.0	9	9	100.0
4	Number	17	18	100.0	18	19	100.0
	Algebra	10	11	100.0	11	12	100.0
	Geometry	8	9	100.0	9	10	100.0
	Data	6	7	100.0	7	8	100.0
5	Number	16	17	100.0	17	18	100.0
	Algebra	10	10	100.0	11	11	100.0
	Geometry	8	9	100.0	9	10	100.0
	Data	5	6	100.0	6	7	100.0

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

5.2.3. Item Exposure Rates

Table 5.2 presents the item exposure rates from the engine simulation 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 Table 5.2, “Total” is the total number of items in the operational item pool. “Fixed” is the number of horizontal linking items. “CAT” indicates the pool of adaptive items that students may be administered during the test. “Unused” shows the number and percentage of unused items that were never administered to students. All horizontal linking items were also part of the item exposure rate calculation. Horizontal Form 1 (i.e., the core form) given to all students had a 100% exposure rate and is therefore included in the 81–100% exposure rate bin, and the horizontal linking Set A and Set B each had an approximately 50% exposure rate for Grades 4–7 and are therefore included in the 41–60% exposure rate bin.

In general, ELA had a higher unused percentage compared to mathematics. One possible reason could be that ELA includes passages and the constraint of a minimum of four items per passage while mathematics does not. In particular, ELA Grade 8 shows a high percentage of unused items (81.09%), which was also observed in 2019 (77.5%). One reason was the large proportion of 2-point items. To meet the blueprint, ELA Grade 8 needs at least seven polytomous items for Reading Comprehension and four polytomous items for Writing Skills. This requirement was confounded with DOK requirements and selected anchor items, plus all Reading Comprehension items were associated with passages. The issue got worse in 2020, resulting in failed simulations. NWEA recommended making both DOK 2 and DOK 3 constraints guidelines at the test level, considering that DOK at the indicator level were controlled as guidelines and that it would not require the adjustments to the blueprint (email communication, December 13, 2019). After NDE approved it at the subsequent weekly meeting with NWEA, the change was implemented, resulting in a similar result of a high percentage of unused items.

Table 5.2. Item Exposure Rates—Simulations

						Exposure Rate											
	#Items					0–20%		21–40%		41–60%		61–80%		81–99%		100%	
Grade	Total	Fixed	Used	Unused	Unused %	N	%	N	%	N	%	N	%	N	%	N	%
ELA																	
3	605	21	274	331	54.71	223	81.39	23	8.39	2	0.73	3	1.09	2	0.73	21	7.66
4	598	28	337	261	43.65	279	82.79	23	6.82	21	6.23	–	–	–	–	14	4.15
5	508	28	389	119	23.43	336	86.38	19	4.88	18	4.63	2	0.51	–	–	14	3.60
6	552	28	311	241	43.66	247	79.42	20	6.43	21	6.75	9	2.89	–	–	14	4.50
7	520	28	302	218	41.92	252	83.44	7	2.32	21	6.95	4	1.32	4	1.32	14	4.64
8	587	21	107	480	81.77	58	54.21	11	10.28	2	1.87	6	5.61	2	1.87	28	26.17
Mathematics																	
3	553	21	515	38	6.87	483	93.79	9	1.75	2	0.39	–	–	–	–	21	4.08
4	432	28	400	32	7.41	364	91.00	6	1.50	15	3.75	1	0.25	–	–	14	3.50
5	444	28	418	26	5.86	375	89.71	14	3.35	15	3.59	–	–	–	–	14	3.35
6	550	28	511	39	7.09	475	92.95	7	1.37	15	2.94	–	–	–	–	14	2.74
7	480	28	436	44	9.17	399	91.51	7	1.61	16	3.67	–	–	–	–	14	3.21
8	451	21	422	29	6.43	394	93.36	6	1.42	1	0.24	–	–	–	–	21	4.98

5.2.4. Score Precision and Reliability

The pre-administration simulation study provided precision ability estimations that showed how well the constraint-based engine recovered students' true ability based on the item pool. It included the standard deviation of estimated theta, mean SEM, SEM by deciles, and marginal reliability. The following indexes were used to examine the functionality of the constraint-based engine during the simulations:

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

Table 5.3 presents the results of the precision ability estimation from the simulations. The 2020 simulations included a sample nine times larger than previous years (9,000 vs. 1,000), which makes it more likely to detect significant p-values between true and estimated theta. Because this study did not involve an actual test administration, the constraint-based engine is not scoring student responses but is instead simulating whether a student got items correct or incorrect based on the student's ability. Because a student's true theta is known, the engine should be able to recover the student's theta after administering all the items. This is the estimated theta. The null hypothesis is that there is no difference between true and estimated theta.

For the overall scores across all students, the mean biases are small, ranging from -0.01 to 0.00 for both ELA and mathematics, and the p-value for the z-test supports the null-hypothesis that there is not a significant difference between the simulated students' true and final estimated thetas. The MSE is also relatively small, showing that the constraint-based engine typically recovered a value near the student's true theta.

Table 5.3. Mean Bias of the Ability Estimation (True - Estimated)—Simulations

	Bias		*P*-Value for Z-Test		95%	99%
Grade	Mean	SE		MSE	Coverage	Coverage
ELA						
3	0.00	0.00	0.80	0.11	5.01	0.83
4	-0.01	0.00	0.35	0.11	4.78	0.81
5	-0.01	0.00	0.26	0.10	4.84	0.90
6	0.00	0.00	0.99	0.10	5.23	0.88
7	0.00	0.00	0.67	0.11	4.69	0.72
8	0.00	0.00	0.98	0.10	4.97	0.94
Mathematics						
3	-0.01	0.00	0.67	0.12	5.28	1.16
4	0.00	0.00	0.98	0.11	5.03	0.89
5	-0.01	0.00	0.60	0.12	5.37	1.01
6	0.00	0.00	0.88	0.10	5.03	0.90
7	0.00	0.00	0.82	0.11	5.09	0.97
8	0.00	0.00	0.73	0.11	4.76	0.94

Table 5.4 presents the score precision and reliability estimates for the simulation study, 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 for the overall scores is from 0.90 to 0.91 for ELA and 0.93 to 0.94 for mathematics. These results indicate that, overall, the score precision is relatively good.

Table 5.4. Score Precision and Reliability—Simulations

Grade	Average #Items	SD of Estimated Theta	Mean SEM	RMSE	Reliability
ELA					
3	41	1.07	0.32	0.33	0.91
4	41	1.09	0.32	0.33	0.91
5	41	1.03	0.32	0.32	0.90
6	41	1.01	0.30	0.30	0.91
7	41	1.07	0.32	0.33	0.91
8	41	1.02	0.31	0.32	0.90
Mathematics					
3	41	1.34	0.33	0.33	0.94
4	41	1.26	0.33	0.33	0.93
5	41	1.36	0.33	0.34	0.94
6	41	1.29	0.32	0.32	0.94
7	41	1.22	0.32	0.32	0.93
8	41	1.36	0.33	0.33	0.94

Table 5.5 presents the average SEM by decile of the true overall proficiency score, including the overall student ability distribution, for the simulation study. 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 0.30 to 0.33.

Table 5.5. SEM by Deciles—Simulations

	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	0.35	0.31	0.31	0.30	0.30	0.30	0.31	0.32	0.33	0.39	0.32
4	0.33	0.30	0.29	0.30	0.30	0.31	0.31	0.33	0.35	0.43	0.32
5	0.32	0.30	0.30	0.30	0.30	0.31	0.31	0.32	0.33	0.38	0.32
6	0.34	0.30	0.28	0.27	0.27	0.27	0.28	0.29	0.31	0.37	0.30
7	0.36	0.32	0.30	0.30	0.29	0.29	0.30	0.31	0.33	0.41	0.32
8	0.35	0.31	0.30	0.29	0.29	0.29	0.30	0.31	0.32	0.38	0.31
Mathematics											
3	0.34	0.32	0.31	0.31	0.31	0.31	0.32	0.33	0.35	0.42	0.33
4	0.36	0.32	0.31	0.30	0.30	0.31	0.31	0.32	0.34	0.38	0.33
5	0.35	0.31	0.30	0.30	0.30	0.31	0.32	0.33	0.36	0.44	0.33
6	0.35	0.31	0.30	0.30	0.30	0.30	0.31	0.32	0.34	0.37	0.32
7	0.38	0.33	0.31	0.30	0.30	0.30	0.30	0.31	0.32	0.35	0.32
8	0.39	0.32	0.31	0.30	0.30	0.30	0.31	0.32	0.33	0.38	0.33

5.3. Engine Simulations: Science Field Test

Engine simulations were also conducted for the NSCAS Science assessment for Grades 5 and 8 that was to be field tested in Spring 2020 (now moved to Spring 2021). The results are presented in the sections below. Based on these simulation results, NDE will be able to administer the NSCAS Science field test using the NWEA constraint-based engine as planned. The engine's population exposure control works as intended to ensure that each test form will be administered to a representative sample of Nebraska students as defined by gender and ethnicity demographic characteristics. The engine also administers the fixed forms as intended. Prompts within a task will be administered in a fixed pre-specified order, survey questions will be administered at the end of the test forms, and the position of most tasks on a form varies across students to reduce task position effect.

5.3.1. Population Exposure Control

Table 5.6 presents the number and percentage of simulated students who received each form by gender and ethnicity. Based on a comparison of the Nebraska general population demographic distributions, each form was delivered to a representative sample of Nebraska students, demonstrating that the proportions set in the engine population exposure control are representative of the Nebraska general student population in terms of gender and ethnicity. Because of this, and because each task and its associated prompts appear on at least two of the three forms based on the test design, it can be reasonably assumed that each task and its associated prompts were also delivered to a representative sample of Nebraska students. These results suggest that the population exposure control function of the constraint-based engine worked well.

Table 5.6. Fixed-Form Demographic Distribution

Demographic Sub-Group	Grade 5						Grade 8					
	Form A		Form B		Form C		Form A		Form B		Form C	
	N	%	N	%	N	%	N	%	N	%	N	%
Total N	1,001	100.0	999	100.0	1,000	100.0	1,000	100.0	999	100.0	1,001	100.0
Gender												
Female	489	48.9	488	48.8	488	48.8	489	48.9	488	48.8	489	48.9
Male	512	51.1	511	51.2	512	51.2	511	51.1	511	51.2	512	51.1
Ethnicity												
American Indian	13	1.3	11	1.1	14	1.4	13	1.3	11	1.1	16	1.6
Asian	31	3.1	27	2.7	24	2.4	28	2.8	23	2.3	29	2.9
Black	63	6.3	67	6.7	68	6.8	72	7.2	68	6.8	70	7.0
Hispanic	203	20.3	202	20.2	202	20.2	197	19.7	197	19.7	196	19.6
White	647	64.6	649	65.0	647	64.7	657	65.7	656	65.7	652	65.1
Two+ Races	44	4.4	43	4.3	45	4.5	33	3.3	44	4.4	38	3.8

5.3.2. Test Design Accuracy

Table 5.7 presents the number of students who received each task during the simulations, including the number of total responses (i.e., number of prompts × the total number of students who received the task). Based on a comparison of the field test design in Table 2.4, the tasks and their associated prompts, including the survey questions, were distributed as designed for all three forms at each grade level. No task was missing from a form, no extra task was administered, and the number of responses matched the number of prompts for each task times the number of students that took the form. Thus, it is reasonable to assume that all the prompts within a task were administered as designed.

Table 5.7. Task Distribution on Forms

Grade 5				Grade 8			
Task Code	#Prompts	#Students	#Responses	Task Code	#Prompts	#Students	#Responses
Form A							
2135	7	1,001	7,007	2133	5	1,000	5,000
2136	6	1,001	6,006	2151	5	1,000	5,000
2142	4	1,001	4,004	2154	6	1,000	6,000
2144	4	1,001	4,004	2156	6	1,000	6,000
2146	6	1,001	6,006	2158	6	1,000	6,000
2147	6	1,001	6,006	2160	7	1,000	7,000
2149	8	1,001	8,008	2161	5	1,000	5,000
Survey Q1	1	1,001	1,001	Survey Q1	1	1,000	1,000
Total #Prompts	42	–	–	–	41	–	–

Grade 5				Grade 8			
Task Code	**#Prompts**	**#Students**	**#Responses**	**Task Code**	**#Prompts**	**#Students**	**#Responses**
Form B							
2136	6	999	5,994	2133	5	999	4,995
2139	4	999	3,996	2150	6	999	5,994
2143	8	999	7,992	2154	6	999	5,994
2144	4	999	3,996	2155	5	999	4,995
2145	5	999	4,995	2156	6	999	5,994
2147	6	999	5,994	2158	6	999	5,994
2149	8	999	7,992	2160	7	999	6,993
Survey Q1	1	999	999	–	–	–	–
Total #Prompts	42	–	–	–	41	–	–
Form C							
2135	7	1,000	7,000	2133	5	1,001	5,005
2139	4	1,000	4,000	2150	6	1,001	6,006
2142	4	1,000	4,000	2151	5	1,001	5,005
2143	8	1,000	8,000	2155	5	1,001	5,005
2145	5	1,000	5,000	2156	6	1,001	6,006
2146	6	1,000	6,000	2160	7	1,001	7,007
2147	6	1,000	6,000	2161	5	1,001	5,005
Survey Q1	1	1,000	1,000	Survey Q1	1	1,001	1,001
Survey Q2	1	1,000	1,000	Survey Q2	1	1,001	1,001
Total #Prompts	42	–	–	–	41	–	–

Table 5.8 presents the number of prompts administered to students across all tasks for each grade level based on prompt position during the simulation as compared to the intended prompt position based on the test design (i.e., it shows how the position of a prompt from the simulation results compared to the designed position of the prompt within a task). If a prompt is designed to be the nth prompt in a task, the engine is expected to administer this prompt as the nth prompt within the task. If the engine performs as expected, the nonzero numbers will appear on the diagonal line. If, for example, a prompt is designed as the first prompt within a set but the engine administers it as the second prompt, the table would show a nonzero number in the position with row =1 and column=2, which is not on the diagonal line.

As shown in Table 5.8, all the nonzero numbers are on the diagonal line, which indicates that the simulation position matched the intended designed position perfectly. For example, 21,000 prompts were administered as the first prompt within a task in the simulation (i.e., 3,000 simulated students × 7 tasks on a form = 21,000 prompts administered as the first prompt within a task). All these 21,000 prompts have IDs that matched the IDs of the first prompt within a task based on the test design. The total number of prompts for some positions are smaller because some tasks have fewer than eight prompts.

Table 5.8. Prompt Position: Test Design vs. Simulations

Design Position	#Prompts by Simulated Position							
	1	2	3	4	5	6	7	8
Grade 5								
1	21,000	0	0	0	0	0	0	0
2	0	21,000	0	0	0	0	0	0
3	0	0	21,000	0	0	0	0	0
4	0	0	0	21,000	0	0	0	0
5	0	0	0	0	15,000	0	0	0
6	0	0	0	0	0	13,001	0	0
7	0	0	0	0	0	0	6,000	0
8	0	0	0	0	0	0	0	3,999
Grade 8								
1	21,000	0	0	0	0	0	0	–
2	0	21,000	0	0	0	0	0	–
3	0	0	21,000	0	0	0	0	–
4	0	0	0	21,000	0	0	0	–
5	0	0	0	0	21,000	0	0	–
6	0	0	0	0	0	11,998	0	–
7	0	0	0	0	0	0	3,000	–

Task positions were also checked to see if their position on a test form is roughly balanced across students. Figure 5.3 and Figure 5.4 present the results at each grade level, respectively. The x-axis represents the position of prompts on a form, which ranges from 0 to about 40. The position of the first prompt of a task is used as the position of the task. The y-axis represents the number of students. Each circle represents the number of students that took the task at a particular position.

As shown in the figures, most tasks have a balanced number of students at various positions. Four tasks are in fixed positions (i.e., Tasks 2135 and 2136 for Grade 5 and Tasks 2156 and 2161 for Grade 8), with one task at the beginning and one near the end of the form because of the content similarity concern. In addition, the position of the survey questions on each form was also checked. The survey questions all appear at the end of the test forms as designed. For example, Survey Question 1 for Grade 5 appears at position 42 for 2,000 students that took Forms A and B and at position 41 for 1,000 students that took Form C. Survey Question 2 for Grade 5 appears at position 42 for 1,000 students that took Form C.

Figure 5.3. Simulated Task Position—Science Grade 5

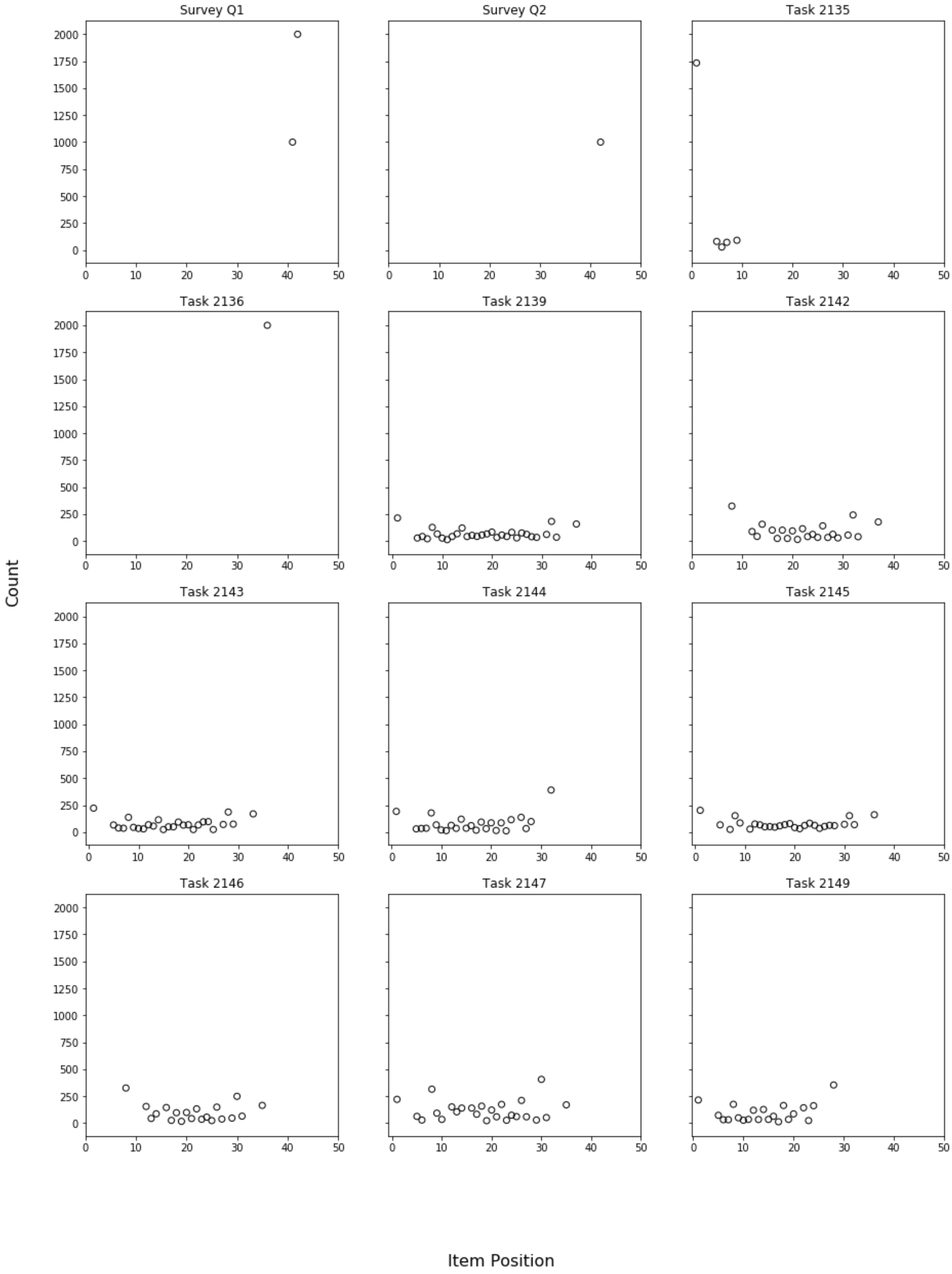
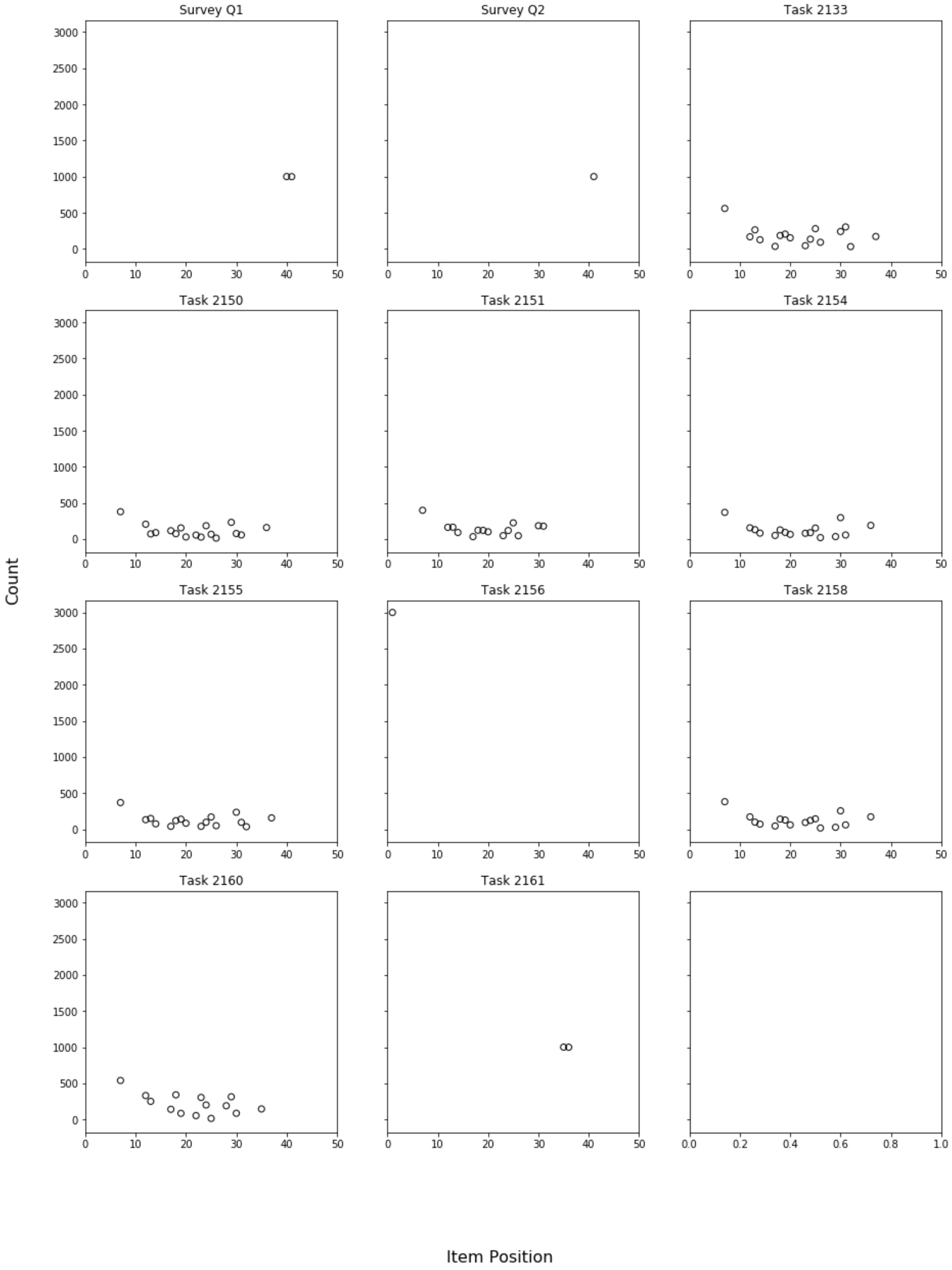


Figure 5.4. Simulated Task Position—Science Grade 8



Section 6: Psychometric Analyses

Psychometric analyses were not conducted for Spring 2020 due to the administration cancellation.

Section 7: Standard Setting

No standard setting was held in 2019–2020. If testing and scoring had occurred in 2020, the cut scores would have been the same as in 2018 and 2019. Nebraska's statewide assessment system for ELA and mathematics underwent significant changes between the 2016 and 2017 administrations, so cut scores for ELA and mathematics were set following the Spring 2018 administration at standard setting and cut score review meetings from July 26–28, 2018, using the Item-Descriptor (ID) Matching method to delineate the Developing, On Track, and CCR Benchmark achievement levels. The purpose of the standard setting was to set new cut scores for mathematics, whereas the purpose of the cut score review was to validate the existing cut scores for ELA. This section summarizes the process and results from those meetings. For more in-depth information, please refer to the full standard setting and cut score review reports (EdMetric, 2018a, 2018b). Standard setting will take place for the new NSCAS Science assessment following the first operational administration.

7.1. Overview

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

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

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

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

7.2. ID Matching Method

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

7.3. Meeting Process

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

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

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

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

7.4. ALD Revision

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

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

7.5. Final Results

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

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

	Cut Scores		Impact Data			
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 8: Test Results

Test results are not provided for Spring 2020 due to the administration cancellation.

Section 9: Reliability

The reliability/precision of the 2020 NSCAS assessments is not able to be properly evaluated due to the cancellation of the Spring 2020 administration. Please reference Section 5.2.4 for score precision and reliability results from the constraint-based engine simulations for ELA and mathematics.

Section 10: Validity

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

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). Because the Spring 2020 administration was cancelled, validity evidence based on the technical quality of the assessments is not available. Therefore, this chapter focuses on validity evidence based on test content and response processes.

10.1. Intended Purposes and Uses of Test Scores

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

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

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

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

The unintended uses of the NSCAS are as follows:

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

10.2. Sources of Validity Evidence

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

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

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

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

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

10.3. Evidentiary Validity Framework

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

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

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

Table 10.2. Sources of Validity Evidence based on Test Content

Test Purpose	Summary of Evidence	Tech Report Sections
1. Measure and report Nebraska students’ depth of achievement regarding the Nebraska College and Career Ready Standards	• Bias is minimized through Universal Design and accessibility resources. • Blueprint, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity. • The item pool and item selection procedures adequately support the test design.	2,9
2. Report if student achievement is sufficient academic proficiency to be on track for achieving college readiness	• Nebraska’s College and Career Ready Standards are based on skills leading to college and career readiness across grades. • Blueprint, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity.	2
3. Measure students’ annual progress toward college and career readiness	• Nebraska’s College and Career Ready Standards are based on skills leading to college and career readiness across grades. • Blueprint, passage specifications and item specifications are aligned to grade-level content, process skills, and associated cognitive complexity.	2
4. Inform teachers how student thinking differs along different areas of the scale as represented by the ALDs as information to support instructional planning.	• Blueprint, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity. • Blueprint and ALDs were developed in consultation with Nebraska educators. • Reporting categories align with the structure of the Nebraska standards to support the interpretation of the test results.	2,4,7

Test 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	• Bias is minimized through Universal Design and accessibility resources. • Assessments are administered with appropriate accommodations.	2,3

Table 10.3. Sources of Validity Evidence based on Response Process

Test Purpose	Summary of Evidence	Tech Report Sections
1. Measure and report Nebraska students' depth of achievement regarding the Nebraska College and Career Ready Standards	• Bias is minimized through Universal Design and accessibility resources. • Blueprint, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity. • Achievement levels were set consistent with best practice.	2
2. Report if student achievement is sufficient academic proficiency to be on track for achieving college readiness	• Blueprint, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity. • Achievement levels were vertically articulated.	2
3. Measure students' annual progress toward college and career readiness	• Blueprint, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity. • Achievement levels were vertically articulated.	2
4. Inform teachers how student thinking differs along different areas of the scale as represented by the ALDs as information to support instructional planning.	• Blueprint, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity. • Range and Policy ALDs were developed in consultation with Nebraska educators with the goal of providing information to Nebraska educators.	2
5. Assess students' construct relevant achievement in ELA, mathematics, and science for all students and subgroups of students	• Bias is minimized through Universal Design and accessibility resources. • Assessments are administered with appropriate accommodations.	2,3

Table 10.4. Sources of Validity Evidence based on Internal Structure

Test Purpose	Summary of Evidence	Tech Report Sections
1. Measure and report Nebraska students' depth of achievement regarding the Nebraska College and Career Ready Standards	N/A for 2020	
2. Report if student achievement is sufficient academic proficiency to be on track for achieving college readiness	N/A for 2020	
3. Measure students' annual progress toward college and career readiness	N/A for 2020	
4. Inform teachers how student thinking differs along different areas of the scale as represented by the ALDs as information to support instructional planning.	• Range and Policy ALDs were developed in consultation with Nebraska educators with the goal of providing information to Nebraska educators. • Reporting categories align with the structure of the Nebraska standards to support the interpretation of the test results. • Items aligned with ALDs to support item writing processes.	2,7
5. Assess students' construct relevant achievement in ELA, mathematics, and science for all students and subgroups of students	N/A for 2020	

Table 10.5. Sources of Validity Evidence based on Other Variables

Test Purpose	Summary of Evidence	Tech Report Sections
1. Measure and report Nebraska students' depth of achievement regarding the Nebraska College and Career Ready Standards	N/A for 2020	
2. Report if student achievement is sufficient academic proficiency to be on track for achieving college readiness		

Test Purpose	Summary of Evidence	Tech Report Sections
3. Measure students' annual progress toward college and career readiness	N/A for 2020	
4. Inform teachers how student thinking differs along different areas of the scale as represented by the ALDs as information to support instructional planning.		
5. Assess students' construct relevant achievement in ELA, mathematics, and science for all students and subgroups of students		

10.4. Interpretive Argument Claims

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

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

Arguments	Tech Report Section(s)	Evidence
Careful test and item development through iteration occurred to ensure that the test measured the College and Career Ready standards.	2.Test Design and Development	Description of the development and review process for item, passage, and test
Test score interpretations are comparable across students.	N/A for 2020	
Test administrations were secure and standardized.	N/A for 2020	
Scoring was standardized and accurate.	N/A for 2020	
Achievement standards were rigorous and technically sound.	7. Standard Setting	Documentation of the mathematics standard setting procedures and ELA cut score review process, including the methodology, identification of workshop participants, and implementation process, and ALD development and validation
Assessments were accessible to all students and fair across student subgroups.	N/A for 2020	

10.5. NSCAS Validity Argument

Evidence based on internal structure or other variables was not available for the 2020 NSCAS, as testing was cancelled and no test result data were collected. The test development and technical quality of the NSCAS General Summative assessments supports the intended test score interpretations that are provided through the Reporting ALDs and scale scores. The blueprint, passage specifications, item specifications, and ALD development process show that the NSCAS assessments are aligned to grade-level content. For ELA and mathematics, there is evidence that the student response processes associated with cognitive complexity specified in the standards and blueprint is behaving as intended. As an added dimension for adaptive testing, the NSCAS ELA and Mathematics assessments demonstrated that the tests administered to students conform to the blueprint during the constraint-based engine simulation studies.

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

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

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

References

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

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

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

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

EdMetric. (2019). *Alignment study for Nebraska Student-Centered Assessment System, mathematics grades 3–8*. Report provided to NDE.

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

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

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. Committee on a Conceptual Framework for New K–12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education Washington, DC: The National Academies Press.

Nebraska Department of Education (NDE). (2019, August). *Nebraska Student-Centered Assessment System (NSCAS) summative & alternate accessibility manual*. Retrieved from https://cdn.education.ne.gov/wp-content/uploads/2019/02/NSCAS-Summative-and-Alternate-Accessibility-Manual-2.8.19.pdf.

NWEA. (2020, January). *Constraint-based engine simulation report for the Spring 2020 NSCAS ELA and mathematics assessments*. Report provided to NDE. Portland, OR: NWEA.

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

Plake, B. S., Huff, K., & Reshetar, R. (2010). Evidence-centered assessment design as a foundation for achievement-level descriptor development and for standard setting. *Applied Measurement in Education, 23*(4), 342–357.

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

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

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

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

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments on mathematics and science education*. (Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6). Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (1999). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. Washington, DC: Council of Chief State School Officers.

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